Randomisierte Dimensionsreduktion und binäre Klassifikation

David Molitor

Geboren am 17. August 1992 in Bonn

6. Februar 2015

Bachelorarbeit Mathematik

Betreuer: Dr. Tino Ullrich

Institut für Numerische Simulation

Mathematisch-Naturwissenschaftliche Fakultät der

Rheinischen Friedrich-Wilhelms-Universität Bonn

Inhaltsverzeichnis

Zusammenfassung

Benötigte Rechenzeit und Speicherplatz von Klassifikationsalgorithmen hängt von der Dimension der Daten ab. Daher möchte man häufig vor Beginn der Klassifikation die Dimension reduzieren. Eine Technik die immer häufiger verwendet wird, sind die zufälligen Johnson-Lindenstrauss Einbettungen. In den Abhandlungen von Bandeira et al. und Dasgupta wurde die Klassifikation nach Anwendung einer solchen Einbettung auf zwei verschiedene Arten beleuchtet. Bandeira et al. wählen eine geometrische Sichtweise, Dasgupta hingegen eine probabilistische. Wir stellen die Ergebnisse beider Abhandlungen vor und vergleichen sie. Dann führen wir einige numerische Experimente durch, die die Aussagen der Abhandlungen auf ihre Übertragbarkeit in den klassifikatorischen Rahmen überprüfen sollen. Wir stellen fest, das weder die Theorie von Bandeira et al. noch die von Dasgupta ausreicht um die Beobachtungen vollständig zu beschreiben. Unsere Experimente stehen sogar der von Bandeira et al. postulierten Übertragbarkeit entgegen.

Einleitung

Klassifikation ist ein Problem des maschinellen Lernens. Man möchte aufgrund von Trainingsdaten die echten Daten möglichst den richtigen Klassen zuordnen. Da die Daten auf reellen Objekten beruhen, sucht man eine Repräsentation für ihre Eigenschaften. Dies geschieht durch die Darstellung als Vektoren in einem eukildischen Raum. Seine Dimension entspricht der Anzahl der Eigenschaften, diese kann unter Umständen sehr hoch sein. Bei binärer Klassifikation existieren nur zwei Klassen. Als Algorithmus werden häufig Stützvektormaschinen benutzt. Man versucht zwischen die beiden Klassen eine trennende Hyperebene zu legen. Dafür müssen auf den Trainingsdaten paarweise innere Produkte berechnet werden. Die Kosten für die Berechnung der Hyperebene skalieren dann wie $O(n^2 \cdot d)$, wobei n die Anzahl der Daten und d ihre Dimension ist. Sollte d sehr groß sein, kann es sich rentieren vor Beginn der Klassifikation die Dimension der Daten zu reduzieren. Die von Johnson und Lindenstrauss 1984 vorgestellten Abbildungen (Auch Johnson-Lindenstrauss Einbettungen) sind in den letzten Jahren zu einem beliebten Werkzeug geworden. Das liegt unter anderem daran, dass diese Abbildungen relative ε-Isometrien sind.
Hat man nun die Daten mittels einer solchen Abbildung in einen Raum der Dimension $k < d$ projiziert, so muss man erneut eine Hyperebene finden, die die beiden Datenwolken trennt. Das eine solche existiert kann jetzt aber nicht mehr garantiert werden. Das heißt die Vorhersagequalität des Klassifikators wird durch vorherige Dimensionsreduktion beeinträchtigt. Wie klein man k wählen kann, ohne dass die Vorhersagequalität zu stark abnimmt, hängt intuitiv davon ab wie sehr sich die Datenwolken durch die Dimensi-

onsreduktion annähern. Diese Problemstellung wurde in den Abhandlungen [4] und [5] untersucht.
Die neuere Abhandlung von Bandeira, Mixon und Recht [4] beschäftigt sich mit disjunkten, konvexen Mengen und sucht nach unteren Schranken für k so das diese nach Dimensionsreduktion noch separiert bleiben. Für den Fall von zwei Bällen erhalten sie die Grenze $k \geq ((r_1+r_2)/\|c_1-c_2\|_2)^2 d+O(\sqrt{d})$.
In der Abhandlung von Dasgupta [5] wird die Fragestellung probabilistisch betrachtet. Für mehrdimensionale Normalverteilungen definiert man ihre c-Separiertheit. Dann zeigt man, dass randomisierte Dimensionsreduktion in einen Raum der Dimension $k > O(\varepsilon^{-2})$ die Separiertheit höchstens auf $c\sqrt{1-\varepsilon}$ verschlechtert.
Ob sich die Ergebnisse aus [4] und [5] tatsächlich auf den klassifikatorischen Rahmen übertragen lassen, wollen wir anhand numerischer Experimente überprüfen. Das erste Experiment überprüft ob und wie die kleinste Dimension k_0 vom Abstand der Bälle abhängt. Im zweiten überprüfen wir dann die Abhängigkeit von der Ursprungsdimension d.
Wir beginnen im ersten Abschnitt damit einige Grundlagen einzuführen, die für manche Beweise benötigt werden. Haben wir dies getan, so stellen wir im zweiten Abschnitt die Johnson-Lindenstrauss Einbettungen vor. Der dritte Abschnitt widmet sich Techniken der binären Klassifikation. In Abschnitt 4 stellen wir die Abhandlung [4] vor. Wir konzentrieren uns dabei insbesondere auf das Ergebnis das man für zwei disjunkte Bälle erhält. Im folgenden, fünften, Abschnitt möchten wir dann noch kurz das Ergebnis aus [5] vorstellen. Der sechste Abschnitt enthält die Ergebnisse unserer Experimente. Unsere Ergebnisse rekapitulieren wir nocheinmal im letzten Abschnitt.

1 Grundlagen

1.1 Wahrscheinlichkeitstheoretische Grundlagen

Im Folgenden ist $(\Omega, \mathcal{F}, \mathbb{P})$ ein Wahrscheinlichkeitsraum. Für die folgenden Grundlagen verweisen wir auch auf die Bücher von Foucart und Rauhut [14] und Georgii [7].

Definition 1.1 (Erwartungswert)**.** Sei Z eine $\mathbb{P}$-integrierbare Zufallsvariable. Dann ist der Erwartungswert von Z definiert über

$$\mathbb{E}[Z] = \int_\Omega Z \, d\mathbb{P}$$

Im Fall von reellen Zufallsvariablen, kann man die Formel umformen zu

$$\mathbb{E}[Z] = \int_{\mathbb{R}} x\phi(x) \, dx,$$

wobei $\phi(x)$ die Wahrscheinlichkeitsdichte von Z ist.

Definition 1.2 (Normalverteilte Zufallsvariablen)**.** Eine Zufallsvariable Z auf $(\Omega, \mathcal{F}, \mathbb{P})$ heißt normalverteilt, falls für ihre Verteilungsfunktion $F_Z(x) = \mathbb{P}(Z \leq x)$ gilt, dass

$$F_Z(t) - F_Z(s) = \int_s^t \phi_{m,\sigma^2}(x)\, dx,$$

wobei $\phi_{m,\sigma^2}(x)$ die Wahrscheinlichkeitsdichte der Normalverteilung mit Mittelwert $m \in \mathbb{R}$ und Standardabweichung $\sigma > 0$ ist. ϕ_{m,σ^2} hat die Formel

$$\phi_{m,\sigma^2} = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{(x-m)^2}{2\sigma^2}\right).$$

Man schreibt $Z \overset{(D)}{=} N(m, \sigma^2)$

Definition 1.3 (Mehrdimensionale Normalverteilung)**.** Eine d-dimensionale reelle Zufallsvariable Z ist normalverteilt mit Erwartungswert $\mu \in \mathbb{R}^d$ und positiv definiter Kovarianzmatrix $\Sigma \in \mathbb{R}^{d\times d}$, wenn sie die Dichtefunktion

$$f_Z(x) = \frac{1}{(2\pi)^{d/2} \cdot \det(\Sigma)^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right) \quad , x \in \mathbb{R}^d$$

besitzt.

Lemma 1.4. *Sei X eine reellwertige Zufallsvariable mit Verteilung $\mathbb{P}$ und sei $f : (0, \infty) \to \mathbb{R}$ eine monoton wachsende Funktion, dann gilt*

$$\mathbb{P}[X > x] \leq \frac{\mathbb{E}[f(X)]}{f(x)} \ .$$

Beweis

$$\mathbb{P}[X > x] = \mathbb{E}[\mathbb{1}_{X>x}] \leq \mathbb{E}[\mathbb{1}_{X>x}] \frac{f(X)}{f(x)} \leq \frac{\mathbb{E}[f(X)]}{f(x)} \ . \quad \square$$

Satz 1.5 (Jensen Ungleichung)**.** *Sei $f : \mathbb{R} \to \mathbb{R}$ eine konvexe Funktion und Z eine integrierbare, reellwertige Zufallsvariable. Dann gilt*

$$f(\mathbb{E}[Z]) \leq \mathbb{E}[f(Z)].$$

Satz 1.6 (Erste Bonferroni Ungleichung)**.** *Seien $\{E_1, \ldots, E_n\}$ Ereignisse in einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{F}, \mathbb{P})$. Dann gilt*

$$\mathbb{P}\left[\bigcup_{i=1}^n E_i\right] \leq \sum_{i=1}^n \mathbb{P}[E_i] \ .$$

Definition 1.7 (Momente)**.** Sei Z eine Zufallsvariable dann definieren wir das p-te Moment M^p von X per

$$M^p(Z) = \mathbb{E}\left[Z^p\right].$$

Definition 1.8 (Momenterzeugende Funktion)**.** Sei Z eine normaverteilte Zufallsvariable mit Erwartungswert $\mu \in \mathbb{R}$ und Standardabweichung $\sigma > 0$. Dann gilt für

$$\Psi(x) = exp\left(x\mu + \frac{\sigma^2 x^2}{2}\right),$$

die folgende Gleichung

$$M^p(Z) = \mathbb{E}\left[Z^p\right] = \frac{\partial^p}{\partial x^p}\Psi(0).$$

Wir nennen Ψ die momenterzeugende Funktion von Z.

Definition 1.9 (Subexponentielle Verteilung)**.** Sei Z eine Zufallsvariable, $\beta, c, r > 0$. Z heißt subexponentiell verteilt, falls

$$\mathbb{P}\left[|Z| \geq r\right] \leq \beta e^{-cr} \ .$$

Definition 1.10 (Subgaußsche Verteilung)**.** Sei Z eine Zufallsvariable. Dann besitzt Z eine subgaußsche Verteilung, falls es β, c gibt, so dass für alle $t > 0$ gilt

$$\mathbb{P}\left[|Z| \geq t\right] \leq \beta e^{-ct^2} \ .$$

Eine Matrix $A \in \mathbb{R}^{k \times d}$ heißt subgaußsche Matrix, falls ihre Einträge subgaußsche Verteilungen besitzen.

Der folgende Satz kann in [14, 7.32] gefunden werden.

Satz 1.11 (Bernstein Ungleichung)**.** *Seien $Z_1, \ldots, Z_n$ unabhängige subexponentielle Zufallsvariablen mit Erwartungswert* 0*. Dann gilt*

$$\mathbb{P}\left[\left|\sum_{i=1}^{n} Z_i\right| \geq t\right] \leq 2\exp\left(-\frac{(ct)^2/2}{2\beta n + ct}\right) \ ,$$

wobei c, β wie in Definition 1.9.

1.2 Weiteres

Lemma 1.12 (Cauchy-Schwarz Ungleichung)**.** *Seien x, y Elemente eines reellen Vektorraumes mit Skalarprodukt $\langle \cdot, \cdot \rangle$. Sei $\| \cdot \|$ die von dem Skalarprodukt induzierte Norm $\|x\| = \sqrt{\langle x, x \rangle}$. Dann gilt*

$$\langle x, y \rangle \leq \|x\| \cdot \|y\| \ .$$

Definition 1.13 (Gammafunktion)**.** Sei x eine positive reelle Zahl, dann ist die Gammafunktion von x, also $\Gamma(x)$ definiert durch

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t}\, dt \ .$$

Definition 1.14 (O-Notation)**.** Die O-Notation gehört zu den Landau Symbolen. Diese dienen dazu, das asymptotische Verhalten von Funktionen anzugeben.
Seien f, g zwei Funktionen. Dann bedeutet $f = O(g)$, dass f nicht wesentlich schneller wächst als g. Genauer gesagt

$$f = O(g) \Longleftrightarrow \limsup_{x \to \infty} \left| \frac{f(x)}{g(x)} \right| < \infty.$$

2 Johnson-Lindenstrauss Einbettungen

Hier führen wir die Johnson-Lindenstrauss [12] Einbettungen (kurz: JLE) ein.

