TECHNISCHE UNIVERSITÄT DARMSTADT

Pareto optimale lineare Klassifikation

Vesselina Poulkova

Betreuer: Eneldo Loza Mencía

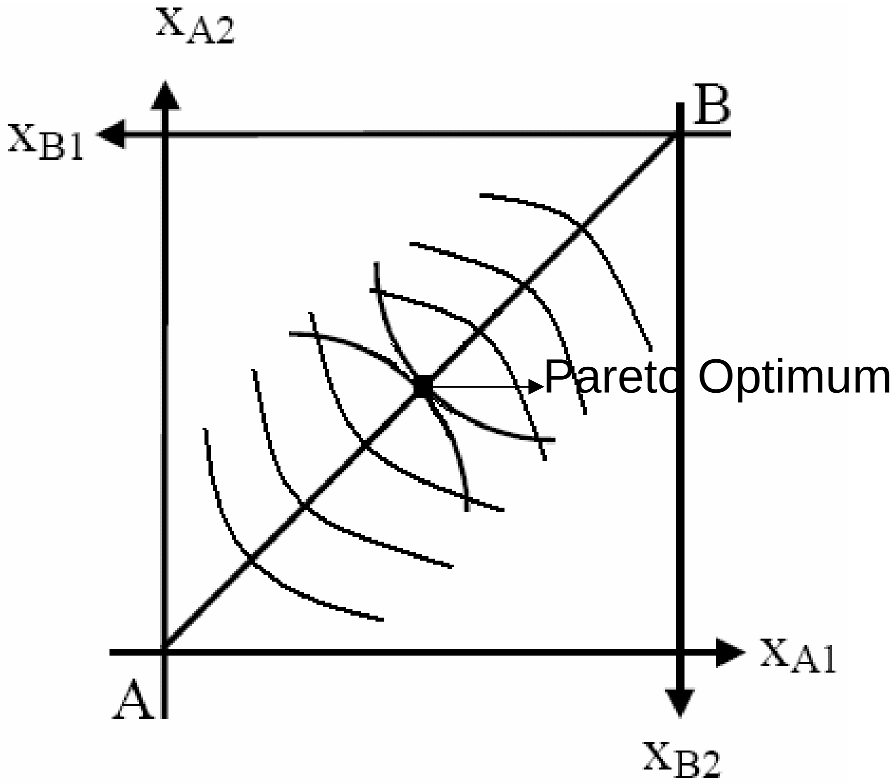
Gliederung

TECHNISCHE UNIVERSITÄT DARMSTADT

1. Einleitung
2. Pareto optimale lineare Klassifizierer
3. Generelle Voraussetzung für Pareto – Optimalität
4. Classification with Scale Mixtures of Normal Distributions
5. Robuste lineare Klassifikation
6. Kernelbasierte Klassifikation
7. Empirische Trade - off Analyse
8. Fazit

1. Einleitung

- Pareto Effizienz in der Wirtschaft:
  - Pareto optimales Gleichgewicht ist eine Allokation, in der es keine Möglichkeit gibt ein Subjekt besser zu stellen, ohne mindestens ein anderes Subjekt schlechter zu stellen.



1. Einleitung

TECHNISCHE UNIVERSITÄT DARMSTADT

- Zwei - Klassen - Problem: Einen linearen Klassifizierer finden, der Wahrscheinlichkeiten für eine Falschklassifikation von Instanzen minimiert.
- Annahmen:
  - Zwei-klassen Klassifikation, in der der Input Raum X $\mathbb{R}^n$ ist, und der Output Menge Y {-1, +1} ist. Das Trainingspaar (x, y), wo x є X und y є Y, wird Beispiel genannt. Ein Beispiel wird negativ (positiv) genannt, falls sein Klassenattribut -1(+1) ist.
  - Die negativen (positiven) Beispiele haben die Verteilung $D_-(D_+)$
  - Idee: $D_-(D_+)$ sind normalverteilt

