**WSI für Informatik** **an der Karls-Eberhardt Universität Tübingen**

# Machine Learning

## Übungsblatt 5

Lea Bey - Benjamin Çoban - Thomas Stüber

15. Juni 2017

# 5

**Aufgabe 5.1.**

a) Warum verwendet man die PCA?
Die Aufgabe von PCA besteht darin, Merkmale zusammenzufassen und somit Dimensionen meiner Daten zu reduzieren. Das bringt zwar in jedem Fall einen Fehler mit sich, allerdings ist der PCA so konzipiert, dass der entstehende Fehler bei minimiert wird. Das ist dann nützlich, wenn ich (**1**) zu viele Dimensionen für relativ wenig Daten habe. Dann ist das einfach undurchschaubar. Somit hilft die PCA, indem die Streumatrix den Rang und somit auch $\min\{n, p\}$ Eigenwerte (Komponenten) besitzt, wobei $n$ die Anzahl der Daten und $p$ die Anzahl der Merkmale sind. Ein weiterer Grund:
(**2**) Wenn einige Merkmale eine gewisse Redundanz besitzen, das heißt, dass sie stark korrelierend sind. Dann lassen sie sich zusammenfassen und die PCA arbeitet dabei schnell, einfach und effizient, sodass wir eine gute Repräsentation der Daten bekommen im Sinne der Varianz.

b) Ziel von PCA: Der Algorithmus möchte die Daten auf die größte Varianz reduzieren. Dazu berechnet er die Eigenwerte und -vektoren der Streumatrix aller Daten und schaut, welche groß - also relevant - sind. Die entscheiden dann die einzelnen Komponenten in absteigender Reihenfolge und maximieren somit die Varianz der Daten. Insofern liegt da der große Unterschied zu LDA
LDA berechnet sowohl die Mittelwerte, als auch die within-class und between-class Streumatrizen. Die Eigenwerte von $A_{within}^{-1} \cdot A_{between}$ geben nun die Diskriminanten für den neuen Unterraum an und dann schiebt LDA alles rüber. Maximiert wird hiermit die Streuung zwischen den Klassen, wobei die Streuung innerhalb einer Klasse minimiert wird ($J(w)$ im Skript). Es reicht nicht, sie einzeln zu maximieren, da ansonsten schlicht und einfach nichts neues passiert. Somit ergibt sich im neuen Schaubild eine Klassifizierungsmöglichkeit.

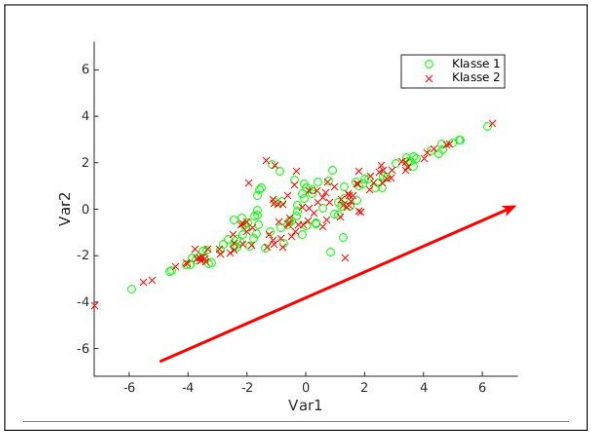
c) Die within-class Streumatrix ist die Summe der einzelnen Streumatrizen pro Klasse ($\sum(x-m)(x-m)^t$) und beschreibt, wie stark die Daten innerhalb einer Klasse variieren.

d) Die between-class Streumatrix beschreibt, wie weit die einzelnen Klassenmittelwerte vom globalen Mittelwert abweichen und gibt somit an, wie die Klassen beieinander liegen.