2.1 Das Johnson-Lindenstrauss Lemma

Sei $X \subset \mathbb{R}^d$ eine Menge von n Punkten und $0 < \varepsilon < 1$. Dann existiert für $k \geq O\left(\varepsilon^{-2} \log n\right)$ eine Abbildung $f : \mathbb{R}^d \to \mathbb{R}^k$, so dass für alle $u, v \in X$ gilt

$$(1-\varepsilon)\|u-v\|_2^2 \leq \|f(u)-f(v)\|_2^2 \leq (1+\varepsilon)\|u-v\|_2^2.$$

Betrachtet man die Abbildung f als lineare Abbildung $R \in \mathbb{R}^{k \times d}$, so ergibt sich die folgende Formulierung

$$(1-\varepsilon)\|u-v\|_2^2 \leq \|Ru-Rv\|_2^2 \leq (1+\varepsilon)\|u-v\|_2^2 \ .$$

Eine solche Abbildung nennt man Johnson-Lindenstrauss Einbettung. Man nennt eine solche Abbildung *klassische Johnson-Lindenstrauss Einebttung*, falls die Einträge von R, $N\left(0, 1/k\right)$ verteilt sind.

Satz 2.1. *Sei $X \subset \mathbb{R}^d$ eine endliche Menge, $\varepsilon > 0$, $\eta > 0$ und $R \in \mathbb{R}^{k \times d}$ eine Matrix mit $N(0, 1/k)$ Einträgen. Dann gilt*

$$\mathbb{P}\left[\forall u, v \in X \ : \ \|u-v\|_2^2(1-\varepsilon) \leq \|Ru-Rv\|_2^2 \leq \|u-v\|_2^2(1+\varepsilon)\right] \geq 1-\eta \, ,$$

falls

$$k \geq O\left(\varepsilon^{-2} \log(n/\eta)\right) .$$

Besonders interessant hieran ist die relative isometrische Eigenschaft. Für uns ist es wichtig im Hinterkopf zu behalten welche Parameter vorkommen und in welchen Abhängigkeiten sie zueinander stehen. Diese Abhängigkeiten werden in den Experimenten auf ihre Bedeutung im Rahmen binärer Klassifikation untersucht.

2.2 Johnson-Lindenstrauss Einbettungen mit Münzwurf

Anstelle der klassischen Johnson-Lindenstrauss Einbettungen, verwenden wir Matrizen, deren Einträge aus Einsen und Nullen bestehen. Diese, in [2] eingeführten, Matrizen benötigen weniger Speicherplatz und Matrix-Vektor Multiplikation ist schneller durchführbar.

Satz 2.2. *Seien X , ε, η und k wie in Satz 2.1. Sei $\tilde{R} = (r_{ij})_{i=1,j=1}^{k,d}$ eine $k \times d$ Matrix deren Einträge folgendermaßen aussehen*

$$r_{ij} = \sqrt{3} \cdot \begin{cases} +1 & \textit{mit Wahrscheinlichkeit } 1/6 \\ 0 & \textit{mit Wahrscheinlichkeit } 2/3 \\ -1 & \textit{mit Wahrscheinlichkeit } 1/6. \end{cases} \tag{1}$$

oder

$$r_{ij} = \begin{cases} +1 & \textit{mit Wahrscheinlichkeit } 1/2 \\ -1 & \textit{mit Wahrscheinlichkeit } 1/2 \ . \end{cases} \tag{2}$$

Sei $R = \frac{1}{\sqrt{k}}\tilde{R}$. *Dann ist mit Wahrscheinlichkeit größer als* $1 - \eta$

$$(1-\varepsilon)\|u-v\|_2^2 \leq \|Ru - Rv\|_2^2 \leq (1+\varepsilon)\|u-v\|_2^2 \ ,$$

für alle $u, v \in X$, *falls* $k \geq O(\varepsilon^{-2}\log n + \log \eta^{-1})$.

Falls man die Einbettung also über eine Matrix $\tilde{R}$ mit der ersten Verteilung (1), also $r_{ij} \in \{-\sqrt{3}, 0, +\sqrt{3}\}$, definiert, erhält man eine dünnbesetzte Matrix. Verwendet man diese, kann Laufzeit eingespart werden. Insbesondere erfüllt sie die gleichen Schranken wie die Matrizen mit $r_{ij} \in \{-1, +1\}$.

Beweis

Zu zeigen ist, dass die Länge eines Bildvektors um den Erwartungswert konzentriert ist. Das heißt die Wahrscheinlichkeit, dass die Länge außerhalb des gewünschten Bereiches liegt, ist sehr gering. Dann könnte man von einem Vektor auf die möglichen Kombinationen von Vektoren in der Menge X übergehen und hätte das gewünschte Resultat. Zuerst zeigen wir, dass im Erwartungswert, die Länge eines Vektors unter Projektion in etwa gleich bleibt. Dann wäre lediglich noch zu zeigen, dass für jede mögliche Kombination aus der Menge der Punkte die ε-Isometrie erfüllt ist.

1. Schritt Zuerst wollen wir $\mathbb{E}\left[\|Rx\|_2^2\right]$ bestimmen. Dazu sei $\tilde{R}_i$ die i-te Zeile der Matrix $\tilde{R}$. Für $Rx = f(x)$ ergibt sich dann

$$f(x) = \sqrt{1/k}\left(\tilde{R}_1 \cdot x, \tilde{R}_2 \cdot x, \ldots, \tilde{R}_k \cdot x\right) \ .$$

Jetzt berechnen wir $\mathbb{E}\left[\tilde{R}_i \cdot x\right]$.

$$\mathbb{E}\left[\tilde{R}_i \cdot x\right] = \mathbb{E}\left[\sum_{j=1}^{d} x_j \cdot r_{ij}\right] = \sum_{j=1}^{d} \mathbb{E}[r_{ij}]x_j = 0,$$

wobei im letzten Schritt benutzt wurde, dass die Erwartungswerte der r_{ij} (r_{ij} ist der Eintrag, der in der i-ten Zeile, in Spalte j der Matrix R steht) per

Definition 0 sind. Es folgt die Berechnung von $\mathbb{E}[(\tilde{R}_i \cdot x)^2]$. Es gilt

$$\begin{aligned}
\mathbb{E}\left[(\tilde{R}_i \cdot x)^2\right] &= \mathbb{E}\left[\left(\sum_{j=1}^{d} x_j r_{ij}\right)^2\right] \\
&= \mathbb{E}\left[\sum_{j=1}^{d}(x_j r_{ij})^2 + \sum_{l,m=1}^{d} 2x_l x_m r_{il} r_{im}\right] \\
&= \left(\sum_{j=1}^{d} x_j^2 \mathbb{E}[r_{ij}^2] + \sum_{l,m=1}^{d} x_l x_m \mathbb{E}[r_{il}]\mathbb{E}[r_{im}]\right) \\
&= \|x\|_2^2,
\end{aligned}$$

wobei wir benutzen, dass die Einträge r_{ij} von $\tilde{R}$ Erwartungswert 0 und Varianz 1 besitzen und unabhängig sind. Insgesamt gilt für $f(x)$ Folgendes:

$$\mathbb{E}[\|f(x)\|_2^2] = \mathbb{E}[\|\sqrt{1/k}(x \cdot \tilde{R}_1, \ldots, x \cdot \tilde{R}_k)\|_2^2) = \frac{1}{k}\sum_{i=1}^{k} \mathbb{E}[(x \cdot \tilde{R}_i)^2] = \|x\|_2^2.$$

Damit haben wir gezeigt, dass unsere Matrix isotrop ist. Wir müssen noch die Abschätzung aus dem Johnson-Lindenstrauss Lemma (Satz 2.1) zeigen.

2.Schritt Wir werden also zeigen, dass für die Distanz von je zwei festen, der n Vektoren aus X gilt

$$\mathbb{P}\left[(1-\varepsilon)\|u-v\|_2^2 \leq \|Ru - Rv\|_2^2 \leq (1+\varepsilon)\|u-v\|_2^2\right] > 1 - \eta^*.$$

So dass sich insgesamt bei n Punkten in X und somit $\binom{n}{2}$ möglichen Kombinationen, die Wahrscheinlichkeit **keine** JLE zu erhalten genau

$$\binom{n}{2}\eta^* \leq \eta$$

beträgt. Wir werden jetzt eine Konzentrationsungleichung beweisen, die uns dann die Aussage aus 2.1 liefert. Sicherlich sind die durch (1) und (2) definierten Matrizen subgaußsche Matrizen. Nun können wir das folgende Lemma aus dem Buch von Foucart und Rauhut [14] anwenden. Es liefert das gewünschte Resultat.

Lemma 2.3 (Konzentrationsungleichung)**.** *Sei R eine $k \times d$ Matrix mit unabhängigen, isotropen und subgaußschen Einträgen. Es gibt ein $\tilde{c}$, so dass für alle $x \in \mathbb{R}^d$ und jedes $\varepsilon \in (0,1)$ gilt*

$$\mathbb{P}\left[\left|\frac{1}{k}\|Rx\|_2^2 - \|x\|_2^2\right| \geq \varepsilon\|x\|^2\right] \leq 2\exp(-\tilde{c}\varepsilon^2 k)\ . \tag{3}$$

Beweis. Wie bereits im ersten Schritt seien die Zeilen von R R_i. Wir wissen bereits

$$\mathbb{E}\left[(R_i \cdot x)^2 - \|x\|_2^2\right] = 0 \ .$$

Da $(R_i \cdot x)$ subgauß ist, ist $(R_i \cdot x)^2 - \|x\|_2^2$ subexponentiell verteilt (siehe 1.11). Das liefert

$$\mathbb{P}\left[(R_i \cdot x)^2 - \|x\|_2^2 \geq r\right] \leq \beta e^{-cr}.$$

Als letztes beobachten wir, dass

$$\frac{1}{k}\|Rx\|_2^2 - \|x\|_2^2 = \frac{1}{k}\sum_{i=1}^{k}(|(R_i \cdot x)^2 - \|x\|_2^2) = \frac{1}{k}\sum_{i=1}^{k} Z_i \ ,$$

wir definieren also $Z_i = (R_i \cdot x) - \|x\|_2^2$. Sicherlich sind die Z_i aufgrund der Unabhängigkeit der R_i ebenfalls unabhängig. Daher folgt mit der Bernsteinungleichung 1.11, dass

$$\begin{aligned}\mathbb{P}\left(\left|\frac{1}{k}\sum_{i=1}^{k} Z_i\right| \geq \varepsilon\right) &= \mathbb{P}\left(\left|\sum_{i=1}^{k} Z_i\right| \geq k\varepsilon\right) \\ &\leq 2\exp\left(-\frac{c^2k^2\varepsilon^2}{4\beta k + 2ck\varepsilon}\right) \leq 2\exp\left(\frac{c^2}{4\beta + 2c}k\varepsilon^2\right) \ .\end{aligned}$$

Die Wahl $\tilde{c} = c^2/(4\beta + 2c)$ liefert das gewünschte Resultat. □

Nun können wir die erhaltene Formel so umformen, dass wir die untere Schranke k_0 erhalten.

Lemma 2.4 (Abschätzung)**.** *Wollen wir mit Wahrscheinlichkeit mindestens $1 - \eta$ eine Johnson-Lindenstrauss Abbildung nach 2.1 erhalten, so müssen wir*

$$k \geq \left(\varepsilon^{-2}\log(n/\eta)\right)$$

wählen.