1. Einleitung

- True negative rate: wie oft ein negatives Beispiel korrekt klassifiziert wird
- True positive rate: wie oft ein positives Beispiel korrekt klassifiziert wird
- Trade-off in zwei-klassen Klassifikation
  - Klassifizierer h: X →Y ordnet jeder Instanz von X eine binäre Klasse zu.
  - Das Klassifikationsergebnis von h ist das Paar: $(P_{\mathrm{tn}}(h), P_{\mathrm{tp}}(h))$
  - $P_{\mathrm{tn}}(h)$ ist die Wahrscheinlichkeit von der „true negative“ Rate
  - $P_{\mathrm{tp}}(h)$ ist die Wahrscheinlichkeit von der „true positive“ Rate

$$\begin{aligned} P_{\mathrm{tn}}(h) &= \Pr(h(x) = -1 \mid y = -1) \\ P_{\mathrm{tp}}(h) &= \Pr(h(x) = +1 \mid y = +1) \end{aligned}$$

  - False positive Rate: $P_{\mathrm{fp}}(h) = 1 - P_{\mathrm{tn}}(h)$
  - False negative Rate: $P_{\mathrm{fn}}(h) = 1 - P_{\mathrm{tp}}(h)$

2. Pareto optimale lineare Klassifizierer

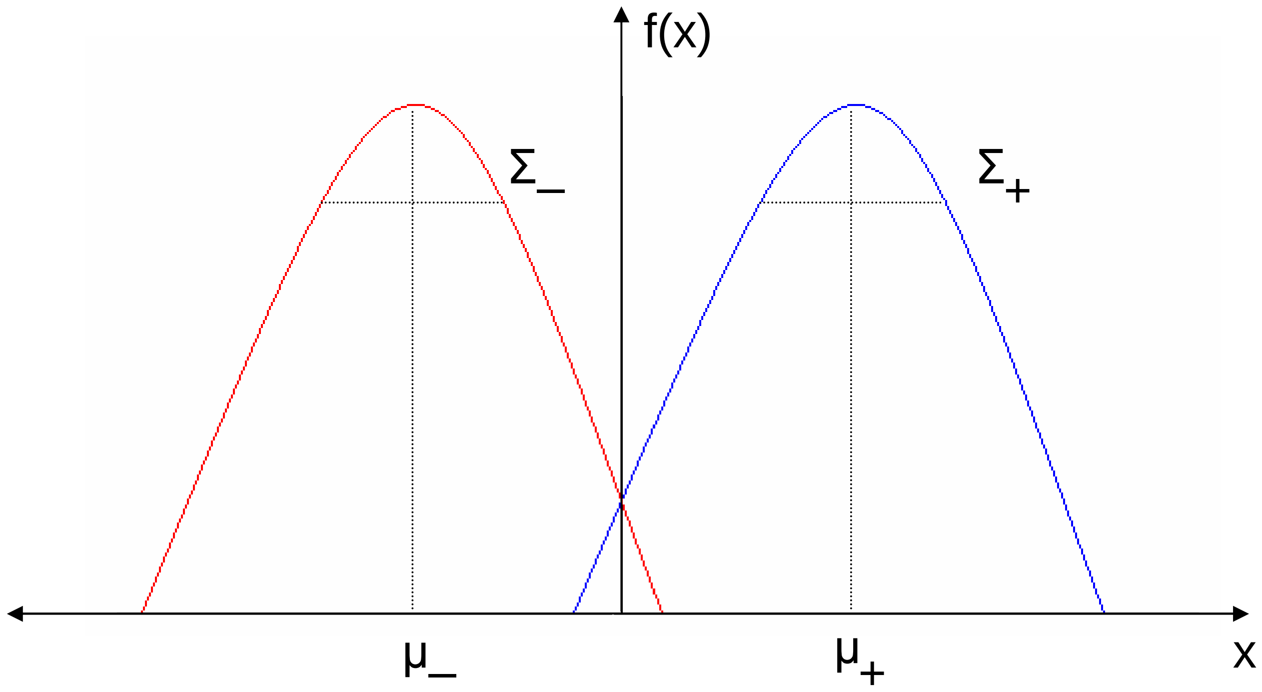
TECHNISCHE UNIVERSITÄT DARMSTADT

- Standard Klassifikationsproblem:
  - Gegeben: Familie Klassifizierer H
  - Finde einen Klassifizierer in H, der die Fehlerrate minimiert.