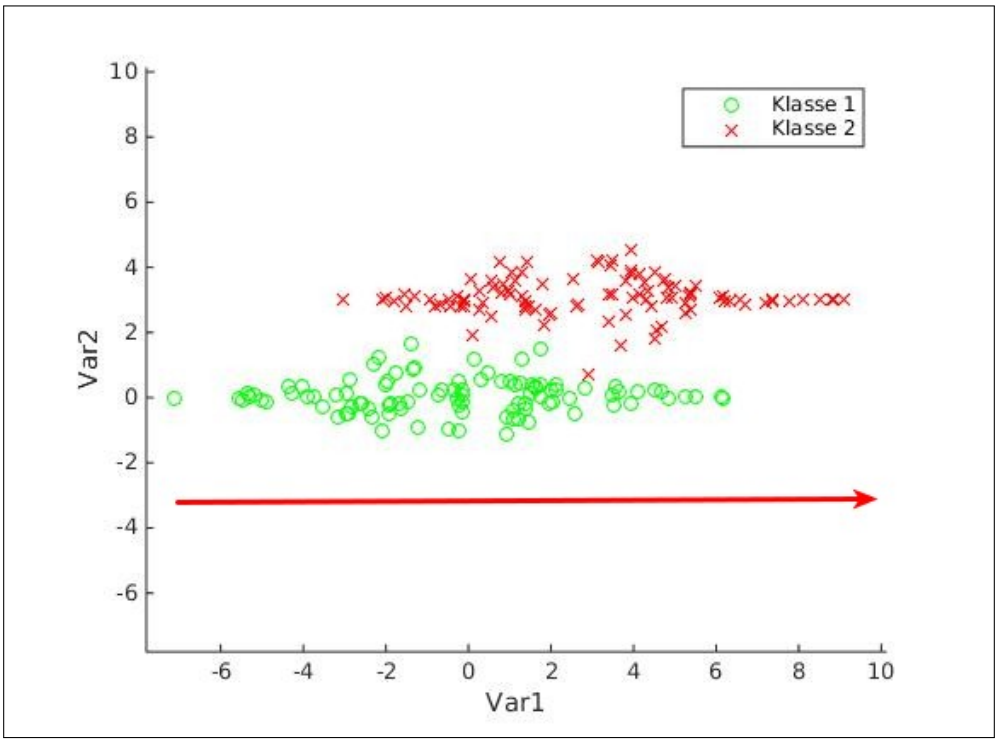
e) PCA gilt hierbei als ein *unsupervised* Verfahren, da es sich ausschließlich für die Varianzmaximierung interessiert. Im Gegensatz dazu berechnet LDA die Richtungen (Lineare Diskriminanten), um die Daten möglichst weit in Klassen aufzuteilen. Somit gilt LDA als ein *supervised* Verfahren.

**Aufgabe 5.2.**

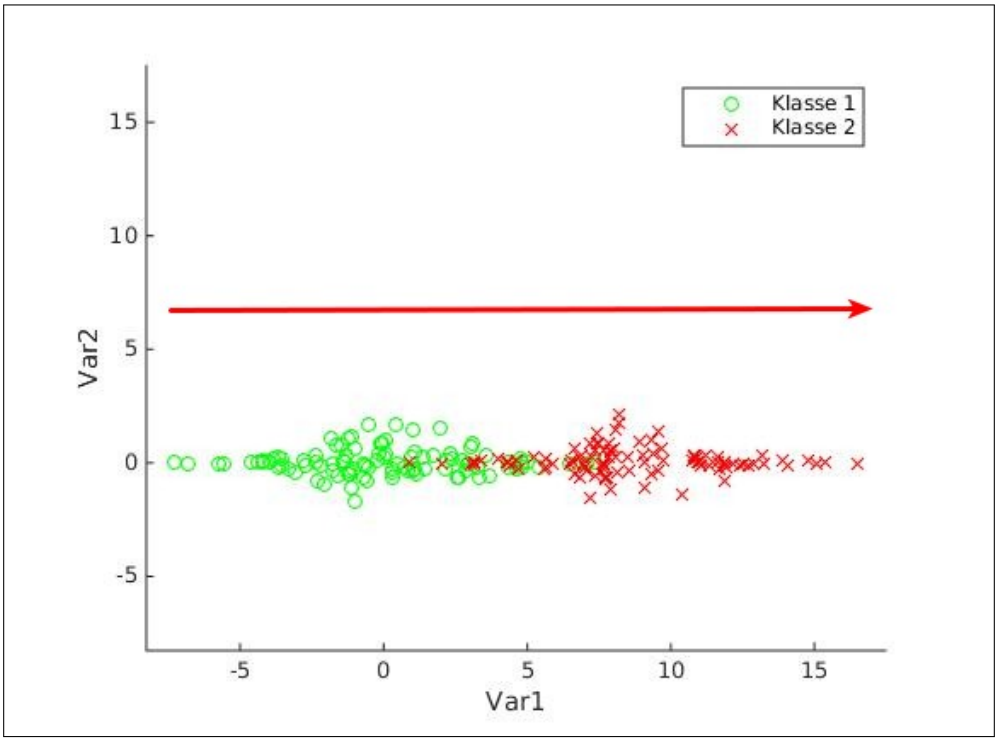
Einzuzeichnen sind in die Grafiken die erste Komponente nach PCA, also die Richtung, in der erst mal am stärksten gestreut wird. Das macht dann nämlich den ersten Eigenvektor laut dem PCA Algoritmus aus. In Rot wurde der erste Komponentenvektor eingetragen.



