Übungsblatt – Principal Component Analysis (PCA)

Aufgabe 1: Konzeptuelles Verständnis

a) Benenne Sie ein Problem, das bei hochdimensionalen Daten auftreten kann. Erklären Sie, wie die PCA dabei hilft.

b) Beschreiben Sie mit eigenen Worten die Funktionsweise der PCA

c) Warum müssen die Daten vor der PCA zentriert werden? Erläutern Sie, welches Problem auftritt, wenn man diesen Schritt überspringt.

d) Wie viele PCs wählt man für 90% Varianzerhalt, wenn die Eigenwerte 4, 2.5, 1.8, 0.9, 0.5 und 0.3 sind?

Aufgabe 2: Kovarianz und Kovarianz-Matrix

Gegeben sei folgender Datensatz:

a	4	-6	10	4
b	6	-4	8	6

a) Berechnen Sie die jeweiligen Mittelwerte $\bar{a}$ und $\bar{b}$ der Daten. Zentrieren Sie dann die Daten, indem Sie die jeweiligen Mittelwerte von den einzelnen Beobachtungen abziehen.

b) Die Kovarianz zweier Stichproben a und b mit jeweils i Beobachtungen berechnet sich wie folgt:

$$cov(a,b) = \frac{1}{n-1}\sum_{i=1}^{n}(a_i - \bar{a})(b_i - \bar{b}).$$

Berechnen Sie die Kovarianz zwischen a und b.

c) Die Varianz von a ist: $var(a) = 33$, die Varianz von b ist: $var(b) = 22$. Geben Sie die Kovarianzmatrix an.

d) Angenommen die Variable a hätte invertierte Vorzeichen, also $a = -4, 6, -10, -4$. Überlegen Sie was das für die Beziehung zwischen den Variablen bedeutet und welchen Einfluss das auf die Kovarianz hätte.

Aufgabe 3: Anwendung auf Brustkrebs-Daten

Im unten verlinkten GitHub finden Sie im Ordner `02_exercise` das Jupyter Notebook `PCA_breast-cancer-empty.ipynb`. Verwenden Sie dieses Notebook als Arbeitsgrundlage für die Aufgabe.

a) Die Daten werden zu Beginn des Notebooks bereits geladen. Führen Sie auf den Daten eine Hauptkomponenten-Analyse durch indem Sie die Schritte einzeln bearbeiten. Verwenden Sie die Jupyter Notebooks im Ordner `03_examples` als Hilfestellung. Wenn Sie nicht weiter wissen, dann finden Sie im Ordner `02_exercise` auch das Notebook `PCA_breast-cancer.ipynb` in dem die PCA bereits durchgeführt wird.

b) in **Schritt 1** ist eine Variable definiert über die festgelegt wird, ob die Daten standardisiert oder nur zentriert werden. Führen sie die PCA in beiden Versionen durch. Welche Auswirkungen hat diese Einstellung auf die Kovarianz-Matrix, den Scree-Plot und die Anzahl der Hauptkomponenten? Versuchen Sie den Unterschied zu erklären.

Alle Materialen zur Veranstaltung finden Sie auf GitHub:
https://github.com/MartinaEchtenbruck/Principal-Component-Analysis