- Linearer Klassifizierer $h(x)= \text{sgn}(a^T x-b)$
- a - Gewichtsvektor
- b - Threshold
- $H= \{(a, b) | a \in \mathbb{R}^n \setminus \{0\}, b \in \mathbb{R}\}$
- Die klassebedingte Normalverteilung $D_-= N(\mu_-, \Sigma_-)$ und $D_+= N(\mu_+, \Sigma_+)$
- Erwartungswert μ, Kovarianz Matrix Σ (positiv definit)

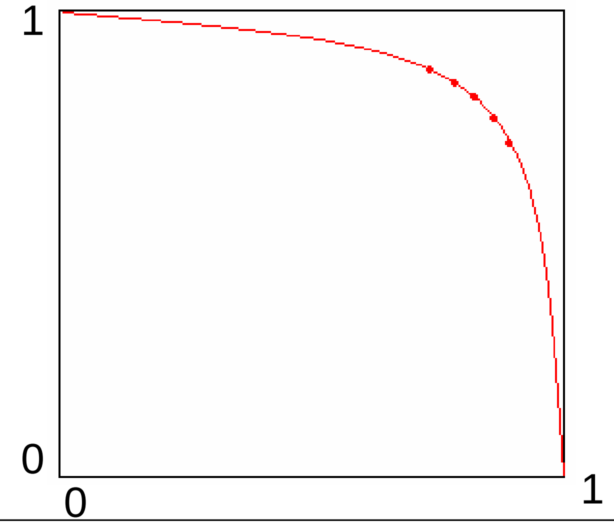
2. Pareto optimale lineare Klassifizierer

- Die Trainingsdaten werden in zwei Klassen unterteilt



2. Pareto optimale lineare Klassifizierer

- Ein Paar von richtig klassifizierten Wahrscheinlichkeiten (α, β) ist erreichbar von H, falls es einen Klassifizierer gibt h є H, so dass $P_{\mathrm{tn}}(h) \geq \alpha$ und $P_{\mathrm{tp}}(h) \geq \beta$.
- Alle erreichbaren Paare (α, β) definieren eine Fläche [0, 1] x [0, 1]
- Die Kurve entlang der oberen Grenze von diesem Viereck wird optimale Tausch - Kurve genannt (optimal trade-off curve) => ein Klassifizierer wird **Pareto optimal** genannt, falls das Paar $(P_{\mathrm{tn}}(h), P_{\mathrm{tp}}(h))$ auf der optimalen Tausch - Kurve liegt.



