Linear nichtseparable Probleme

Mustererkennung und Klassifikation, Vorlesung No. 10[1]

M. O. Franz

20.12.2007

[1] falls nicht anders vermerkt, sind die Abbildungen entnommen aus Duda et al., 2001.

Übersicht

1. Linear nicht trennbare Probleme
2. Kleinste Quadrate
3. Widrow-Hoff-Methode

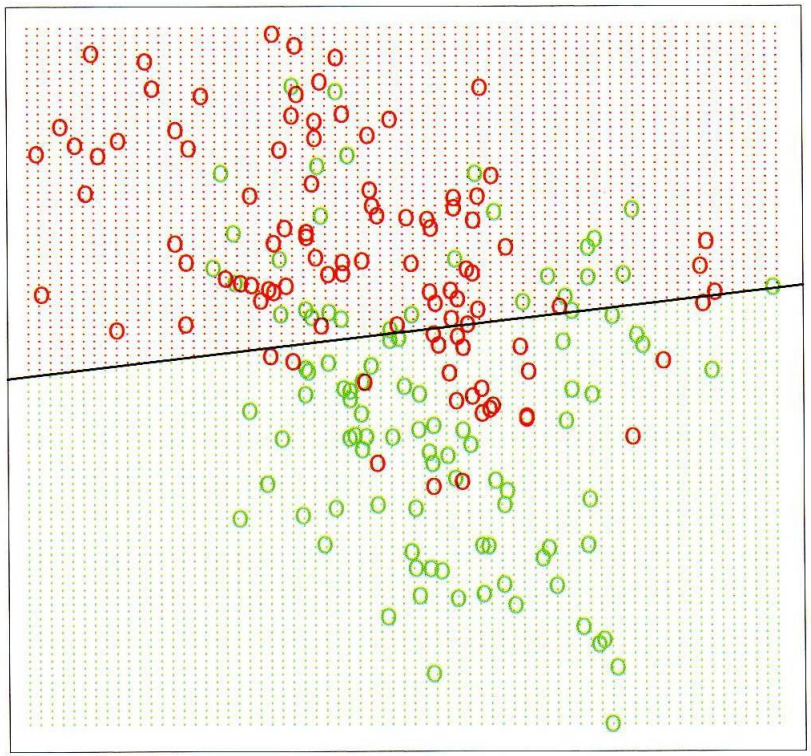
Übersicht

1 Linear nicht trennbare Probleme

2 Kleinste Quadrate

3 Widrow-Hoff-Methode

Linear nicht trennbare Probleme



Die bisher betrachteten linearen Methoden basieren auf der Korrektur von Fehlern: Der Gewichtsvektor wird nur dann verändert, wenn ein Fehler auftritt.

In der Praxis sind auf der linearen Trennbarkeit beruhende Methoden nur einsetzbar, wenn die Fehlerrate der optimalen linearen Diskrimimantenfunktion sehr niedrig ist.

Bei nicht linear trennbaren Problemen wird der Fehler nie Null $\Rightarrow$ fehlerkorrekturbasierte Methoden laufen unendlich weiter.

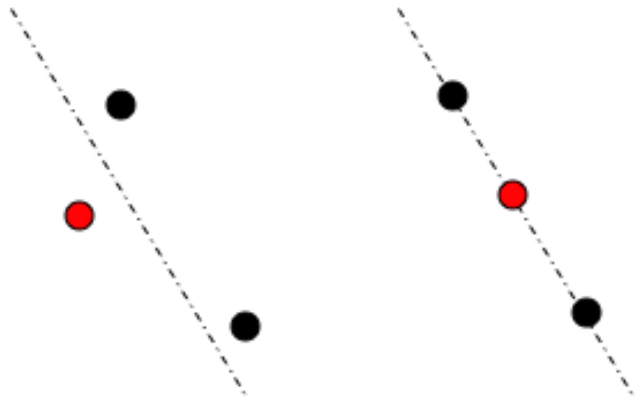
Lineare Trennbarkeit in höheren Dimensionen

- Bei d-dimensionalen Daten sind weniger als $2d$ Datenpunkte fast immer linear trennbar.
- Selbst wenn die Trainingsdaten linear trennbar sind, folgt daraus nicht, dass der resultierende Klassifikator gut auf unabhängigen Testdaten funktioniert.
- Man sollte also ein Mehrfaches von $2d$ an Trainingsdaten haben, um den Klassifikator zu überbestimmen und sicherzustellen, daß Trainings- und Testdaten ähnlich genug sind.