Beweis. Die Menge X enthalte genau n Punkte. $\mathfrak{A}$ bezeichne das Ereignis, das man eine JLE erhält. Das heißt

$$\mathfrak{A} = \forall u, v \in X \ : \ \left|\|Ru - Rv\|_2^2 - \|u - v\|_2^2\right| \leq \varepsilon\|u - v\|_2^2.$$

Daraus folgt, das das Gegenereignis $\overline{\mathfrak{A}}$ folgendes ist

$$\overline{\mathfrak{A}} = \exists u, v \in X \ : \ \left|\|Ru - Rv\|_2^2 - \|u - v\|_2^2\right| > \varepsilon\|u - v\|_2^2.$$

Wir möchten $\mathbb{P}\left[\overline{\mathfrak{A}}\right]$ berechnen. Dazu benutzen wir die Bonferroni Ungleichung (Satz 1.6). Jetzt können wir anstelle des Existenzquantors, die einzelnen Summe der einzelnen Wahrscheinlichkeiten schreiben. Das führt zu

$$\mathbb{P}\left[\overline{\mathfrak{A}}\right] \leq \sum_{u,v \in X} \mathbb{P}\left[\left|\|Ru - Rv\|_2^2 - \|u - v\|_2^2\right| > \varepsilon \|u - v\|_2^2\right].$$

Jetzt schätzen wir die einzelnen Wahrscheinlichkeiten mit der Konzentrationsungleichung 2.3 ab. Dann erhalten wir

$$\mathbb{P}\left[\overline{\mathfrak{A}}\right] \leq \binom{n}{2} 2\exp(-\tilde{c}k\varepsilon^2).$$

Das heißt insgesamt beläuft sich die Wahrscheinlichkeit keine JLE zu erhalten auf

$$\begin{aligned} &\binom{n}{2} 2\exp(-\tilde{c}k\varepsilon^2) \leq \eta \\ &\Rightarrow k \geq O\left(\varepsilon^{-2}\log(n/\eta)\right). \end{aligned}$$

□

Insgesamt haben wir jetzt also gezeigt, dass die Funktion R die wir in 2.2 definiert haben, für jede der beiden Verteilungen, die gewünschten Schranken aus 2.1 einhält. Diese Art der Einbettung werden wir in unseren Experimenten benutzen. Einerseits sind sie nicht schlechter als die herkömmlichen JLE und andererseits entsteht mit Verteilung (1) eine dünnbesetzte Matrix. Das führt dazu, dass unser Algorithmus eine geringere Laufzeit hat.

3 Binäre Klassifikation

An dieser Stelle wollen wir binäre Klassifikation einführen. Besonders interessant sind dabei die Maße zur Fehlerbeurteilung und die Arten der Fehler. Wir stellen uns unsere Daten typischer Weise als Vektoren in einem euklidischen Raum vor, also $x \in \mathbb{R}^d$. Dabei ist d, die Dimension dieses Raumes, auch die Anzahl der Attribute, beziehungsweise der Eigenschaften unserer Daten. Für den Fall der binären Klassifikation muss man also entscheiden, ob ein Datenpunkt x zur Klasse -1 oder zur Klasse $+1$ ($\mathbb{K} = \{-1, +1\}$) gehören soll.

Definition 3.1 (Klassifikator)**.** Sei $\mathbb{R}^d$ der Raum unserer Daten, $\mathbb{K}$ die Menge der Klassen. Dann ist eine Funktion κ ein Klassifikator, wenn

$$\kappa : \mathbb{R}^d \to \mathbb{K}.$$

Damit ein Klassifikator eine sinnvolle Anwendung ist, muss die Entscheidung der Funktion κ, welcher Klasse ein Element x aus $\mathbb{R}^d$ zugeordnet wird, nachvollziehbar sein.

Eigentlich besitzt jeder unserer Datenpunkte eine vorher definierte Klasse, die wir allerdings nicht kennen. Das heißt es existiert eine **echte** Zuordnung der Daten $y : \mathbb{R}^d \to \mathbb{K}$, die uns die richtigen Klassen liefert. Für jeden Punkt $x \in \mathbb{R}^d$ gilt dann entweder

$$\kappa(x) \neq y(x)$$

oder Gleichheit. Unser Klassifikator, beziehungsweise unsere Entscheidungsfunktion, sollte es vermeiden x der falschen Klasse zuzuordnen. Nehmen wir an, das Daten und Zugehörigkeiten zufällig sind. Dann können wir das Problem folgendermaßen formulieren.

$$\text{Wähle } \kappa \text{ ,so dass für alle } x \in \mathbb{R}^d : \mathbb{P}\left[\kappa(x) \neq y(x)\right] \text{ minimal wird.}$$

Da wir die Verteilung nicht kennen, benötigen wir die folgenden Werkzeuge zur Fehleranalyse, die uns eine möglichst gute Abschätzung für dieses Problem liefern sollten.

3.1 Fehleranalyse

In diesem Abschnitt, werden wir die Werkzeuge zur Fehleranalyse vorstellen. Sei dazu $X \subset \mathbb{R}^d$ eine Menge an Daten. $X = \{x_1, \ldots, x_m\}$ mit zugehörigen Klassen $y_i \in \{+1, -1\} = \mathbb{K}$. Für einen Punkt können wir nun die Null-Eins Verlustfunktion definieren.

Definition 3.2 (Null-Eins Verlustfunktion)**.** Sei $x \in \mathbb{R}^d$ ein Punkt, $y \in \mathbb{K}$ sei seine wahre Klasse und $\kappa : \mathbb{R}^d \to \mathbb{K}$ eine Entscheidungsfunktion. Dann hat die Null-Eins Verlustfunktion folgende Gestalt

$$\begin{aligned} &l : \mathbb{R}^d \to \{0, 1\} \\ &l(\kappa(x), y) = \begin{cases} 0 & \text{wenn } \kappa(x) = y \\ 1 & \text{wenn } \kappa(x) \neq y \end{cases} . \end{aligned}$$

Die Funktion ist also genau dann 1, wenn der Datenpunkt nicht der richtigen Klasse zugeordnet wurde.

Hieraus kann man direkt den mittleren Trainingsfehler ableiten.

Definition 3.3 (Trainingsfehler)**.** Ein Trainingsfehler ist ein Fehler auf den Trainingsdaten. Den Trainingsfehler definiert man über 3.2 wie folgt. (Hierbei ist X die Menge der m Trainingsdaten x_i, die die echten Klassen y_i besitzen.)

$$R_{emp,\,X} = \frac{1}{m} \sum_{i=1}^{m} l(\kappa(x_i), y_i) \ .$$

Also der Anteil der Trainingsdaten die den falschen Klassen zugeordnet worden sind. Ziel sollte es sein, den Trainingsfehler möglichst gering zu halten. Er ist eine Annäherung an den unbekannten Testfehler.

Definition 3.4 (Testfehler)**.** Sei $X = \{x_1, \ldots, x_m\} \subset \mathbb{R}^d$ die Menge der Trainingsdaten. Sei κ eine Entscheidungsfunktion, die durch das Training entstanden ist. Sei $\mathbb{P}(x, y)$ die unbekannte Zufallsverteilung auf $\mathbb{R}^d \times \mathbb{K}$ der Klassenzugehörigkeiten $y \in \mathbb{K}$.

$$\mathbb{P}(x, y) = \mathbb{P}\left(x \text{ gehört zu der Klasse } y\right) \ .$$

Der Testfehler ist

$$R_X = \int\limits_{\mathbb{R}^d \times \mathbb{K}} l\left(\kappa(x), y\right) \, d\mathbb{P}(x, y).$$

Sicherlich können wir den Testfehler angeben, wenn wir die Verteilung $\mathbb{P}$ kennen. Dann lässt sich obiges Integral einfach ausrechnen. Kennt man die Verteilung aller Daten nicht, bedient man sich entweder dem empirischen Trainingsfehler (der ist ja auf jeden Fall bekannt), oder man berechnet für einen größeren Datensatz den empirischen Testfehler. Die Erfolgsrate, $\mathcal{S}$ für "Score", erhalten wir in dem wir $1-$Testfehler rechnen.

Damit ein Klassifikator sinnvoll ist, bedarf es also zwei Kriterien. Zum einen

muss die Entscheidung welcher Klasse ein Punkt x zugeordnet wird nachvollziehbar und reproduzierbar sein. Zum anderen sollte es keinen Klassifikator geben, der auf den Trainingsdaten einen kleineren empirischen Fehler erzeugt. Wäre das der Fall, so könnte man nicht begründen, warum man diesen (besseren) Klassifikator nicht gewählt hat. Eine intuitive Möglichkeit einen Klassifikator zu finden sind Abstandsklassifikatoren. Also den Abstand des neuen Punktes zu denen der beiden Klassen zu messen und ihm diejenige zuzuordnen, die ihm am nächsten liegt.

3.2 Abstandsklassifikatoren

Ein Abstandsklassifikator misst den Abstand eines neuen Datenpunktes zu allen Klassen und weist ihm dann die Klasse zu, zu der er den geringsten Abstand hat.

Definition 3.5 (Abstandsklassifikator)**.** Sei $X \subset \mathbb{R}^d$ die Menge der Trainingsdaten, mit zugeordneten Klassen -1 oder $+1$. Sei weiterhin $X_{+1} \subset X$ die Menge der Punkte aus X, mit Klasse $+1$ und $X_{-1} \subset X$ die Menge der Punkte mit Klasse -1. Sei $x \in \mathbb{R}^d$ ein neuer Datenpunkt, der einer der beiden Klassen zugeordnet werden soll.
Wir berechnen den Abstand von x zu den beiden Mittelpunkten der Klassen und weisen x diejenige Klasse zu, deren Mittelpunkt am nächsten liegt. Also sei $x^*_{-1} = \sum_{x_i \in X_{-1}} \frac{x_i}{|X_{-1}|}$, der Mittelpunkt der Klasse -1 ($|X_{-1}|$ ist die Anzahl an Elementen in X_{-1}.) und x^*_{+1} der Mittelpunkt der Klasse $+1$. Dann lautet die Funktion κ:

$$\kappa(x) = \begin{cases} -1 & \iff \|x - x^*_{-1}\| < \|x - x^*_{+1}\| \\ +1 & \text{sonst.} \end{cases}$$

Definition 3.6 (Nächste Nachbarn Klassifikator)**.** Wenn wir anstelle der Mittelpunkte der beiden Klassen, den nächsten Nachbarn als Bezugspunkt wählen, so erhalten wir den Nächsten-Nachbarn-Klassifikator.

$$\kappa(x) = \kappa(y^*), \;\; y^* = arg\min_{y \in X} \|y - x\| \; .$$

Man beachte, dass wir bei dieser Methode die bereits bekannten Datenpunkte (Trainingsdaten) und ihre Klassen auch in der Funktion κ speichern. Was können wir aus diesen einfachen Beispielen lernen? Sicherlich sind die Abstände der Trainingsdaten wichtig. Für möglichst gute Klassifikation sollte man also versuchen diese zu erhalten. Außerdem kann man sich auch vorstellen welche Probleme bei Dimensionsreduktion auftreten können. Das schlichte „Vergessen“ von Einträgen könnte ein Problem sein, wenn sich die Klassen nur in wenigen Einträgen unterscheiden.
Der folgende Abschnitt beschäftigt sich mit einer der wichtigsten Methode

im Bereich der Klassifikation. Die vorgestellten Techniken werden auch in unseren Experimenten verwendet. Wir sollten also ein besonderes Augenmerk auf Fehlerquellen haben.

3.3 Stützvektormaschinen (Support Vector Machines)

Die Idee der Stützvektormaschinen (SVM) ist es, die Klassen durch eine Hyperebene zu trennen. Diese wird durch die Trainingsdaten erzeugt.

Definition 3.7 (Hyperebene)**.** Man kann eine Hyperebene über ein Tupel (w, b) definieren, wobei $w \in \mathbb{R}^d$ und $b \in \mathbb{R}$. Die Hyperebene besteht dann aus allen Punkten $x \in \mathbb{R}^d$ mit

$$\langle w, x \rangle + b = 0,$$

w ist der Normalenvektor der Ebene und b die Verschiebung.

Die SVM berechnet also zuerst basierend auf den Trainingsdaten eine Hyperebene, die die Klassen trennt. Dazu ist es notwendig, dass eine solche Hyperebene überhaupt existiert. Das heißt, die Klassen sollen linear separierbar sein.