2. Pareto optimale lineare Klassifizierer

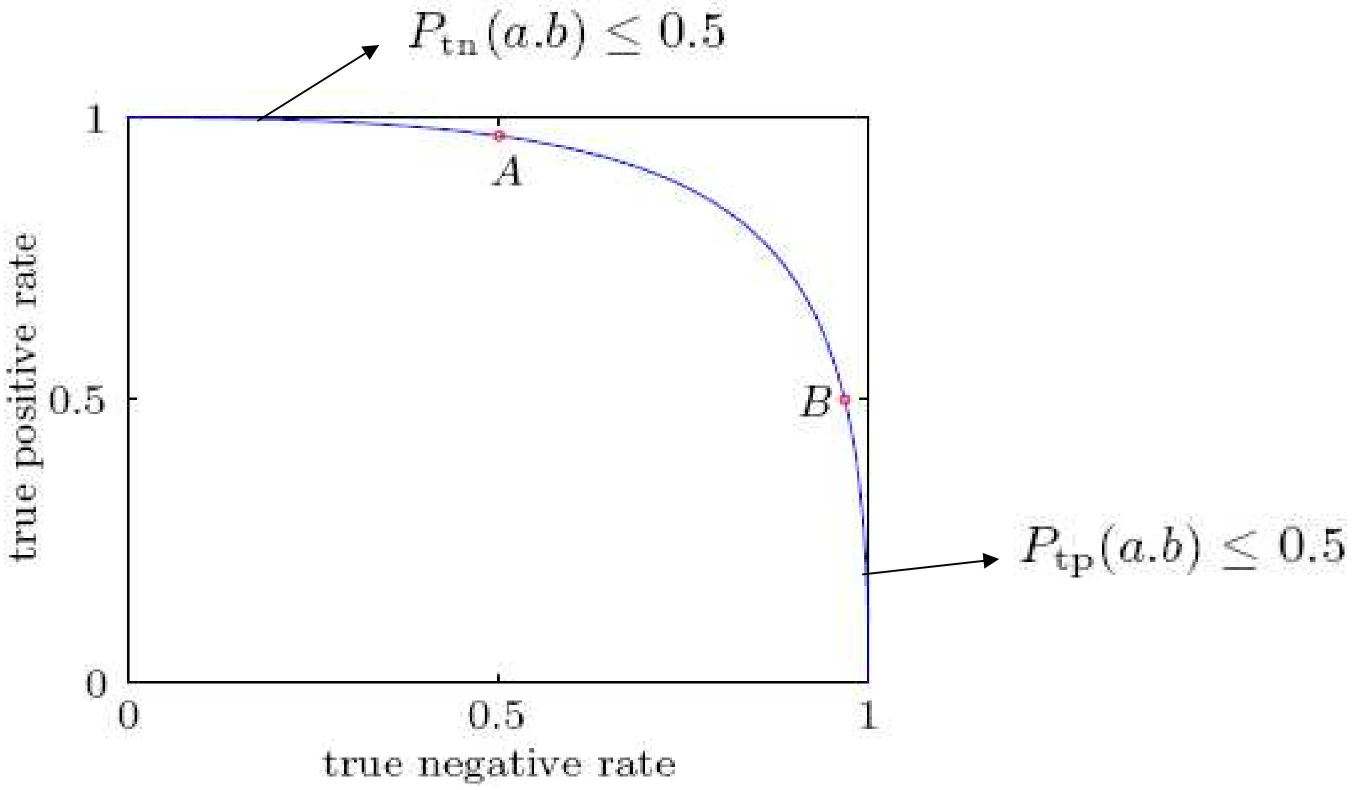
- Mit der **Normalverteilung** können die „true negative" und „true positive" Rate von jedem linearen Klassifizierer (a,b) ausgerechnet werden:

$$\begin{aligned}
\Pr(a^T x < b \mid y = -1) &= \Phi\left(\frac{b - a^T \mu_-}{\sqrt{a^T \Sigma_- a}}\right), \\
\Pr(a^T x > b \mid y = +1) &= \Phi\left(\frac{a^T \mu_+ - b}{\sqrt{a^T \Sigma_+ a}}\right),
\end{aligned}$$

- Φ ist die kumulative Verteilungsfunktion von der Standardnormalverteilung
- $\mathcal{L}(\mu_-, \Sigma_-, \mu_+, \Sigma_+)$ ist die gefundene Menge von pareto - optimalen Klassifizierern

2. Pareto optimale lineare Klassifizierer

- Die optimale Trade-off Kurve



2. Pareto optimale lineare Klassifizierer

TECHNISCHE UNIVERSITÄT DARMSTADT

- Berechnung von $\mathcal{L}(\mu_-, \Sigma_-, \mu_+, \Sigma_+)$:
  - Trade-off Analyse mittels konvexer Optimierung:

    Min $\sqrt{a^T \Sigma_+ a} + \lambda \sqrt{a^T \Sigma_- a}$

    Nebenbedingung: $a^T(\mu_+ - \mu_-) = 1$,wo a ϵ $\mathbb{R}^n$ und Parameter λ >0.

- Das Problem ist streng konvex => eine einzige Lösung a_λ
- b kann dann aus a berechnet werden: $b_\lambda = \mu_+^T a_\lambda - d_\lambda \left(a_\lambda^T \Sigma_+ a_\lambda\right)^{1/2}$
- Ergebnis: Pareto optimale Klassifizierer $(a^\star, b^\star)$ mit:

$P_{\mathrm{tn}}(a^\star, b^\star), P_{\mathrm{tp}}(a^\star, b^\star) > 0.5$, wo $\{(a_\lambda, b_\lambda) \mid 0 < \lambda < \infty\}$