Lineare Trennbarkeit wird um so unwahrscheinlicher, je größer die Trainingsmengen werden.

Lineare Trennbarkeit in höheren Dimensionen

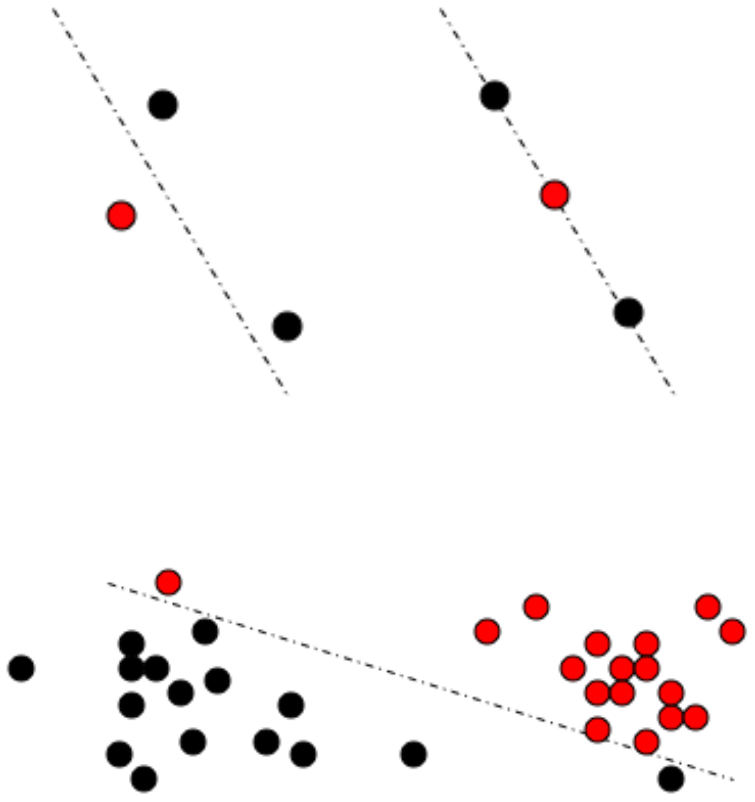
- Bei d-dimensionalen Daten sind weniger als $2d$ Datenpunkte fast immer linear trennbar.
- Selbst wenn die Trainingsdaten linear trennbar sind, folgt daraus nicht, dass der resultierende Klassifikator gut auf unabhängigen Testdaten funktioniert.
- Man sollte also ein Mehrfaches von $2d$ an Trainingsdaten haben, um den Klassifikator zu überbestimmen und sicherzustellen, daß Trainings- und Testdaten ähnlich genug sind.



Lineare Trennbarkeit wird um so unwahrscheinlicher, je größer die Trainingsmengen werden.