Definition 3.8 (Linear separierbar)**.** Zwei Mengen $A, B \subset \mathbb{R}^d$ heißen linear separierbar, falls reelle Zahlen $w_1, \dots w_n$und $b \in \mathbb{R}$ existieren, so dass für alle $a = (a_1, \dots a_n) \in A, b = (b_1, \dots b_n) \in B$ gilt

$$\sum_{i=1}^{n} w_i \cdot a_i < b < \sum_{j=1}^{n} w_j \cdot b_j \ .$$

Man beachte: (w, b) spannt eine Hyperebene auf!

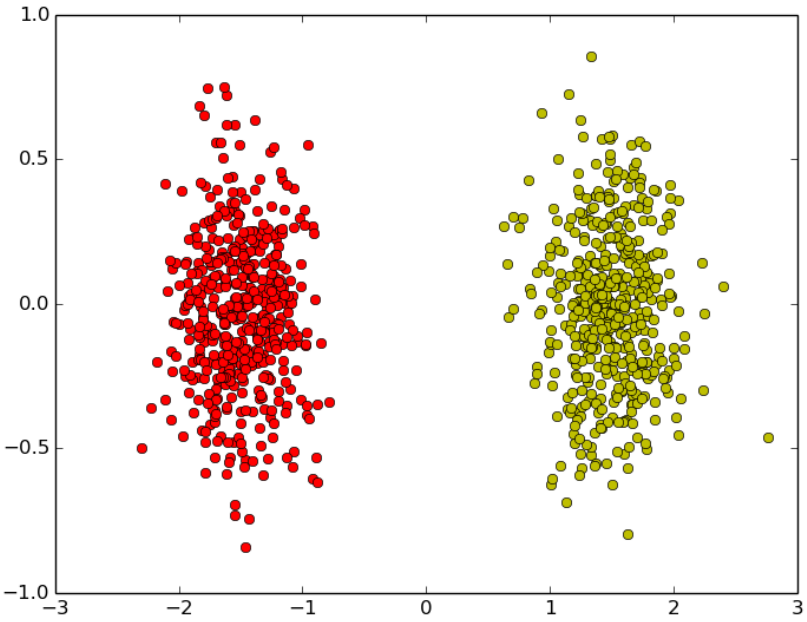

Abbildung 1: Linear separierbare Daten

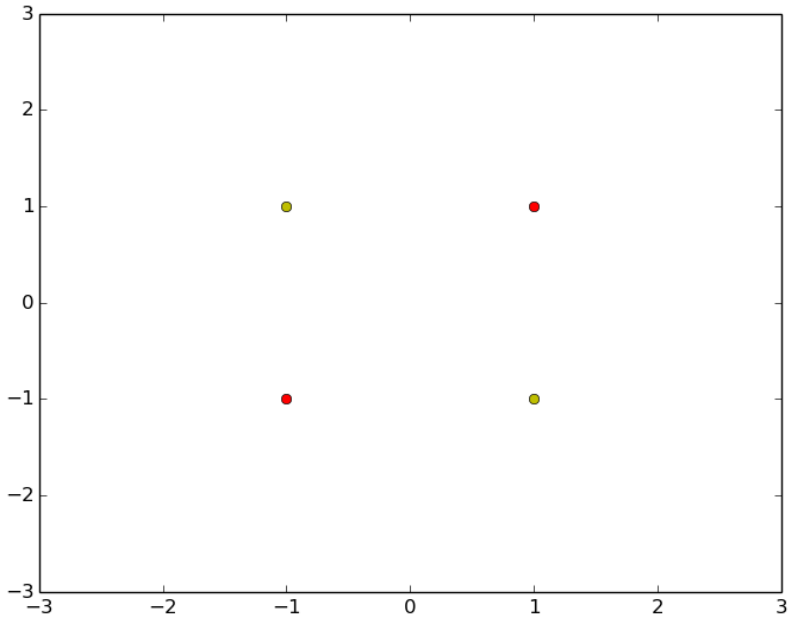

Abbildung 2: Nicht linear separierbare Daten

Hier liegt eine mögliche Fehlerquelle für unsere Klassifikatoren. Was passiert wenn die Daten nicht linear separierbar sind? Es können zwei Arten von Fehlern entstehen:

1. Die Trainingsdaten überschneiden sich (d.h. die SVM kann schon gar nicht richtig lernen.)

2. Die Testdaten überschneiden sich (d.h. die SVM hat zwar richtig gelernt, aber eine Testdaten liegen jetzt näher an der falschen Klasse und werden deshalb auch falsch zugeordnet.)

Haben wir nun mit Hilfe der Trainingsdaten eine passende Hyperebene gefunden, so können wir die Entscheidungsfunktion κ definieren.

Definition 3.9 (Entscheidungsfunktion)**.** Sei $X \in \mathbb{R}^d$ die endliche Menge der Trainingsdaten, $\{-1, +1\}$ die Menge der Klassen und (w, b) eine Hyperebene die anhand der Trainingsdaten gewählt worden ist. Dann können wir eine Funktion $\kappa_X : \mathbb{R}^d \to \{-1, +1\}$ definieren per

$$\kappa_X(x) = sgn(\langle w, x \rangle + b) \ .$$

Eine mögliche Wahl der Hyperebene ist die sogenannte kanonische Hyperebene.

Definition 3.10 (kanonische Hyperebene)**.** Die kanonische Hyperebene bezüglich einer endlichen Menge $X = \{x_1 \dots x_m\}$ ist definiert durch

$$\min_{i=1,\dots,m} (|\langle w, x_i \rangle + b|) = 1 \ .$$

Also die Hyperebene, die dadurch charakterisiert wird, dass der ihr am nächstgelegene Punkt $x \in X$ zu ihr den Abstand 1 besitzt.

3.4 Optimierungsproblem

Es kann jedoch unendlich viele Hyperebenen geben, die zwei linear separierbare Mengen trennen. Die SVM wählen unter all diesen diejenige aus, die erfüllt, dass $\|w\|^2$ minimal ist. Die Nebenbedingung lautet, dass $\| \langle w, x_i \rangle + b\| \geq 1$ für alle $x_i \in X$.

Definition 3.11 (Primales Problem)**.** Sei X eine Menge von Trainingsdaten, die zwei linear separierbare Klassen enthält. Suche die Hyperebene (w^*, b^*), die die beiden Klassen trennt und dabei folgendes erfüllt.

$$w^* = arg \min_{w \in \mathbb{R}^d} \|w\|_2^2$$
$$\forall x \in X : \ |\langle x, w^* \rangle + b^*| \geq 1$$

Dieses Optimierungsproblem wird von der kanonischen Hyperebene gelöst.

Berechnung des Randes Mithilfe der kanonischen Hyperebene können wir nun den Abstand (engl. Margin) der Klassen zur Hyperebene berechnen. Sei x_1 ein Punkt der die Minimumsbedingung $\langle w, x_1 \rangle + b = 1$ erfüllt, wir wollen nun zeigen, dass dieser den Abstand $\frac{1}{\|w\|}$ zur Hyperebene E hat. Dazu sei

$$\begin{aligned} & x_1 - \frac{1}{\|w\|^2} \cdot w \in E \\ \iff & \left\langle w, x_1 - \frac{1}{\|w\|^2} w \right\rangle + b = 0 \\ \iff & \langle w, x_1 \rangle - \left\langle w, \frac{1}{\|w\|^2} w \right\rangle + b = 0 \\ \iff & 1 - 1 = 0. \end{aligned}$$

Also hat die Hyperebene, die durch eine SVM entstanden ist den Abstand $\frac{1}{\|w\|}$ zu den nächsten Punkten aus X (Rand). Da wir im Optimierungsproblem versuchen die Norm möglichst klein zu halten, erhalten wir also einen möglichst breiten Rand. Daher zählen SVMs zu den sogenannten „Large Margin Classifier “. Sind die Trainingsdaten schlecht positioniert, so kann es dazu kommen, dass eine Hyperebene gewählt wird, die nicht optimal für die Testdaten ist. Das wäre ein möglicher Grund für einen Testfehler.

4 Das „Rare Eclipse Problem“

Hier wollen wir das Ergebnis der Abhandlung „Compressive Classification and the Rare Eclipse Problem“ von Dustin Mixon, Alfonso Bandeira und Benjamin Recht [4] vorstellen.

4.1 Allgemeine Überlegungen

„Rare Eclipse Problem“ Seien $A, B \subset \mathbb{R}^d$ zwei disjunkte, geschlossene Mengen und sei $\eta > 0$. Finde das kleinste $k \leq d$, so dass eine Matrix $R \in \mathbb{R}^{k \times d}$ mit zufälligen Einträgen

$$\mathbb{P}\left[RA \cap RB = \emptyset\right] \geq 1 - \eta$$

erfüllt.

Man erkennt den Zusammenhang zwischen diesem Problem und binärer Klassifikation deutlicher, wenn man sich A und B als Klassen vorstellt und bemerkt, das lineare Separierbarkeit eine stärkere Forderung als Disjunktheit darstellt. Die Anforderung, eine Überschneidung der Klassen zu vermeiden ist sinnvoll.

Wir stellen nun fest, dass man, anstelle von $RA \cap RB = \emptyset$ auch

$$Ker(R) \cap (A - B) = \emptyset$$

schreiben kann. Hierbei bezeichnet $A - B$ die Minkowski Differenz, also die Menge

$$\{a - b \,:\, a \in A, b \in B\}.$$

Sicherlich ist $Ker(R)$ unter Skalarmultiplikation abgeschlossen. Daher reicht es anstelle des Schnittes mit $A - B$ nur den Schnitt mit $S \subset (A - B)$, $S = (A - B) \cap S^{d-1}$, wobei S^{d-1} die Einheitssphäre in $\mathbb{R}^d$ bezeichnet, zu betrachten. Insgesamt ergibt sich also

$$RA \cap RB = \emptyset \iff Ker(R) \cap S = \emptyset.$$

Sind die Einträge unserer Matrix R nun $N(0,1)$ verteilt, wie zum Beispiel bei klassischen JLE (siehe: 2.1) so ist der Nullraum ein zufälliger Unterraum der Dimension $d-k$. An dieser Stelle können wir das „Rare Eclipse Problem“ mit einem geometrischen Ansatz lösen. Hierzu benötigen wir die sogenannte Gaußsche Weite und das danach folgende Theorem.

Definition 4.1 (Gaußsche Weite)**.** Sei g ein Vektor mit i.i.d. $N(0,1)$ Einträgen. Und sei $S \subset S^{d-1}$ eine abgeschlossene Teilmenge der Einheitssphäre. Dann ist die Gaußsche Weite $\omega(S)$ gegeben durch

$$\omega(S) = \mathbb{E}_g \left[\sup_{z \in S} \langle z, g \rangle\right] \,.$$

Satz 4.2 („Gordon's Escape through a Mesh Theorem"). *Sei $S \subset \mathbb{R}^d$ eine abgeschlossene Teilmenge der Einheitssphäre, $g \in \mathbb{R}^k$ ein Vektor mit i.i.d. $N(0,1)$ Einträgen. Desweiteren sei $\lambda_k = \mathbb{E}[\|g\|_2]$. Falls $\omega(S) < \lambda_k$ dann gilt für eine Matrix R mit i.i.d. $N(0,1)$ Einträgen folgende Ungleichung*

$$\mathbb{P}\left[Ker(R) \cap S = \emptyset\right] \geq 1 - exp\left(-\frac{1}{2}(\lambda_k - \omega(S))^2\right). \tag{4}$$

Beweis. Der Beweis beruht auf folgender Aussage, welche als Korollar in [8] auftaucht. Wir lassen den Beweis des Korollars aus.

Korollar 4.3 („Gordon's Comparison Theorem"). *Sei S eine abgeschlossene Untermenge von S^{d-1}. Sei R eine $k \times d$ Matrix mit i.i.d. $N(0,1)$ Einträgen. Dann gilt*

$$\mathbb{E}\left[\min_{x \in S} \|Rx\|_2\right] \geq \lambda_k - \omega(S),$$

wobei λ_k und $\omega(S)$ wie eben definiert sind.