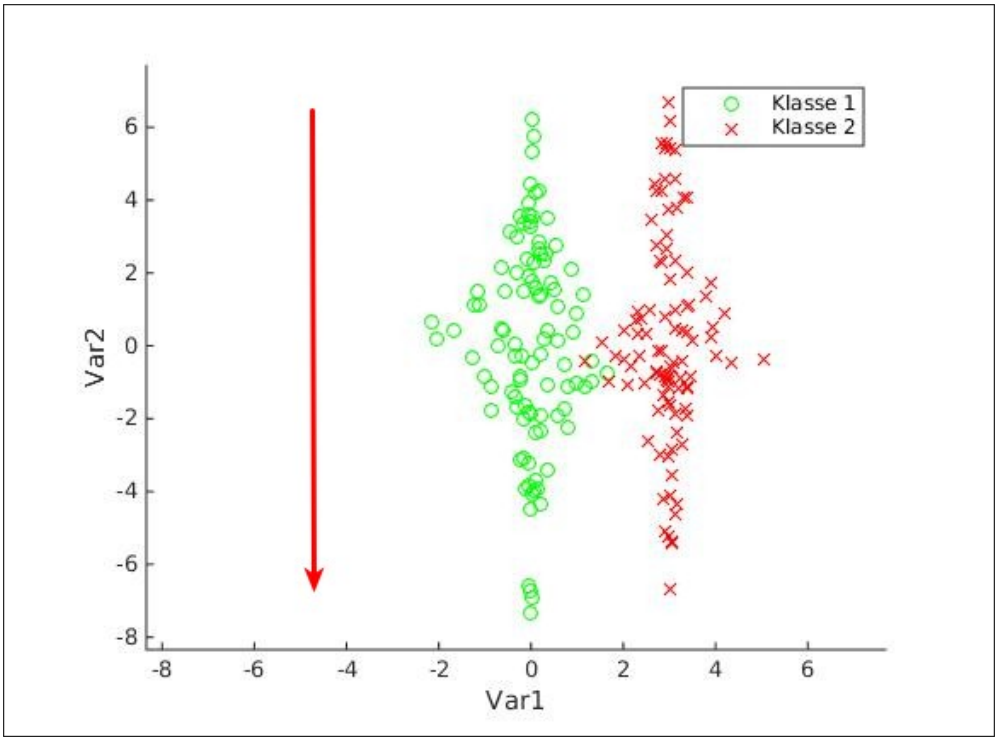

Hier sind die Daten verstreut in einer Punktemenge, welche einer Gerade angenähert gleicht. Es würde sich also PCA anbieten, da die Abweichungen zu dieser Gerade (roter Pfeil) nicht allzu groß sind in den meisten Fällen. Für die Klassifizierung wäre es allerdings auch nicht förderlich.




Hier haben wir große interne Streuungen, welche die Cluster beide besitzen. Ein PCA wäre somit nicht förderlich. Die Reduktion auf die erste Komponente würde auch die weitere Klassifizierung beeinträchtigen.




Hier wäre ein PCA angebracht, da die Streuung wie bei (a), somit könnte man auf die erste Komponente reduzieren. Es ist allerdings auf den ersten Anblick nicht förderlich für die Klassifizierung.




Siehe (b)