Lineare Trennbarkeit in höheren Dimensionen

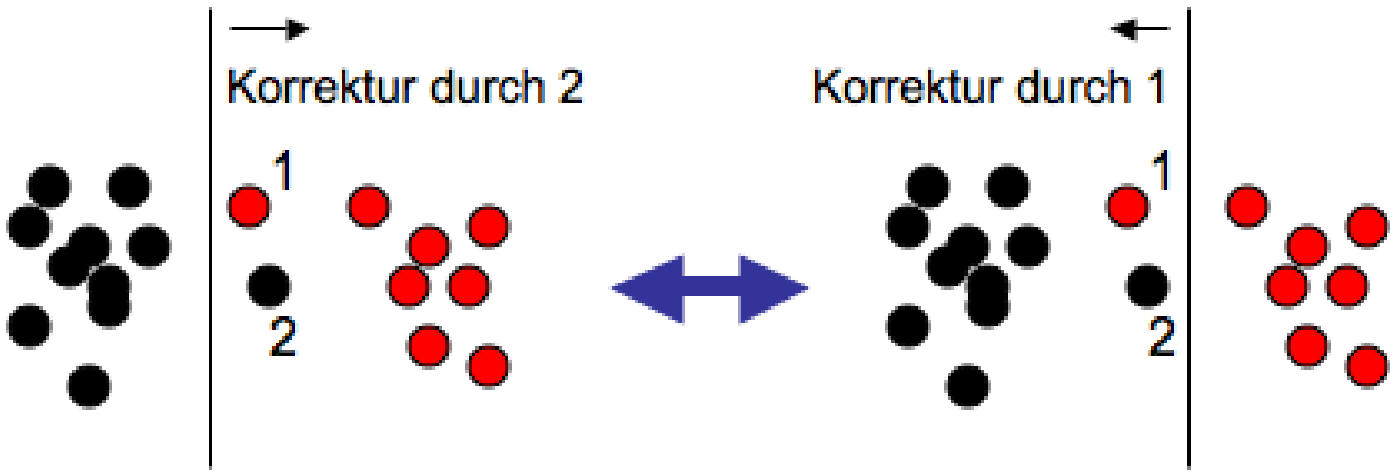
- Bei d-dimensionalen Daten sind weniger als $2d$ Datenpunkte fast immer linear trennbar.
- Selbst wenn die Trainingsdaten linear trennbar sind, folgt daraus nicht, dass der resultierende Klassifikator gut auf unabhängigen Testdaten funktioniert.
- Man sollte also ein Mehrfaches von $2d$ an Trainingsdaten haben, um den Klassifikator zu überbestimmen und sicherzustellen, daß Trainings- und Testdaten ähnlich genug sind.



Lineare Trennbarkeit wird um so unwahrscheinlicher, je größer die Trainingsmengen werden.

Verhalten von Fehlerkorrekturmethoden bei nicht linear trennbaren Problemen

Bei nicht linear trennbaren Problemen hört die Fehlerkorrektur nie auf.

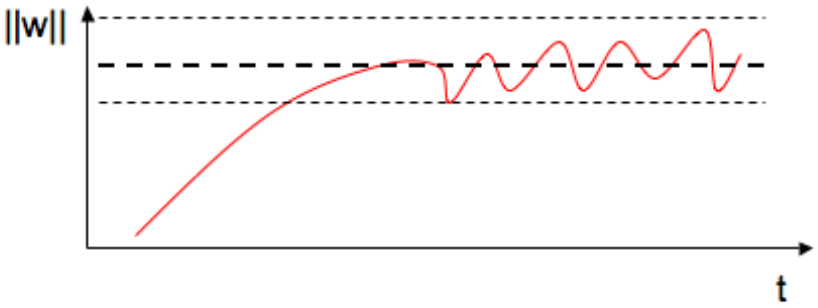


Beispiel Perzeptron mit Einzelbeispielkorrektur: $a \leftarrow a + y_k$

$\Rightarrow$ Gewichtsvektor oszilliert zwischen 2 Werten hin und her.

Heuristiken für nichtseparable Probleme

Ziel: Erträgliche Leistung in nichtseparablen Problemen unter Beibehaltung der Fähigkeit, trennbare Problem zu separieren.



- Detektion des Gleichgewichtszustandes durch Überwachung der Länge des Gewichtsvektors

- Um nicht zufällig im schlechtesten Endzustand stehenzubleiben, wird über alle Endzustände gemittelt.
- Häufig wird die Lernrate nicht fest gewählt, sondern über die Zeit abnehmend. damit wird eine Konvergenz auch im nichtseparablen Fall erzwungen. Problem: Wahl der richtigen Abnahmegeschwindigkeit (oft abhängig von Lernleistung).