2. Pareto optimale lineare Klassifizierer

- Die optimale Trade-off Kurve mit Verteilung D_ und D_+ ist monoton fallend
- Für λ>0 ist die Kurve $\alpha = \Phi(\lambda\Phi^{-1}(\beta))$ monoton steigend
- Die zwei Kurven schneiden sich im Punkt:

  $(\Phi(\lambda\Phi^{-1}(\beta_\lambda)), \beta_\lambda)$
- Falls $\mu_+ \neq \mu_-$: $\Phi(\lambda\Phi^{-1}(\beta_\lambda)) > 0.5$, $\beta_\lambda > 0.5$
- Alle Punkte auf der Kurve zwischen Punkt A und B können gefunden werden
- Lösen des Problems: Max β

  NB:
$$\Phi\left(\frac{b - a^T\mu_-}{a^T\Sigma_+ a}\right) = \alpha,$$
$$\Phi\left(\frac{a^T\mu_+ - b}{a^T\Sigma_+ a}\right) = \beta,$$
$$\Phi^{-1}(\alpha) = \lambda\Phi^{-1}(\beta),$$

- Min $\sqrt{a^T\Sigma_+ a} + \lambda\sqrt{a^T\Sigma_- a}$
- NB: $a^T(\mu_+ - \mu_-) = 1$,

  wo a є $\mathbb{R}^n$ und Parameter λ >0

2. Pareto optimale lineare Klassifizierer

TECHNISCHE UNIVERSITÄT DARMSTADT

- Optimale Lösung $(a_\lambda, b_\lambda, \alpha_\lambda, \beta_\lambda)$
- (a_λ, b_λ) ist der pareto optimale Klassifizierer mit $P_{\mathrm{tn}}(a_\lambda, b_\lambda) = \alpha_\lambda = \Phi(\lambda\Phi^{-1}(\beta_\lambda))$ und $P_{\mathrm{tp}}(a_\lambda, b_\lambda) = \beta_\lambda$
- Φ^{-1} ist streng wachsend
- Max $\Phi^{-1}(\beta)$
- NB: $$\begin{aligned} \mu_-^T a + \Phi^{-1}(\alpha)\sqrt{a^T\Sigma_- a} &= b, \\ \mu_+^T a - \Phi^{-1}(\beta)\sqrt{a^T\Sigma_+ a} &= b, \\ \Phi^{-1}(\alpha) &= \lambda\Phi^{-1}(\beta). \end{aligned}$$

=> Max $$\frac{a^T(\mu_+ - \mu_-)}{\sqrt{a^T\Sigma_+ a} + \lambda\sqrt{a^T\Sigma_- a}}$$

NB: $a \neq 0$

- Min $\sqrt{a^T\Sigma_+ a} + \lambda\sqrt{a^T\Sigma_- a}$
- NB: $a^T(\mu_+ - \mu_-) = 1$, wo a є $\mathbb{R}^n$ und Parameter λ >0

3. Generale Voraussetzung für pareto Optimalität

TECHNISCHE UNIVERSITÄT DARMSTADT

- Keine Normalverteilung => Pareto Optimum kann auch gefunden werden
- Behauptung:

$$
\begin{aligned}
P_{\mathrm{tn}}(a,b) &= \kappa_{-}\left(\frac{b - a^T\mu_{-}}{\sqrt{a^T\Sigma_{-}a}}\right)\\
P_{\mathrm{tp}}(a,b) &= \kappa_{+}\left(\frac{a^T\mu_{+} - b}{\sqrt{a^T\Sigma_{+}a}}\right)
\end{aligned}
$$

  - k_ und k_+ sind streng wachsend in $\mathbb{R}$. Die Menge von pareto - optimalen Klassifizieren ist: $\mathcal{L}(\mu_{-}, \Sigma_{-}, \mu_{+}, \Sigma_{+})$

  - $D_{-} = N(\mu_{-}, \lambda_{-}\Sigma_{-})$ und $D_{+} = N(\mu_{+}, \lambda_{+}\Sigma_{+})$