Nun können wir Satz 4.2 („Gordon's Escape thorugh a Mesh Theorem") beweisen.

$$\begin{aligned}
&\mathbb{P}\left[Ker(R) \cap S = \emptyset\right] = \mathbb{P}\left[\min_{x \in S} \|Rx\|_2 > 0\right] \\
&= \mathbb{P}\left[\min_{x \in S} \|Rx\|_2 > (\lambda_k - \omega(S)) - (\lambda_k - \omega(S))\right] \\
&\geq \mathbb{P}\left[\min_{x \in S} \|Rx\|_2 > \mathbb{E}\left[\min_{x \in S} \|Rx\|_2\right] - (\lambda_k - \omega(S))\right].
\end{aligned}$$

Da $R \to \min_{x \in S} \|Rx\|_2$ sicherlich Lipschitz mit Konstante 1 ist, können wir die folgende Gleichung aus [14, 8.40] anwenden

$$\mathbb{P}\left[\min_{x \in S} \|Rx\|_2 > \mathbb{E}\left(\min_{x \in S} \|Rx\|_2\right) - t\right] \geq 1 - e^{-t^2/2}.$$

Die Wahl von $t = \lambda_k - \omega(S)$ liefert das gewünschte Resultat. □

An dieser Stelle können wir aus der Formel eine erste Schranke für k extrahieren. Dazu stellen wir zuerst fest, dass $\lambda_k \geq \sqrt{k}$.

$$\lambda_k = \mathbb{E}\left[\|g\|_2\right] = \mathbb{E}\left[\left(\sum_{i=1}^{k} g_i^2\right)^{1/2}\right] \geq \left(\sum_{i=1}^{k} \mathbb{E}\left[g_i^2\right]\right)^{1/2} = \sqrt{k}. \tag{5}$$

Im dritten Schritt haben wir die Jensensche Ungleichung (Satz 1.5) benutzt. (Wir tauschen zu erst den Erwartungswert mit der Wurzel und können dann über die Unabhängigkeit der g_i den Erwartungswert auch in die Summe ziehen). Dann nutzen wir die Kenntnis der Momente. Nun setzen wir die erhaltene Ungleichung (5) in (4) ein und erhalten so die gewünschte Abschätzung.

Lemma 4.4. *Seien $A, B \subset \mathbb{R}^d$ zwei disjunkte abgeschlossene und konvexe Mengen, $\omega(S)$ die Gaußsche Weite von $S = (A - B) \cap S^{d-1}$ und R eine Matrix mit $N(0,1)$ Einträgen. Dann gilt*

$$k \geq \left(\omega(S) + \sqrt{2\log(1/\eta)}\right)^2 \Longrightarrow \mathbb{P}\left[RA \cap RB = \emptyset\right] = 1 - \eta.$$

Beweis. Aus (5) wissen wir, dass

$$\lambda_k \geq \sqrt{k}.$$

Auf der anderen Seite liefert (4), dass

$$\mathbb{P}\left[RA \cap RB \neq \emptyset\right] \leq \exp\left(-\frac{1}{2}\left(\lambda_k - \omega(S)\right)^2\right) \leq \eta.$$

Zusammen ergibt sich dann

$$\begin{aligned} &\eta \geq \exp\left(-\frac{1}{2}\left(\sqrt{k} - \omega(S)\right)\right) \\ &\Rightarrow 2\log(1/\eta) \leq \left(\sqrt{k} - \omega(S)^2\right) \\ &\Rightarrow \sqrt{2\log(1/\eta)} \leq \sqrt{k} - \omega(S) \\ &\Rightarrow k \geq \left(\omega(S) + \sqrt{2\log(1/\eta)}\right)^2 \quad , \end{aligned} \tag{6}$$

womit das geforderte bewiesen wäre. □

Jetzt wollen wir für den konkreten Fall der zwei Bälle die Gaußsche Weite $\omega(S)^2$ berechnen.

4.2 Zwei disjunkte Bälle

In diesem Abschnitt ist $B_1 = B(c_1, r_1)$ und $B_2 = B(c_2, r_2)$. Wir beginnen damit die Minkowski Differenz $B_1 - B_2$ zu berechnen.

Lemma 4.5. *Sei $B_1 = \{x \in \mathbb{R}^d | \, \|x - c_1\|_2 \leq r_i\}$, $B_2 = \{x \in \mathbb{R}^d | \, \|x - c_2\|_2 \leq r_2\}$, so dass $r_1 + r_2 < \|c_1 - c_2\|_2$ gilt. Dann hat die Minkowski Differenz folgende Gestalt*

$$B_1 - B_2 = Circ(\alpha) = \{z : \langle z, c_1 - c_2\rangle \geq \|z\|_2 \|c_1 - c_2\|_2 \cos(\alpha)\} \tag{7}$$

wobei $\alpha \in (0, \pi/2)$ der Winkel ist, der $\sin(\alpha) = \frac{r_1 + r_2}{\|c_1 - c_2\|}$ erfüllt.

Beweis. Sei $C_- \in \mathbb{R}^d = B_1 - B_2$. Wir wollen zeigen $C_- = Circ(\alpha)$.
Schritt 1: $C_- \subset Circ(\alpha)$.

Wir suchen das kleinste α, so dass C_- in $Circ(\alpha)$ liegt. Das ist äquivalent zu

$$\cos(\alpha) \leq \inf_{z \in C_-} \frac{\langle z, c_1 - c_2 \rangle}{\|z\| \|c_1 - c_2\|}$$

In dem wir $\delta = \|c_1 - c_2\|_2$ setzen ergibt sich $S_1 - S_2 = (r_1 + r_2)B(0,1) + \delta$ und das vorige Minimierungsproblem lässt sich mit

$$f(y) = \frac{\langle y + \delta, \delta \rangle}{\|y + \delta\| \|\delta\|}$$

umformen zu

$$\min \ f(y)$$

unter Beachtung von $\|y\|_2 \leq r_1 + r_2$. Nun zeigen wir, dass $f(y)$ auf dem Rand minimiert wird. Das heißt, dass für den Minimierer y gilt: $\|y\|_2 = r_1 + r_2$. Sei dazu P_{orth} die Projektion auf das orthogonale Komplement von $span(\delta)$. Desweiteren sei $\|y\|_2 < r_1 + r_2$ und $t > 0$ so dass $\|y + tP_{orth}y\|_2 = r_1 + r_2$. Dann ergibt sich

$$f\left(y + tP_{orth}y\right)) = \frac{\langle y + tP_{orth}y + \delta, \delta \rangle}{\|y + tP_{orth}y + \delta\|_2 \|\delta\|_2} = \frac{\langle y + \delta, \delta \rangle}{\|y + tP_{orth}y + \delta\|_2 \|\delta\|_2}$$
$$< \frac{\langle y + \delta, \delta \rangle}{\|y + \delta\|_2 \|\delta\|_2} = f(y).$$

Also genügt es den Minimierer unter $\|y\|_2 = r_1 + r_2$ zu suchen. Wir verwenden jetzt Lagrange Multiplikatoren um einen Ausdruck für $\nabla f(y)$ zu finden. Sicher gilt für $g(y) = \|y\|_2^2$ dass der Gradient der Funktion $\nabla g(y) = 2y$ beträgt. Und mit einem Lagrange Multiplikator $\lambda \in \mathbb{R}$ folgt, dass der Minimierer

$$\nabla f(y) = -2\lambda g(y),$$

erfüllt. Für $\nabla f(y)$ gilt

$$\nabla f(y) = \frac{1}{\|y + \delta\|_2^2} \left(\delta - \left\langle \frac{y + \delta}{\|y + \delta\|_2}, \delta \right\rangle \frac{y + \delta}{\|y + \delta\|_2} \right).$$

Man sieht, dass $\nabla f(y) = 0$ genau dann, wenn $y = k \cdot \delta$ für ein $k > 0$. Also wenn y die Funktion f maximiert. Das gilt, da wir mit der Cauchy-Schwarz Ungleichung angewandt auf f folgendes erhalten

$$\langle y + \delta, \delta \rangle \leq \|y + \delta\| \cdot \|\delta\|$$
$$\implies \frac{\langle y + \delta, \delta \rangle}{\|y + \delta\| \cdot \|\delta\|} \leq 1.$$

Die Gleichheit ist erfüllt, falls $y = k \cdot \delta$.
Des weiteren ist $\langle y + \delta, \nabla f(y) \rangle = 0$ für alle y. Insgesamt haben wir die folgenden drei Aussagen erhalten

1. $\nabla f(y^*) = -2\lambda y^*$ für $\lambda \in \mathbb{R}$,
2. $\nabla f(y^*) \neq 0$,
3. $\langle y^* + \delta, \nabla f(y^*) \rangle = 0$.

Aus 1. und 2. folgern wir, dass $\nabla f(y^*)$ ein vielfaches von y^* ist. Mit 3. ergibt sich dann, $\langle y^* + \delta, y^* \rangle = 0$. Das bedeutet, dass der Ursprung, y^* und $y^* + \delta$ ein rechtwinkliges Dreieck bilden. Die Länge der Hypotenuse beträgt $\|\delta\|$. Der kleinste Winkel α der die Forderung erfüllt, ist der zwischen $y^* + \delta$ und y^*. Es gilt $\sin(\alpha) = \frac{\|y^*\|}{\|\delta\|} = \frac{r_1 + r_2}{\|c_1 - c_2\|}$. Womit gezeigt wäre, dass $C_- \subset Circ(\alpha)$. Es bleibt zu zeigen, dass auch die andere Teilmengenrelation erfüllt ist.
Schritt 2: $Circ(\alpha) \subset C_-$ für α wie oben.
Sei dazu :

$$G = \left\{ z \in \mathbb{R}^d \, : \, \langle z, \delta \rangle = \|z\| \|d\| cos(\alpha), \, \|z\| = \|y^* + \delta\| \right\}.$$

Es reicht dann zu zeigen, dass $G \subset C_-$. Sei $z \in G$ beliebig. Wir betrachten das Dreieck mit den Eckpunkten 0, δ und z. Dann ist per Definition der Winkel zwischen δ und z α und $\|z\| = \|y^* + \delta\|$. Das Dreieck ist kongruent zu dem mit Ecken 0, δ und $y^* + \delta$. Daraus folgt, das die Seite zwischen z und δ die Länge $\|y^*\| = r_1 + r_2$ hat. Wir erhalten, dass $z = (z - \delta) + \delta$ in der Differenz $B_1 - B_2$ liegt. □

Jetzt haben wir gezeigt, dass die Minkowski-Differenz $B_1 - B_2 = Circ(\alpha)$ ist. Als nächstes benötigen wir die Gaußsche Weite des Schnittes mit der Einheitssphäre $\omega\left(Circ(\alpha) \cap S_{d-1}\right)$. Diese wurde in [3] berechnet.

4.2.1 Berechnung der Gaußschen Weite

Satz 4.6 (Gaußsche Weite des Kegels)**.** *Sei $Circ(\alpha)$ wie in (7). Sei $S = Circ(\alpha) \cap S^{d-1}$ der Schnitt mit der Einheitssphäre. Dann gilt für die Gaußsche Weite $\omega(S)$ folgendes*

$$\omega(S)^2 = \sin(\alpha)^2 \cdot d + O(1).$$

Beweisskizze. Wir beginnen damit, dass wir die Erkenntnis aus [3] nutzen, die uns erlaubt $\omega(S)$ umzuformulieren zu

$$\omega(S)^2 = \mathbb{E}_{S^{d-1}} \left[\Pi_S(v) \right] \ ,$$

wobei $\Pi_S(v) = arg\min\{\|x - y\|_2 : y \in S\}$ die euklidische Projektion auf den Kegel ist. Im folgenden sei θ das Bild eines Einheitsvektors v auf den Kegel und $\beta = \arccos(\theta_1)$ bezeichne den Winkel zwischen θ und e_1. Weiter definieren wir

$$F(\beta) = \|\Pi_S(v)\|^2 = \begin{cases} 1 & 0 \leq \beta \leq \alpha \\ cos(\beta - \alpha)^2 & \alpha \leq \beta \leq \pi/2 + \alpha \\ 0 & \pi/2 + \alpha \leq \beta \leq \pi. \end{cases}$$

An dieser Stelle erlaubt uns der Transformationssatz [15, 6.5.1.], das wir anstelle des Integrals über Π_S ein Integral über F benutzen können. Dieses hat die Gestalt

$$\omega(S)^2 = d \cdot \frac{\Gamma\left(\frac{d}{2}\right)}{\sqrt{\pi} \cdot \Gamma\left(\frac{d-1}{2}\right)} \int_0^\pi \sin^{d-2}(\beta) F(\beta) d\beta. \tag{8}$$