Übersicht

1 Linear nicht trennbare Probleme

2 Kleinste Quadrate

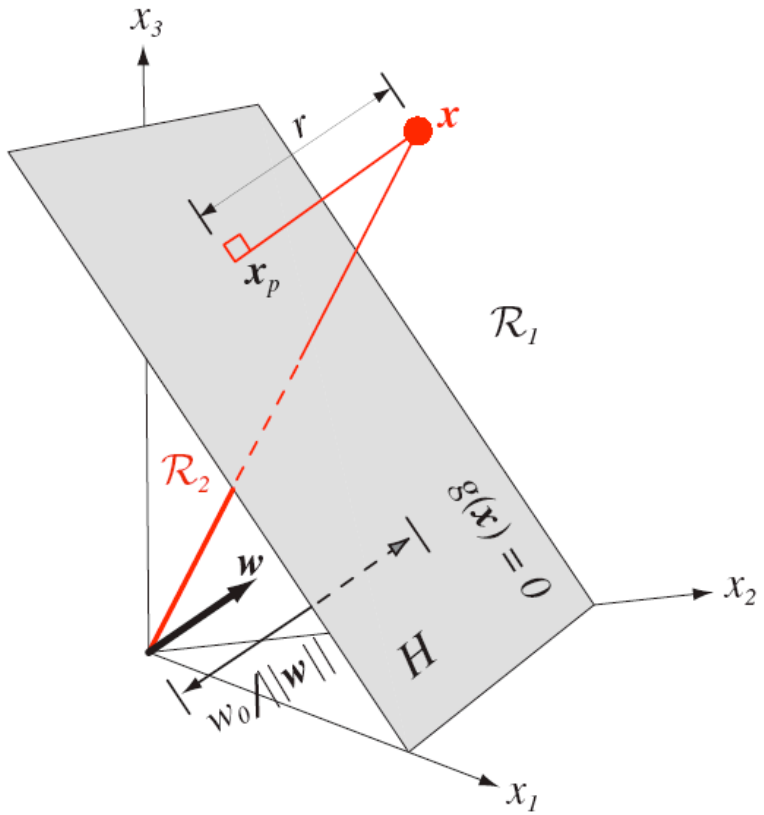
3 Widrow-Hoff-Methode

Methode der kleinsten Quadrate

Vorgehensweise:

- Das Ziel, einen separierenden Gewichtsvektor zu finden, wird aufgegeben. Stattdessen wird eine gute Klassifikationsleistung bei trennbaren und nicht trennbaren Problemen angestrebt.
- Statt Fokussierung auf falsch klassifizierte Beispiele wie bisher werden jetzt alle Beispiele in der Fehlerfunktion berücksichtigt.
- Bisher wurde ein Gewichtsvektor a gesucht, der alle Skalarprodukte $a^\top y_i$ positiv macht. Jetzt suchen wir ein a, das die Gleichung $a^\top y_i = b_i$ erfüllt (b_i sind frei wählbare positive Konstanten).
- Statt einer Lösung für eine Menge von linearen Ungleichungen wird jetzt eine Lösung für ein lineares Gleichungssystem $a^\top y_i = b_i$ gesucht.

Was heißt $a^\top y_i = b_i$?



Wir erinnern uns: Abstand zur Trennebene

$$r = \frac{g(x)}{\|w\|} = \frac{a^\top y_i}{\|w\|}$$

d.h. $a^\top y_i = r\|w\|$ beschreibt den Abstand, skaliert durch den Betrag des Gewichtsvektors.

Durch Wahl einer positiven Konstante b_i in der Gleichung $a^\top y_i = b_i$ verlangen wir, daß Beispiel y_i den skalierten Abstand b_i von der Entscheidungsebene (engl. margin) haben soll.

Klassifikation als Lösung eines linearen Gleichungssystems

Aufgabe: Finde einen Gewichtsvektor a, so daß die zugehörige Entscheidungsebene zu jedem Datenpunkt y_i den skalierten Abstand b_i hat.