4. Classification with Scale Mixtures of Normal Distributions

TECHNISCHE UNIVERSITÄT DARMSTADT

- Mix aus unterschiedlich skalierten Normalverteilungen:
  - Dichtefunktion:

$$p_X(x) = \int \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} e^{-(x-\mu)^T \lambda \Sigma (x-\mu)} p_\Lambda(\lambda)\, d\lambda$$

  - Diese Verteilung wird $S(\mu, \Sigma, p_\Lambda)$ genannt
  - Lemma: Falls x ~ $S(\mu, \Sigma, p_\Lambda)$. Dann:

$$\Pr(a^T x > b) = \kappa \left(\frac{a^T \mu - b}{\sqrt{a^T \Sigma a}} \right)$$, mit: $$\kappa(u) = \int_0^\infty \Phi(u/\sqrt{\lambda}) p_\Lambda(\lambda)\, d\lambda$$

  - Für jede p_ und p_+ kann $\mathcal{L}(\mu_-, \Sigma_-, \mu_+, \Sigma_+)$ berechnet werden mit der klassbedingten Verteilungen $D_- = S(\mu_-, \Sigma_-, p_-)$ und $D_+ = S(\mu_+, \Sigma_+, p_+)$

5. Robuste lineare Klassifikation

TECHNISCHE UNIVERSITÄT DARMSTADT

- D_- und D_+ sind unbekannt, aber wichtige Information darüber ist gegeben.
- Worst- case Wahrscheinlichkeiten:

$$P_{\mathrm{tn}}^{\mathrm{wc}}(h) = \inf\{\Pr(h(x) < 0) \mid x \sim D_- \in \mathcal{D}_-\}$$
$$P_{\mathrm{tp}}^{\mathrm{wc}}(h) = \inf\{\Pr(h(x) > 0) \mid x \sim D_+ \in \mathcal{D}_+\}$$

- Klassifikation mit der Schranke von Tschebyschev
  - Die klassebedingte Verteilung ist nicht vollständig bekannt:

$$D_- \in \mathcal{D}_- = \mathcal{D}(\mu_-, \Sigma_-), \qquad D_+ \in \mathcal{D}_+ = \mathcal{D}(\mu_+, \Sigma_+)$$

5. Robuste lineare Klassifikation

TECHNISCHE UNIVERSITÄT DARMSTADT

- Durch die Tschebyschev Schranke kann die worst-case true negative und positive Rate berechnet werden:

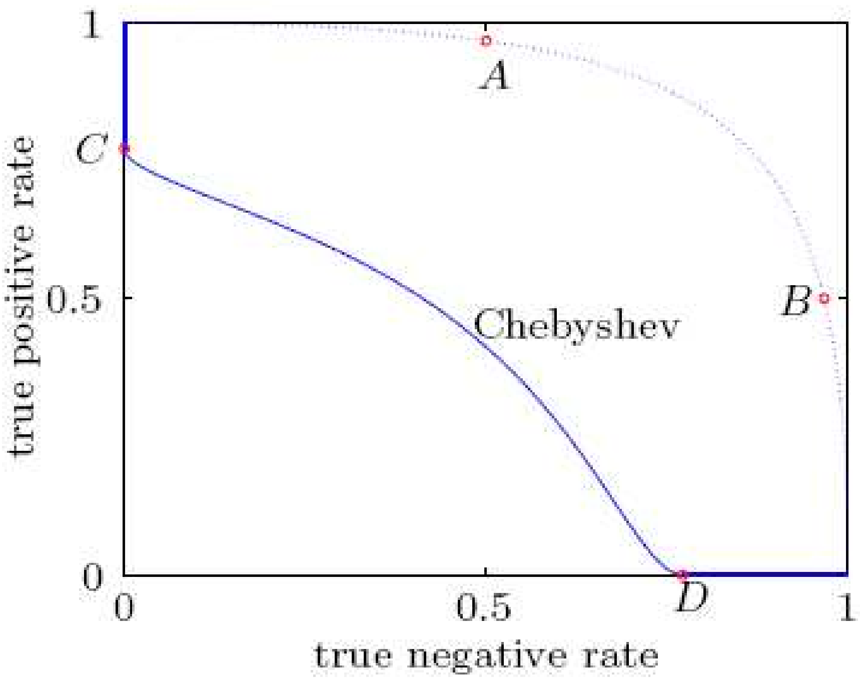
$$\begin{aligned}
P_{\mathrm{tn}}^{\mathrm{wc}}(a,b) &= \inf\left\{\Pr(a^T x < b) \mid x \sim D_- \in \mathcal{D}_-\right\} \\
&= \Psi\left(\frac{b - a^T\mu_-}{\sqrt{a^T\Sigma_- a}}\right), \\
P_{\mathrm{tp}}^{\mathrm{wc}}(a,b) &= \inf\left\{\Pr(a^T x > b) \mid x \sim D_+ \in \mathcal{D}_+\right\} \\
&= \Psi\left(\frac{a^T\mu_+ - b}{\sqrt{a^T\Sigma_+ a}}\right),
\end{aligned}$$