Jetzt berechnen wir das Integral und schätzt dann den Bruch davor ab. Das Integral können wir mit [1, 6.2.3.] ausrechnen und erhalten

$$\int_0^\pi \sin^{d-2}(\beta) F(\beta) d\beta = \sqrt{\frac{2\pi}{d}} F\left(\frac{\pi}{2}\right) + O\left(d^{-3/2}\right).$$

Den Bruch vor dem Integral behandeln wir mit [13, 5.6.4.] und erhalten

$$\begin{aligned} &\sqrt{d-2} < \frac{\sqrt{2} \cdot \Gamma\left(\frac{d}{2}\right)}{\Gamma\left(\frac{d-1}{2}\right)} < \sqrt{d} \\ &\Rightarrow \frac{\Gamma\left(\frac{d}{2}\right)}{\sqrt{\pi} \cdot \Gamma\left(\frac{d-1}{2}\right)} < \sqrt{\frac{d}{2\pi}}\ . \end{aligned}$$

Setzen wir beides in (8) ein, so entsteht

$$\begin{aligned} \omega(S)^2 &= dF\left(\frac{\pi}{2}\right) + d \cdot \sqrt{\frac{d}{2\pi}} \cdot O\left(d^{-3/2}\right) \\ &= dF\left(\frac{\pi}{2}\right) + O(1) \\ &= d\left(\cos^2\left(\frac{\pi}{2} - \alpha\right)\right) + O(1) \\ &= d \cdot \sin^2(\alpha) + O(1). \end{aligned}$$

Für einen detailierten Beweis siehe [3]. □

Wir wissen jetzt, dass für zwei Bälle die Gaußsche Weite

$$\omega(S)^2 = \sin(\alpha)^2 \cdot d$$

beträgt. Aus (7) wissen wir, dass $\sin(\alpha) = r_1 + r_2 / \|c_1 - c_2\|_2$. Also ergibt sich insgesamt

$$\omega(S)^2 = \left(\frac{r_1 + r_2}{\|c_1 - c_2\|_2}\right)^2 \cdot d + O(1). \tag{9}$$

5 Separierung Gaußscher Daten

Einen anderer Ansatz das Problem zu betrachten, liefert die Abhandlung [5] von Dasgupta. In [4] wird davon ausgegangen, dass die Daten in zwei disjunkten Bällen liegen. Nun betrachten wir zwei mehrdimensionale Gaußverteilungen mit unterschiedlichen Erwartungswerten und Varianzen. Da wir unsere Daten selber generieren, können wir die Varianz so wählen, dass die Daten mit hoher Wahrscheinlichkeit in Bällen liegen. Das heißt wir wählen Gaußverteilungen $N(\mu_1, \sigma^2 I_d)$ und $N(\mu_2, \sigma^2 I_d)$. Dabei ist σ so gewählt, dass außerhalb des Balls mit Radius r um den Erwartungswert, nur sehr selten Daten liegen. Es entsteht eine Ausgangssituation ähnlich der in Abschnitt 4. Analog zum Abstand der Bälle definieren wir nun die Separiertheit der Verteilungen.

Definition 5.1 (c-Separiertheit)**.** Seinen $N(\mu_1, \Sigma_1)$, $N(\mu_2, \Sigma_2)$ zwei Gaußverteilungen in $\mathbb{R}^d$. $\lambda_{max}(\Sigma_i)$ und $\lambda_{min}(\Sigma_i)$ beschreiben den größten beziehungsweise den kleinsten Eigenwert der Kovarianzmatrizen. Wir sagen, dass die Gaußverteilungen c-separiert sind, falls

$$\|\mu_1 - \mu_2\| \geq c\sqrt{d \max(\lambda_{max}(\Sigma_1), \lambda_{max}(\Sigma_2))}. \tag{10}$$

Wenn die beiden Verteilungen die Varianz $\Sigma_1 = \Sigma_2 = \sigma^2 I_d$ besitzen, so bedeutet (10)

$$\|\mu_1 - \mu_2\| \geq c\sigma\sqrt{d}.$$

Eine Menge von n Gaußverteilungen heißt c-separiert, wenn alle Verteilungen paarweise c-separiert sind.

Die Konstante c ist ein Maß dafür, wie stark sich die Verteilungen überschneiden, ist c sehr groß, so sind viele σ Umgebungen noch voneinander getrennt. Das folgende Lemma liefert eine Aussage über den Erhalt der c-Separiertheit nach zufälliger Projektion in einen Raum kleinerer Dimension.

Lemma 5.2 (Dimensionsreduktion)**.** *Seien $\{(\mu_i, \Sigma_i)\}$ n verschiedene c-separierte Gaußverteilungen für ein $c > 0$. Dann sind, nach Projektion (per JLE) in einen Raum der Dimension k, mit Wahrscheinlichkeit $1 - \eta$, die Gaußverteilungen $c\sqrt{1-\varepsilon}$-separiert. Falls gilt, dass*

$$k \geq \frac{4}{\varepsilon^2} \log\left(\frac{n^2}{\eta}\right). \tag{11}$$

Seien $\{(\mu_i^*, \Sigma_i^*)\}$ die n Gaußverteilungen die wir nach Anwendung der JLE erhalten. Dann gilt desweiteren

$$\lambda_{max}(\Sigma_i^*) \leq \lambda_{max}(\Sigma_i)$$
$$\lambda_{min}(\Sigma_i^*) \geq \lambda_{min}(\Sigma_i).$$

Das bedeutet, das die Ausdehnung der Daten in gewisser Weise gleichmäßiger wird. Insbesondere gilt, wenn

$$\lambda_{max}(\Sigma_i) = \lambda_{min}(\Sigma_i) = \sigma^2 \ ,$$

so auch

$$\lambda_{max}(\Sigma_i^*) = \lambda_{min}(\Sigma_i^*) = \sigma^2 \ .$$

Für unseren Fall folgt, dass die Daten auch nach Projektion noch in Bällen liegen. Die Separiertheit wurde allerdings verringert. In [5] finden sich auch noch allgemeinere Aussagen über die Varianz und die Ekzentrizitätsverringerung. Für unseren Fall, in dem auch die Gaußschen Daten in zwei Bällen liegen sollen, reicht Lemma 5.2 aus.

6 Beschreibung der Experimente

6.1 Einführung und Idee

Nehmen wir die Daten $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$ als gegeben an. Wir wollen diese nach $\mathbb{R}^k$ projizieren und die Erfolgsrate $\mathcal{S}_d$ der Klassifikation in $\mathbb{R}^d$ mit der Erfolgsrate $\mathcal{S}_k$ in $\mathbb{R}^k$ vergleichen. Insbesondere suchen wir eine untere Schranke k_0 für k, so dass die Erfolgsrate in allen euklidischen Räumen, der Dimension $k \geq k_0$ noch mindestens 95% der ursprünglichen Erfolgsrate beträgt. Die theoretischen Ansätze stammen aus den Abschnitten 4 und 5.

6.2 Mathematischer Rahmen

6.2.1 Erzeugung der Daten

Wir benötigen zufällige Daten in zwei getrennten Bällen. Dies können wir auf zwei Arten tun. Beide Wege führen zu den gleichen Ergebnissen in den Experimenten. Wir stellen sie jedoch beide vor, damit der Zusammenhang zu den Abschnitten 5 und 4 klar wird.

Möglichkeit 1 Zuerst lassen wir uns per d-dimensionaler Normalverteilung (Definition 1.3) Daten ausgeben. Dann normieren wir die zufälligen Vektoren, dies führt dazu, dass alle Vektoren in S^{d-1} liegen. Dann strecken wir die Vektoren um eine Zahl zwischen 0 und dem Radius. Zu letzt verschieben wir sie in der ersten Komponente, so dass für die beiden Zentren c_1 und c_2 der Bälle B_1 und B_2 gilt

$$\|c_1 - c_2\|_2 = 2 \cdot r + \delta \ ,$$

wobei δ der Abstand der Bälle ist.

Möglichkeit 2 Hierbei benutzen wir lediglich die d-dimensionale Normalverteilung (Definition 1.3). Wählen als Mittelwerte die Zentren der Bälle und als Kovarianzmatrix die Matrix $\sigma^2 I_d$, wobei I_d die Einheitsmatrix in $\mathbb{R}^{d \times d}$ und $\sigma^2 = 1/(7rd)$ ist.

Der Vorteil der ersten Methode ist, dass wir garantieren können, das die Daten in den Bällen liegen. Im zweiten Fall hingegen liegen mit sehr geringer Wahrscheinlichkeit einige Daten außerhalb der Bälle. Da die erste Methode jedoch mehr Rechenzeit benötigt, gehen wir dieses Risiko in manchen Fällen ein. Man beachte, dass die Daten unabhängig voneinander generiert werden. Also die Einträge jedes Vektors $x_i \in X$ sind untereinander und zu denen der anderen Vektoren unabhängig. Die Klassenzugehörigkeiten können wir über die Bälle definieren. Liegt ein Punkt x in B_1 so erhält er die Klasse -1, liegt er dagegen in B_2 so gehört er zu Klasse $+1$.

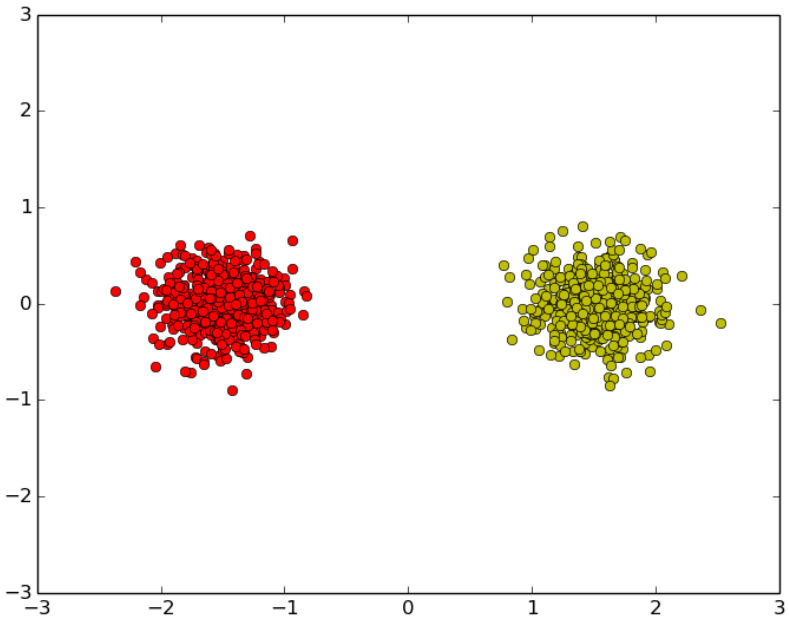

Abbildung 3: Unsere Daten

6.2.2 Grundlegender experimenteller Ansatz

Das Maß für die Güte unserer Klassifikatoren wird der empirische Testfehler (Definiton 3.4) sein. Sei dazu κ die Entscheidungsfunktion (Definition 3.9), die durch das Training entstanden ist. Die Testmenge sei $X = \{x_1, \ldots, x_n\}$. Da wir wissen welche Daten in welchen Bällen liegen, kennen wir auch die echten Klassen y_i. Dann können wir uns eine Null-Eins Verlustfunktion (Definition 3.2) definieren.

$$l(x_i) = \begin{cases} 0 & \text{,wenn } \kappa(x_i) = y_i \\ 1 & \text{,wenn } \kappa(x_i) \neq y_i \end{cases} .$$

Jetzt können wir den empirischen Testfehler $R_{emp,X}$ berechnen, also den empirischen Fehler auf X. Dieser beträgt

$$R_{emp,X} = \frac{1}{n} \sum_{i=1}^{n} l(x_i) .$$

Die Erfolgsrate $\mathcal{S}$ ergibt sich aus $\mathcal{S} = 1 - R_{emp,X}$. Wir wählen die Erfolgsrate als Gütemaß unseres Klassifikators. Im folgenden wird $\mathcal{S}_d$ die Erfolgsrate des Klassifikators im Ursprungsraum $\mathbb{R}^d$ bezeichnen. $\mathcal{S}_k$ ist die Erfolgsrate, die unsere lineare Stützvektormaschine (Definition 3.9), nach Anwendung einer JLE (Satz 2.1) in $\mathbb{R}^k$, noch erbringt.