Mathematisch gesprochen, muß a bei n Datenpunkten ein System von n linearen Gleichungen der Form $a^\top y_i = b_i$ erfüllen.

Matrixschreibweise:

$$\begin{pmatrix} y_{10} & y_{11} & \cdots & y_{1d} \\ y_{20} & y_{21} & \cdots & y_{2d} \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ y_{n0} & y_{n1} & \cdots & y_{nd} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_d \end{pmatrix} = \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ \vdots \\ b_n \end{pmatrix} \quad \text{oder} \quad Ya = b.$$

Methode der kleinsten Quadrate

Da wir ein Mehrfaches von $2d$ Datenpunkten zur Überbestimmung der Entscheidungsebene brauchen und wir von nicht linear trennbaren Problemen ausgehen, ist das Gleichungssystem $Ya = b$ nicht exakt lösbar.

Bei einem überbestimmten Gleichungssystem haben wir mehr Gleichungen als Unbekannte $\Rightarrow$ Die Matrix Y ist nichtquadratisch, somit existiert i.A. die formale Lösung $a = Y^{-1}b$ nicht.

Ansatz: Suche eine Näherungslösung durch Minimierung des quadratischen Fehlers (Methode der kleinsten Quadrate).

Fehlervektor: $e = Ya - b$

Quadratischer Fehler: $J_z(a) = \|e\|^2 = \|Ya - b\|^2 = \sum_{i=1}^{n} (a^\top y_i - b_i)^2$

Pseudoinverse

Am Minimum muß die Ableitung des quadratischen Fehlers $J_z(a) = \sum_{i=1}^{n}(a^\top y_i - b_i)^2$ nach a gleich 0 sein, d.h.

$$\nabla J_z = \sum_{i=1}^{n} 2(a^\top y_i - b_i)y_i = 2Y^\top(Ya - b) = 0$$

Umstellung führt auf Normalengleichung:

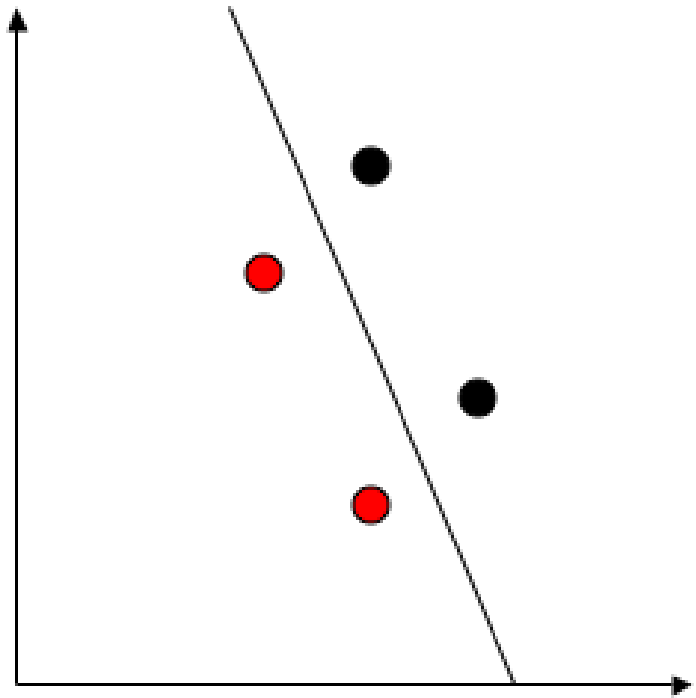
$$Y^\top Ya = Y^\top b,$$

d.h zur Minimierung des quadratischen Fehlers muß nicht mehr die Gleichung $Ya = b$ gelöst werden, sondern $Y^\top Ya = Y^\top b$. Da $Y^\top Y$ immer quadratisch ist, existiert eine Inverse, solange $Y^\top Y$ nicht singulär ist:

$$a = (Y^\top Y)^{-1}Y^\top b = Y^\dagger b$$

mit der Pseudoinversen $Y^\dagger = (Y^\top Y)^{-1}Y^\top$.