- Die Funktion ψ ist streng wachsend über (0,∞):

$$\Psi(u) = u_+^2/(1+u_+^2), \qquad u_+ = \max\{u, 0\}$$

5. Robuste lineare Klassifikation

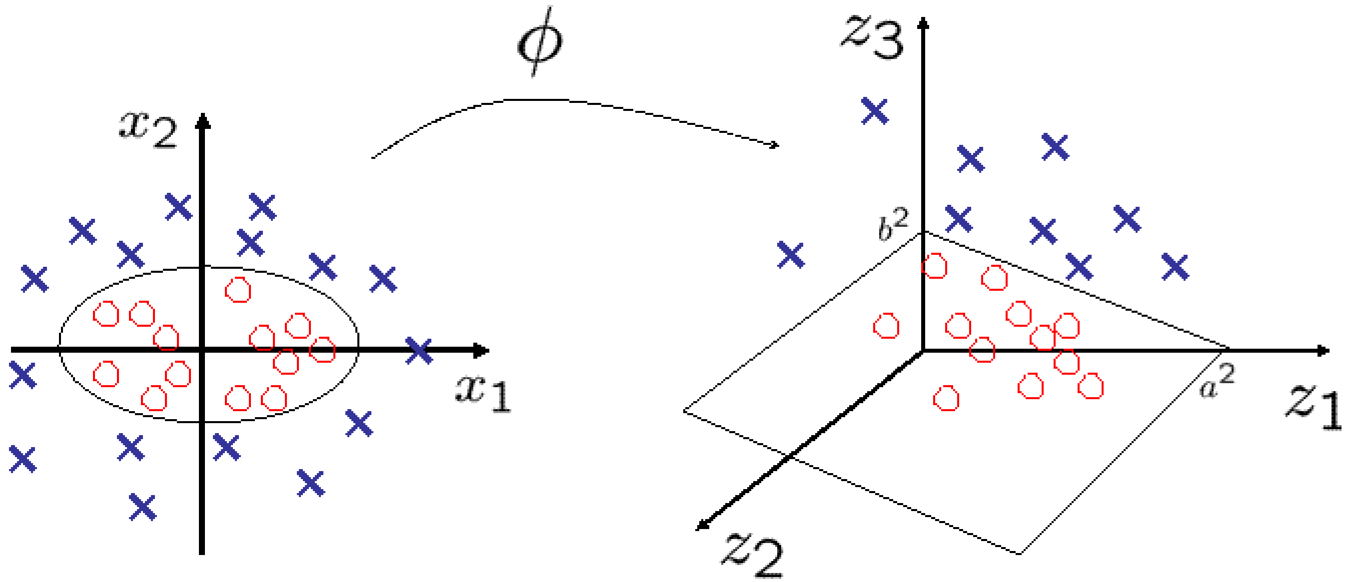
- Die optimale Trade-off Kurve mit der Schranke von Tschebyschev



6. Kernelbasierte Klassifikation

TECHNISCHE UNIVERSITÄT DARMSTADT

- Trade-off Analysis mit kernelbasierten Klassifizierern



$$\phi : (x_1, x_2) \longrightarrow (x_1^2, \sqrt{2}x_1x_2, x_2^2)$$

Abbildung: http://omega.albany.edu:8008/machine-learning-dir/notes-dir/ker1/phiplot.gif