Definition 6.1 ($\mathcal{S}_k$)**.** Sei $X_-, X_+ \subset X$ die Menge der Daten mit Klassen -1 beziehungsweise $+1$ und R eine JLE. Dann sind RX_- und RX_+ die Mengen der Daten mit Klassen -1 und $+1$ nach Projektion. Desweiteren sei κ_k die Entscheidungsfunktion der SVM die wir nach der Projektion anlernen. Mit dieser können wir uns wieder eine Null-Eins Verlustfunktion l_k definieren. Dann berechnen wir den empirischen Testfehler $R_{emp,RX}$ und erhalten die Erfolgsrate durch

$$\mathcal{S}_k = 1 - R_{emp,RX}.$$

Da die Abbildung R aus Satz 2.2 zufällige Einträge besitzt, müssen wir mehrmals projizieren und klassifizieren. Wir bilden dann den Mittelwert der erhaltenen $\mathcal{S}_k$. Das bedeutet wir betrachten eine Näherung an $\mathbb{E}_R[\mathcal{S}_k]$.

Gesucht ist dann das kleinste $k = k_0$, so dass

$$\frac{\mathbb{E}_R\left[\mathcal{S}_k\right]}{\mathcal{S}_d} > 0,95\ . \tag{12}$$

Diese untere Schranke kann von den Parametern d und δ beziehungsweise δ und r abhängen.

Solange wir zu Beginn eines Experimentes nichts anderes erwähnen, wählen wir die Parameter folgender Weise:

Ursprungsdimension $d =$	100
Anzahl der Trainingsdaten $p =$	10
Anzahl der Testdaten $n =$	1000
Radius der Bälle $r =$	1
Abstand der Bälle $\delta =$	1
Anzahl der Wiederholungen von Projektion und Klassifikation	2000

Für Klassifikation und Projektion benutzen wir die Funktionen aus [16].

6.3 Ursprungsdimension fest und Abstand variabel

6.3.1 Experiment 1.1

Wir ermitteln die Erfolgsrate $\mathcal{S}_d$. Dann erfolgt die Projektion und die Annäherung an die Erwartungswerte $\mathbb{E}_R[\mathcal{S}_k]$. Wir tragen dann das Verhältnis $\mathbb{E}_R[\mathcal{S}_k]/\mathcal{S}_d$ gegen die Dimension k auf.

Thesen

These 6.2 (Vermutung nach Bandeira et al)**.** *Seien $B(c_1, r_1)$ und $B(c_2, r_2)$ die zwei Bälle in $\mathbb{R}^d$, die unsere Daten enthalten. Des weiteren sei $\mathcal{S}_d$ die*

Erfolgsrate der linearen SVM in $\mathbb{R}^d$. Wähle $\varepsilon \in (0,1)$ fest. Dann gilt

$$k_0 = O\left(\frac{(r_1+r_2)^2}{\|c_1-c_2\|_2^2}d\right) \ .$$

Für den Fall $r_1 = r_2 = r$ und $\|c_1 - c_2\|_2 = 2r + \delta$ ergibt sich

$$k_0 = O\left(\frac{4r^2}{(2r+\delta)^2}d\right) \ . \tag{13}$$

Wir erhalten diese Vermutung, wenn wir (9) in (6) einsetzen.

These 6.3 (Vermutung nach Dasgupta)**.** *Seien $(\mu_1, \sigma^2 I_d)$ und $(\mu_2, \sigma^2 I_d)$ zwei d-dimensionale Normalverteilungen (Definition 1.3), die c-separiert (Definition 5.1) sind. Mit der Wahl $\sigma = 1/\sqrt{7rd}$ verhindern wir, dass c von d abhängt. Für c ergibt sich nach Definition 5.1 lediglich*

$$c \leq (2r+\delta)\sqrt{7r} \ .$$

Das ist für jede Wahl von delta noch sehr groß. Gehen wir davon aus, dass zufällige Dimensionsreduktion die c-Separiertheit der Daten verringert (was auch durch Lemma 5.2 nahegelegt wird) so müssen wir feststellen, das δ keine Rolle spielt. Wir gehen also davon aus, dass für ein $m \in \mathbb{N}$

$$k_0(\delta, d) = m \ .$$

Als Beispiel werden wir einigen Werten von c einige Werte von $\frac{2r}{2r+\delta}$ gegenüberstellen.

δ	c	$(2r)/(2r+\delta)$
1	$3\sqrt{7}$	2/3
0,5	$2,5\sqrt{7}$	4/5
0	$2\sqrt{7}$	1

Wie man sehen kann, ist die c-Separiertheit stets sehr groß, wohingegen sich der Bruch der 1 nähert. Das heißt, dass in der zweiten Formel k_0 weitestgehend konstant bleiben dürfte.

Durchführung Wir halten d, r, n und p fest. Für jeden in der Graphik angegebenen Werte von δ reduzieren wir k bis auf 1. Zuerst benutzen wir klassische JLE. Also Abbildungsmatrizen der Einträge $N(0,1)$ verteilt sind.

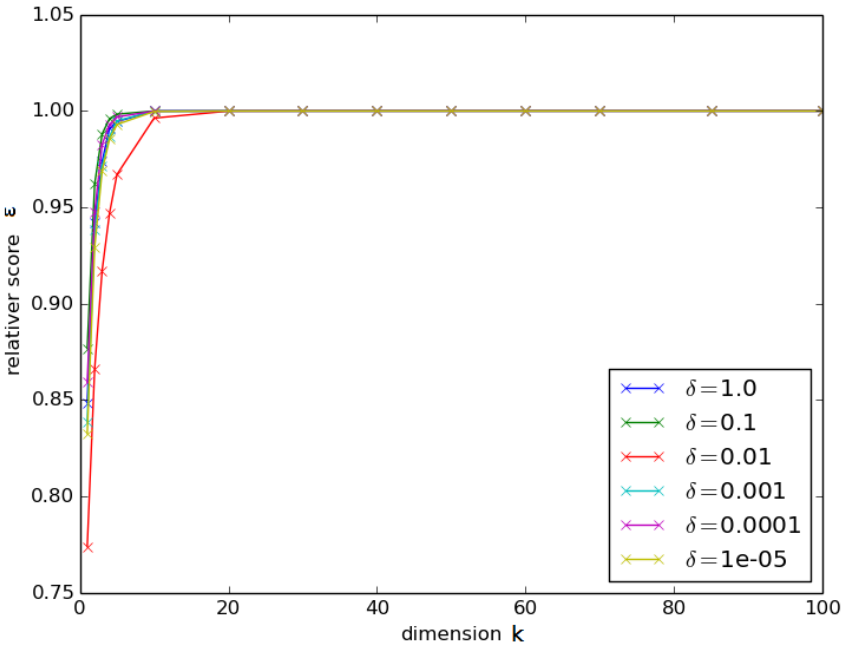

Abbildung 4: Daten die mit klassischer JLE projiziert wurden

Nun vergleichen wir dieses Ergebnis mit dem entsprechenden für Achlioptas Matrizen (1). Man beachte, dass wir die Standardeinstellung von [16] benutzen. Damit erhalten wir im Mittel mehr Nulleinträge als bei (1). Deshalb kommt es zu schlechteren Ergebnissen als bei den klassischen JLE. Allerdings sparen wir Rechenzeit und die beobachteten Effekte sind die selben, daher werden wir auch weiterhin Matrizen benutzen, deren Einträge die Verteilung (1) besitzen.

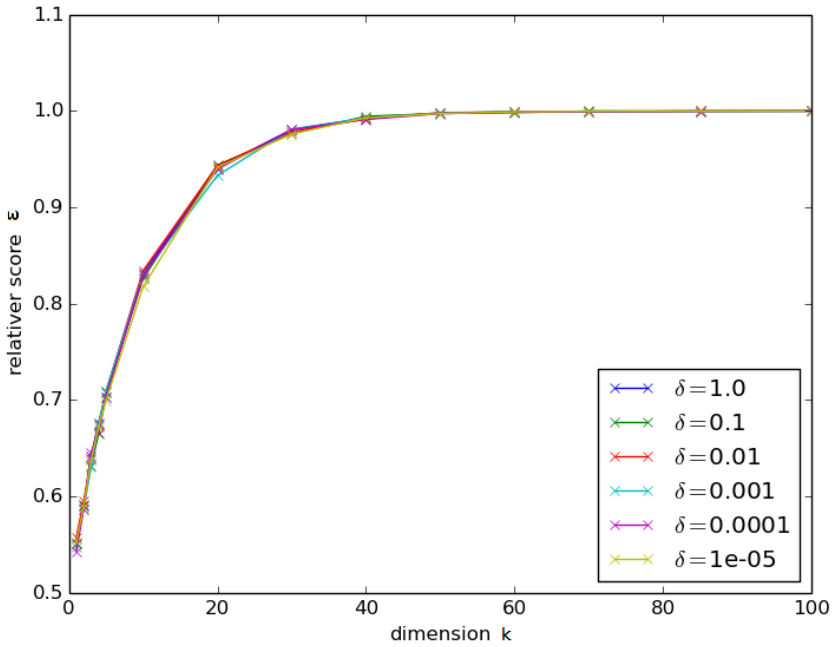

Abbildung 5: Ergebnisse des Experimentes

Auswertung Anscheinend gilt, dass

$$k_0(\delta) = m.$$

Das stimmt nicht mit der ersten These 6.2 überein. Allerdings, kann man sich dieses Verhalten mit 6.3 erklären. Da die c-Separiertheit (5.1) unserer Daten sehr groß ist. Man kann auch erkennen, dass k_0 von ε abhängt. Auch in [4] werden numerische Experimente durchgeführt. Diese offenbaren aber nicht die Ergebnisse des obigen Experimentes. Daher wollen wir die Abhängigkeit $k_0(\delta)$ noch einmal überprüfen.

6.3.2 Experiment 1.2

Diesmal wollen wir uns nur auf das in Experiment 1.1 beobachtete Phänomen konzentrieren. Das bedeutet, dass wir überprüfen wollen, ob k_0 vom Abstand der Bälle δ beziehungsweise von dem Verhältnis $2r/(2r+\delta)$ abhängt. Wir untersuchen diesmal aber nicht die Funktion $\varepsilon_\delta(k)$, sondern die Funktion $k_0(\delta)$ aus (12) für festes d. Für jeden Wert von δ verringern die Dimension mittels Intervallschachtelung. Die Entscheidung ob die neue Dimension k obere oder untere Schranke wird, treffen wir aufgrund der mittleren Erfolgsrate der anschließenden Klassifikation. Ist sie größer als 95% der ursprünglichen

Erfolgsrate, so haben wir eine neue obere Schranke, andernfalls eine untere Schranke.

Thesen Nach Experiment 1.1 gehen wir davon aus, das $k_0(\delta)$ eine konstante Funktion ist. Da wir in (12) die Parameter d und ε festhalten.

Durchführung Wir halten alle Parameter fest und variieren δ von 1 auf $0,000001$.