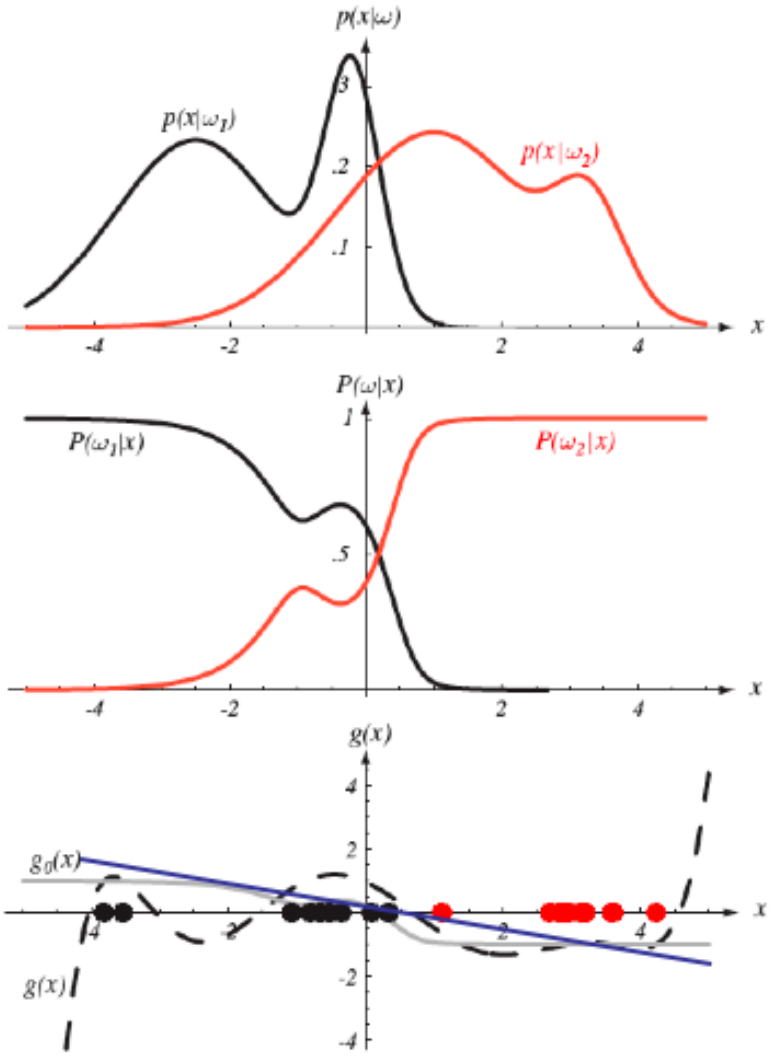
Aufgabe 10.1: Konstruktion eines linearen Klassifikators mit der Pseudoinversen



Geg. 2 Punkte aus Klasse ω_1: (1,2) uns (2,0) und aus Klasse ω_2: (3,1) und (2,3).

Finden Sie die Kleinste-Quadrate-Lösung für den Marginvektor $b = (1, 1, 1, 1)^\top$.

Kleinste Quadrate und Bayes-Klassifikator



- Man kann zeigen (s. Duda, S.243): Die durch kleinste Quadrate gefundene Lösung ist die beste lineare Näherung für die Bayes-Diskriminantenfunktion

$$g_0(x) = p(\omega_1|x) - p(\omega_2|x)$$

- Näherung ist dort am besten, wo
  - viele Datenpunkte sind,
  - $p(x)$ groß ist.

Übersicht

1 Linear nicht trennbare Probleme

2 Kleinste Quadrate

3 Widrow-Hoff-Methode

Widrow-Hoff- oder LMS-Methode (1)

- Statt direkt den quadratischen Fehler $J_z(a) = \|Ya - b\|^2$ zu minimieren, kann man auch hier (wie beim Perzeptron) einen Gradientenabstieg durchführen.
- Vorteil 1: keine Probleme, wenn $Y^\top Y$ singulär wird.
- Vorteil 2: bei sehr hochdimensionalen Problemen muß keine vollständige $d \times d$-Matrix $Y^\top Y$ im Speicher gehalten werden.
- Vorteil 3: robuster bei numerischen Problemen.
- Nachteil: häufig sehr lange Laufzeit.