6. Kernelbasierte Klassifikation

TECHNISCHE UNIVERSITÄT DARMSTADT

- Klassifizierer h: X → Y, $h(x) = \mathrm{sgn}(a^T \phi(x) - b)$
  - a є H ist ein Gewichtsvektor in dem hochdimensionalen Hilbert Raum H
  - ϕ ist eine Abbildung von X nach H
  - Verteilung: $N(\tilde{\mu}_-, \tilde{\Sigma}_-)$ und $N(\tilde{\mu}_+, \tilde{\Sigma}_+)$ für die negative und positive Klasse
  - Trainingsinstanzen: $\{x_1, \ldots, x_{m+}\}$ von der positiven Klasse und $\{x_{m_+ + 1}, \ldots, x_m\}$ von der negativen Klasse, mit $m_- = m - m_+$
  - Erwartungswert:

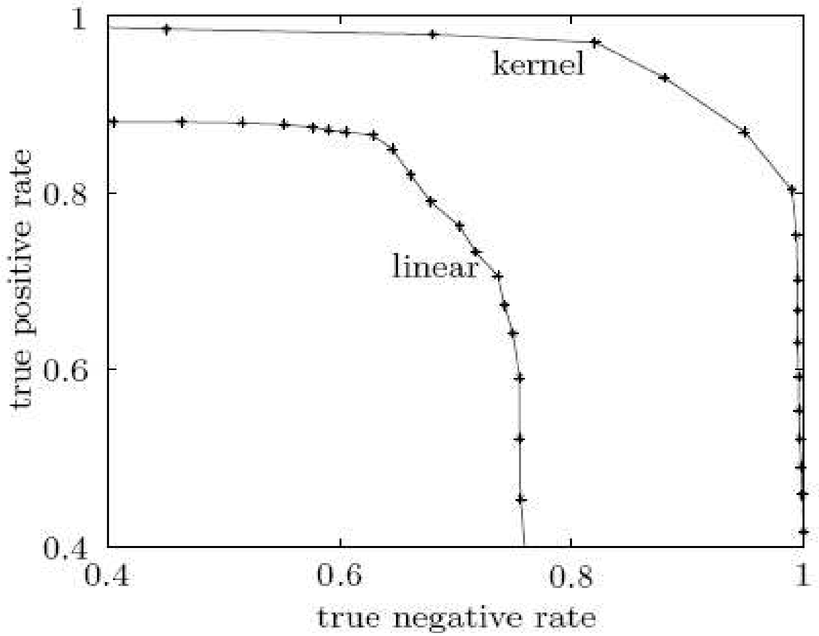
$$\tilde{\mu}_+ = \frac{1}{m_+} \sum_{i=1}^{m_+} \phi(x_i), \quad \tilde{\mu}_- = \frac{1}{m_-} \sum_{i=m_+ + 1}^{m} \phi(x_i)$$

7. Empirische Trade - off Analyse

- Trainingsinstanzen, um die Stichproben Erwartungswert und Varianz abzuschätzen
- Pareto optimalen Klassifizierer finden
- True positive und negative rate finden
- Mehrmals wiederholen und die Resultaten sammeln
- Die Trade-off Kurve kann dann durch Regression der kleinsten Quadrate berechnet werden

7. Empirische Trade-off Analyse

- Ionosphere benchmark data set from the UCI repository:
  - 351 points in $\mathbb{R}^{34}$
  - 70% der Daten werden als Trainingsinstanz benutzt
  - Kernelbasierte Klassifikation ist besser



8. Fazit

- Klassifikation mit der Schranke von Tschebyschev ist vergleichbar mit dem MEMPM Algorithmus.
- Die Klassifizierer, die mit der robusten linearen Klassifikation gefunden werden, kann man mit den Ergebnissen von der Support Vector Machine vergleichen.
- Es wurden aber keine direkten Tests gemacht.
- Die Robustheit - Analyse basiert auf der Behauptung, dass es keinen Schätzfehler gibt bei der Schätzung des Erwartungswerts und der Varianz.
- Die pareto optimalen Klassifizierer können unter dem „small sample problem“ leiden.

Vielen Dank für Ihre Aufmerksamkeit!