

Übungsblatt: Principal Component Analysis

Aufgabe 1) Konzeptuelles Verständnis

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

Hochdimensionale Daten sind schwer zu visualisieren, sie benötigen mehr Speicher und mehr Rechenkapazität bei der Verarbeitung.

In höheren Dimensionen verlieren die Abstände zwischen den Datenpunkten ihre Bedeutung, das Rauschen tritt in den Vordergrund.

Die Performance vieler Machine Learning Verfahren leidet.

Die PCA reduziert die Dimensionalität der Daten, indem sie sie auf die wichtigsten Hauptkomponenten projiziert:

- b) Beschreiben Sie mit eigenen Worten die Funktionsweise der PCA.

Die PCA bestimmt für die Daten Hauptkomponenten, die die Daten beschreiben. Diese Hauptkomponenten werden so bestimmt, das gilt:

- Die erste Hauptkomponente bildet die Richtung der größten Varianz ab.
- Die Hauptkomponenten stehen senkrecht zueinander, d.h. sie sind unkorreliert.

Reduktion der Dimension geschieht durch Auswahl der Hauptkomponenten, unter der Annahme, dass die ersten

Hauptkomponenten den Großteil der Varianz (Information) in den Daten erklärt.

c) Warum müssen die Daten vor der PCA zentriert werden?

Erläutern Sie, welches Problem auftritt, wenn man diesen Schritt überspringt.

Die PCA sucht die Richtungen der größten Varianz. Wenn die Daten nicht zentriert werden, dann kann es passieren, dass weit vom Ursprung entfernte Daten als wichtig eingestuft werden und so die erste Hauptkomponente in Richtung dieser Daten zeigt - und nicht in Richtung der größten Varianz.

d) Wie viele Hauptkomponenten wählt man für 80% Varianzerhalt, wenn die Eigenwerte 4, 2,5, 1,8, 0,9, 0,5 und 0,3 sind?

Die Summe der Eigenwerte ist:

$$4 + 2,5 + 1,8 + 0,9 + 0,5 + 0,3 = 10$$

Um 80% Varianzerhalt zu erzielen, müssen also so viele Hauptkomponenten gewählt werden, dass die Summe der zugehörigen Eigenwerte mindestens 8 ergibt.

1 Hauptkomponente: 4 das entspricht 40% erklärter Varianz

2 Hauptkomponenten: $4 + 2,5 = 6,5 \rightarrow 65\%$ "

3 Hauptkomponenten: $4 + 2,5 + 1,8 = 8,3 \rightarrow 83\%$ "

4 Hauptkomponenten: $4 + 2,5 + 1,8 + 0,9 = 9,2 \rightarrow \underline{92\%}$ "

Es müssen mindestens 4 Komponenten gewählt werden.

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.

$$\bar{a} = \frac{1}{4}(4 - 6 + 10 + 4) = \frac{1}{4}(12) = 3$$

$$\bar{b} = \frac{1}{4}(6 - 4 + 8 + 6) = \frac{1}{4}(16) = 4$$

Zentrierung der Daten ergibt dann:

a'	1	-9	7	1
b'	2	-8	4	2

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

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

Berechnen Sie die Kovarianz zwischen a und b.

In a) wurden die Mittelwerte bereits abgezogen, dies kann hier direkt verwendet werden:

$$\begin{aligned} \text{cov}(a, b) &= \frac{1}{3} (1 \cdot 2 + (-9)(-8) + 7 \cdot 4 + 1 \cdot 2) \\ &= \frac{1}{3} (2 + 72 + 28 + 2) \\ &= \frac{1}{3} 104 = 34 \frac{2}{3} \approx \underline{\underline{34,667}} \end{aligned}$$

c) Die Varianz von a ist $\text{var}(a) = 33$, die Varianz von b ist $\text{var}(b) = 22$.

Geben Sie die Kovarianzmatrix des Datensatzes an.

$$\text{cov}(a,b) = 34,667 \quad \text{var}(a) = 33 \quad \text{var}(b) = 22$$

Kovarianz-Matrix: - Hauptdiagonale enthält die Varianzen

- übrige Positionen enthalten Kovarianzen

$$\Sigma = \begin{pmatrix} \text{Var}(a) & \text{cov}(a,b) \\ \text{cov}(a,b) & \text{var}(b) \end{pmatrix} = \begin{pmatrix} 33 & 34,667 \\ 34,667 & 22 \end{pmatrix}$$

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.

Ursprünglich:

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

a invertiert:

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

Die ursprünglichen Variablen haben sich immer ähnlich verändert: wenn a große Werte angenommen hat, dann hat b auch große Werte angenommen. Hat a kleine Werte angenommen, dann hat auch b kleine Werte angenommen.

Nach der Invertierung von a verändern die Variablen sich gegenläufig: b nimmt kleine Werte an, wenn a große Werte annimmt und umgekehrt.

Bei der Berechnung der Kovarianz

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

macht dies den Unterschied aus, ob die Faktoren in der Summe eher dasselbe oder unterschiedliche Faktoren haben.

Die Invertierung des Vorzeichen von a würde also insgesamt die Invertierung der einzelnen Summanden und somit auch eine Änderung des Vorzeichens bei der Kovarianz bedeuten.

Aufgabe 3) Anwendung auf Brustkrebs-Daten

a) Durchführung der PCA auf den Brustkrebs-Daten

Siehe Jupyter-Notebook PCA_breast_cancer.

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 Auswirkung hat diese Einstellung auf die Kovarianz-Matrix, die Ladungs-Matrix und die Anzahl der Hauptkomponenten?

Versuchen Sie, den Unterschied zu erklären.

Die Kovarianz-Matrix ist mit Standardisierung mit Werten im Bereich von etwa -0.3 bis 1 belegt. Viele der Werte liegen dabei etwa im Bereich von 0.4 - 0.6. Auf der Hauptdiagonale sind einsch. Ohne Standardisierung bewegen sich die Einträge der Kovarianzmatrix etwa im Bereich von 0 bis 320.000. Dabei sind nur vier Einträge deutlich rot markiert, die anderen Werte sind nahe null.

Die erste Hauptkomponente wird mit Standardisierung von allen Variablen ein wenig beladen und erklärt 44,3% der Varianz in den Daten. Ohne Standardisierung, nur mit Zentrierung, wird die erste Hauptkomponente hauptsächlich von zwei Variablen beladen: mean area und worst area. Die verbleibenden Variablen sind entweder nahe null oder null. PC1 erklärt 98,2% der Varianz in den Daten. Dies weist eine parallele mit der Kovarianz-Matrix auf, die auch hauptsächlich bei diese Variablen deutliche Werte hatte. Mit Standardisierung werden für 95% erklärter Varianz 10 Hauptkomponenten benötigt - ohne nur eine.

Das Problem liegt in der fehlenden Skalierung der Daten. Wenn man sich die Original-Daten ausgeben lässt, kann man sehen, dass die unterschiedlichen Variablen ein ganz unterschiedlicher Wertebereiche rangieren.

Mean Area hat Werte zwischen: 143,5 - 2501

Worst Area hat Werte zwischen: 185,2 - 4254

Während viele andere Variable im Bereich zwischen 0 und 1 rangieren, nur wenige liegen noch im Bereich bis max. So. Diese riesige Unterschiede in den Wertebereichen verzerrten die Auswertung.