Gradient des quadratischen Fehlers:

$$\nabla J_z = \sum_{i=1}^{n} 2(a^\top y_i - b_i)y_i = 2Y^\top(Ya - b)$$

Widrow-Hoff- oder LMS-Methode (2)

Offensichtliche Lernregel:

$$a_{\text{neu}} = a_{\text{alt}} - \eta_k \nabla J_z = a_{\text{alt}} - \eta_k Y^\top (Ya - b)$$

Man kann zeigen: Konvergenz für $\eta_k = \eta_1 / k$, auch bei singulärem $Y^\top Y$

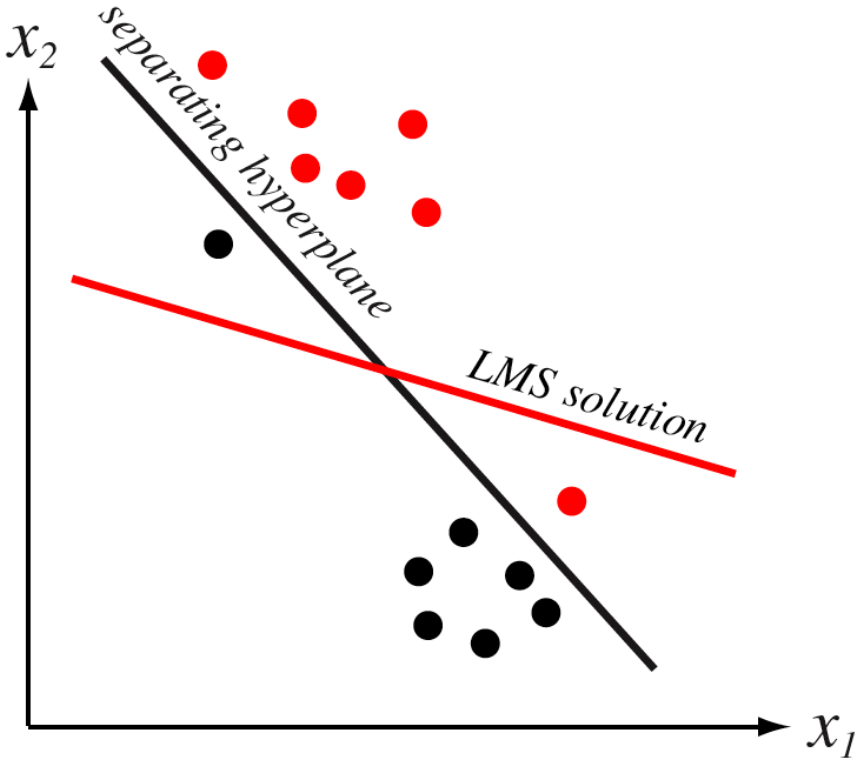
Speicherbedarf wird verringert, wenn a für jeden Datenpunkt einzeln angepaßt wird:

$$a_{\text{neu}} = a_{\text{alt}} + \eta_k (b_k - a_k^\top y_k) y_k$$

Algorithmus: Widrow-Hoff

begin initialize a, b, η_k, Schwellwert θ, $k = 0$
  do $k \leftarrow (k+1) \mod n$
    $a \leftarrow a + \eta_k (b_k - a_k^\top y_k) y_k$
  until $|\eta_k (b_k - a_k^\top y_k) y_k| < \theta$
  return a
end

Kleinste Quadrate und separierende Entscheidungsebene



- Die Kleinste-Quadrate-Lösung (Pseudoinverse oder Widrow-Hoff) ist normalerweise keine separierende Entscheidungsebene, auch wenn eine solche existiert.