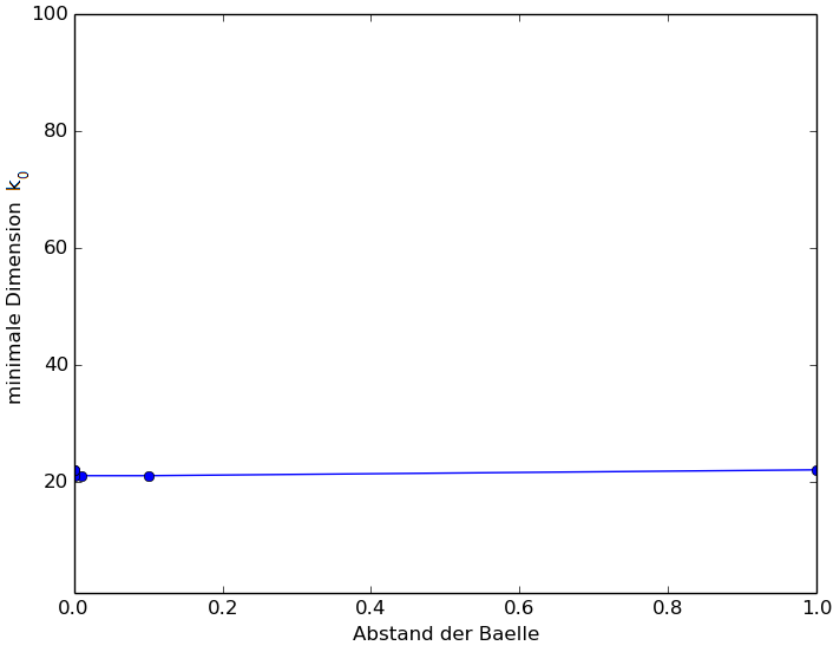

Abbildung 6: Ergebnisse des zweiten Experiments

Auswertung Die kleinstmögliche Dimension k_0 hängt nicht vom Abstand der Bälle, δ ab.
Obwohl die Aussage, dass das Verhältnis von Volumen zu Abstand eine Rolle spielt, sehr sinnvoll scheint. Die Argumente aus Abschnitt 5 können dieses Verhalten besser erklären. Da die c-Separiertheit in jedem Fall sehr groß ist, könnte es sein, das der Abstand der Bälle nicht sehr wichtig ist. Bevor wir zum zweiten Experiment kommen, möchten wir noch einmal rekapitulieren, was wir bis jetzt festgestellt haben.
k_0 scheint **nicht** von dem Verhältnis von Radius zu Abstand der Bälle abzuhängen. Sicherlich hängt sie aber von der vorgegeben Güte ε ab. Die Abhängigkeit vom Radius geht mir der des Abstands einher, da nur das Verhältnis

eine Rolle spielt. Wir wissen also bis jetzt, dass für ε und d fest gewählt, die Funktion $k_0(\delta)$ eine **konstante** Funktion ist.

6.4 Ursprungsdimension variabel und Abstand fest

6.4.1 Experiment 2

Im zweiten Experiment werden wir die Abhängigkeit von der Ursprungsdimension d testen. Sowohl Abschnitt 4 als auch Abschnitt 5 legen unterschiedliche Sichtweisen auf dieses Problem nahe. Irgendeine Form der Abhängigkeit könnte es geben. Eine hohe Dimension des Ursprungsraums bedeutet ja auch, dass die Daten viele Eigenschaften haben. Könnten wir immer in die gleiche Dimension k_0 projizieren, wäre das eine sehr starke Aussage. Wir halten den Radius, den Abstand und die erforderliche Güte fest. Dann variieren wir die Ursprungsdimension d. Für jedes d suchen wir per Intervallschachtelung wie in Experiment 1.2 die Dimension k_0, so dass die Erfolgsrate der Klassifikation noch 95% der ursprünglichen Erfolgsrate beträgt.

Thesen Aus (11) könnte man schließen, dass es keine Abhängigkeit von der Dimension d gibt. Der Abschnitt 4 und insbesondere (9) legen nahe, dass k_0 linear von der ursprünglichen Dimension abhängt. Also aus den beiden Abhandlungen [4] und [5] könnten wir zwei Formen für $k_0(d)$ erahnen. Nutzen wir 6.2, so erhalten wir

$$k_0(d) = a \cdot d + b \ ,$$

für konstante Steigung a und eine Konstante $b \in \mathbb{R}$. Aus 6.3 können wir hingegen

$$k_0(d) = m$$

für konstantes $m \in \mathbb{N}$ schließen.

Durchführung Wir halten alle Parameter fest und variieren d von 100 zu 10000. Wir beginnen damit, k_0/d gegen d aufzutragen.

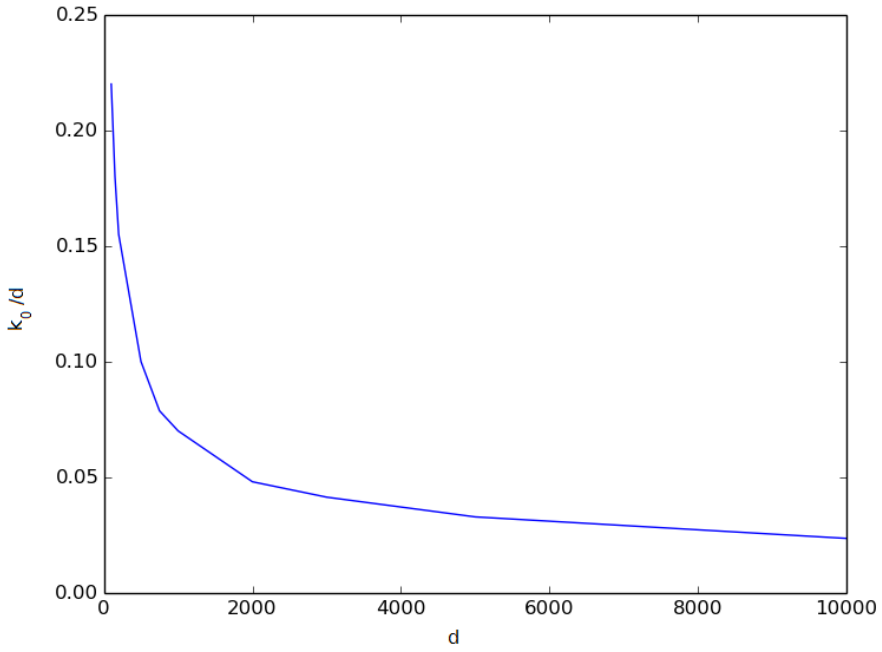

Abbildung 7: Verhältnis relativ

Man erkennt direkt, das es einen Zusammenhang zwischen Ursprungsdimension d und der kleinstmöglichen Dimension k_0 zu geben scheint. Je größer d desto kleiner ist das Verhältnis k_0/d.

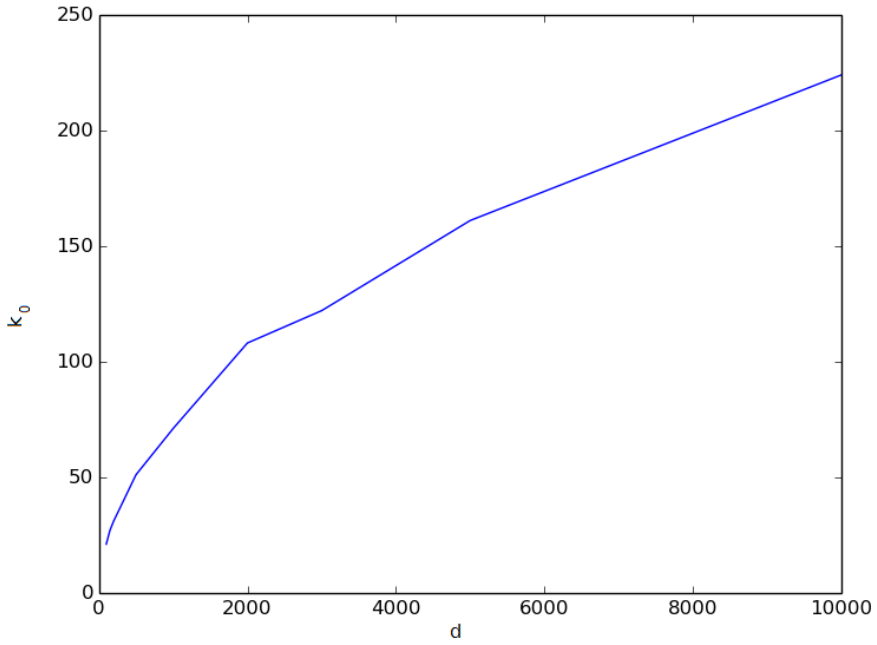

Abbildung 8: Absolutes Verhältnis

Auswertung Wenn wir die Abschätzung aus (9) auf unseren Ansatz übertragen würden, so würden wir erwarten, dass $k_0(d) = a \cdot d + b$ für eine Konstante $b \in \mathbb{R}$ und eine Steigung $a > 0$. Aus (11) könnten wir nur $k_0(d) = m$, für ein $m > 0$, herleiten. In unseren Experimenten passiert jedoch etwas anderes. Es scheint so als ob das tatsächliche Verhalten dazwischen liegt

$$k_0(d) = O\left(d^p\right) \ ,$$

für ein $p \in (0, 1)$. Um dieses Ergebnis abschließend zu belegen müsste man $\log k_0$ gegen $\log d$ auftragen. Würde man eine Gerade mit der Steigung γ erhalten, so hätte man die Vermutung bestärkt.

7 Abschließende Zusammenfassung

Wir haben zwei Experimente durchgeführt, die unser k_0 aus (12) auf die angegebenen Abhängigkeiten überprüfen sollte. Für jede dieser Abhängigkeiten gab es basierend auf den Arbeiten [4],[5] und [12] Gründe, die dafür oder dagegen sprachen. Zu Beginn formulierten wir allgemein und naiv, dass für alle $k \geq k_0(d, r, \delta, \varepsilon)$

$$\mathbb{E}\left[\mathcal{S}_k\right] > \varepsilon \cdot \mathcal{S}_d \ ,$$

gelten sollte.
Wir konnten für den Fall der Daten, die in Bällen liegen jedoch folgende Formulierung erlangen

$$\mathbb{E}\left[\mathcal{S}_k\right] > \varepsilon \cdot \mathcal{S}_d,$$

für alle $k \geq k_0(d, \varepsilon)$. Da wir nach den Experimenten 1.1 und 1.2 die Abhängigkeit von r und δ aufgeben konnten. Experiment 2 diente dazu, die Abhängigkeit k_0 von d zu quantifizieren. Es stellte sich heraus, dass für festes ε eine Relation der folgenden Art besteht:

$$k_0(d) = O(d^\gamma) \ , \ \gamma \in (0, 1).$$

Keine der beiden Abhandlungen [4] und [5] lieferten theoretische Grundlagen um diese Erkenntnis zu fundieren.

8 Literaturverzeichnis

Literatur

[1] M.J. Ablowitz, A.S.Fokas, Complex variables: Introduction and applications, Cambridge Texts in Applied Mathematics, Cambridge University Press, 2003.

[2] D. Achlioptas, Database-friendly random projections: Johnson-Lindenstrauss with binary coins, Journal of computer and system sciences, 2001.

[3] D. Amelunxen, M. Lotz, M.B. McCoy, J.A. Tropp, Living on the edge: A geometric theory of phase transition in convex optimization, Available online: arXiv: 1303.6672

[4] A. S. Bandeira, D. G. Mixon and B. Recht, Compressive classification and the rare eclipse problem, 2014.

[5] S. Dasgupta, Learning mixtures of gaussians, University of California, Berkeley, 1999.

[6] P. Frankl, H. Maehara, The Johnson-Lindenstrauss lemma and the sphericity of some graphs, Journal of Combinatorial Theory, Series B 44, p.355-362, 1988.

[7] H. Georgii, Stochastik, de Gruyter Lehrbuch, Walter de Gruyter & Co., Berlin, 2002

[8] Y. Gordon, On Milman's inequality and random subspaces which escape through a mesh in $\mathbb{R}^n$, Geometric aspects of functional analysis, Israel Seminar 1986-87, Lecture Notes in Mathematics 1317, p. 84-106, 1988.

[9] A. Gupta, S. Dasgupta, An elementary proof of a theorem of Johnson and Lindenstrauss, Random Structures and Algorithms 22, p. 60-65, 2003.

[10] M. Hazewinkel, C. F. Martin, Representations of the symmetric group, the specialization order, systems and Grassmann manifolds, L'Enseignement Mathématique, 1983, p.53-87

[11] M. Ledoux, M. Talagrand, Probability in Banach Spaces, Springer Verlag, 1991.

[12] W. B. Johnson, J. Lindenstrauss, Conference in Modern Analysis and Probability (New Haven, Conn., 1982), Contemporary Mathematics 26, American Mathematical Society, p. 189-206 1984.

[13] F.W.J. Olver, D.W. Lozier, R.F. Boisvert, C.W. Clark, NIST handbook of mathematical functions, U.S. Departement of Commerce, National Institue of Standards and Technology, Washington DC, 2010.

[14] S. Foucart, H. Rauhut, A Mathematical Introduction to Compressive Sensing, Applied and Numerical Harmonic Analysis, Birkenhäuser, 2013.

[15] R. Schneider, W. Weil, Stochastic and Integral Geometry, Springer series in statistics: Probability and its applications, Springer, 2008.

[16] Scikit-learn: Machine Learning in Python, F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Journal of Machine Learning Research, volume 12, 2011, p.2825-2830.