

Maschinelles Lernen und Data Mining
Sommersemester 2008
Übungsblatt 8

Besprechung des Übungsblattes am 10.07.2008

Aufgabe 8-1 PCA
schriftlich

- Beschreiben Sie, unter welchen Umständen eine PCA sinnvoll ist bzw. welchen Zweck die PCA verfolgt.
- Welche möglichen negativen Auswirkungen nehmen Sie in Kauf, wenn Sie die PCA auf einen Datensatz unbekannter Struktur verwenden?

Aufgabe 8-2 PCA
schriftlich

Eine PCA besteht im wesentlichen aus:

Mittelwertbereinigung der Daten, Kovarianzmatrix $E[(X - E(X)) \cdot (X - E(X))^T]$ berechnen, Eigenwerte und Eigenvektoren berechnen, Entfernen der Eigenvektoren mit kleinem Einfluss, Transformation der Eingangsdaten.

Gegeben eine normalverteilte Zufallsvariable

$$X \sim N(\mu, \Sigma), \quad \mu = \begin{pmatrix} 0 \\ 3 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}$$

wobei Σ die Kovarianzmatrix der mittelwertbereinigten Daten, und X die 2dimensionale Designmatrix darstellt.

- Berechnen Sie die Hauptkomponenten von Σ sowie die zugehörigen Eigenwerte und beschreiben Sie, welche Hauptkomponente bei einer Dimensionsreduktion entfallen würde.
- Berechnen Sie nun die Hauptkomponenten und Eigenwerte der unzentrierten Kovarianzmatrix $E(XX^T)$ und vergleichen Sie das Ergebnis mit dem Ergebnis aus Teilaufgabe (a)
Verwenden Sie dazu: $X = (X - \mu) + \mu$ mit $E(X - \mu) = 0$

Anmerkungen:

Ist eine Zufallsvariable $X \sim N(0, \Sigma)$ verteilt, gilt $E(XX^T) = \Sigma$.

Der Erwartungswert E ist linear. D.h. es gilt:

$$E(aX + Y) = a \cdot E(x) + E(Y)$$

Damit gilt auch $E(MX) = ME(X)$, wenn M keine Zufallsvariable ist. Andernfalls gilt $E(MX) = E(M)E(X)$, wenn X und M voneinander unabhängig sind.

Eigenwerte und Eigenvektoren können mit gängigen Softwarepaketen (R, Octave, Maple) einfach berechnet werden, dürfen aber natürlich auch von Hand berechnet werden.

Aufgabe 8-3 PCA
schriftlich

Gegeben ist die folgende Designmatrix X:

1	2	3	4	5	6
1	1	1	5	5	5

Führen Sie eine PCA auf den gegebenen Daten durch.

Geben Sie dabei die Eigenvektoren, Eigenwerte und die Kovarianzmatrix an und visualisieren Sie die Daten vor und nach der PCA.