Maschinelles Lernen
Aufgabenblatt 08

Prof. Dr. Christoph Böhm
Hochschule München

6. Juni 2024

Aufgabe 8.1 (Praxisbeispiel PCA)**.** In dieser Aufgabe führen Sie eine Hauptkomponentenanalyse auf Sport-Daten durch. Die Hoffnung ist interpretierbare Komponenten in den Daten zu finden und die Information in den Daten zu komprimieren. Verwenden Sie die Daten zum **Siebenkampf** (Daten von Michael Fröhlich), die auf der Homepage der Veranstaltung verfügbar sind. Wir fokussieren uns auf die 7 Disziplinen, Hochsprung, Weitsprung, 100m Huerden, 200m Lauf, 800m Lauf, Kugelstoßen und Speerwurf.

1. Lesen Sie die Daten in Python ein und visualisieren Sie die Ergebnisse in den 7 Disziplinen. Was fällt Ihnen auf?
2. Löschen Sie, falls nötig, große Ausreißer aus den Daten.
3. Drehen Sie die Variablen 'Zeit_100m_Huerden', 'Zeit_200m_Lauf', 'Zeit_800m_Lauf_Minute' um, sodass auch bei diesen Disziplinen höhere Werte ein besseres Ergebnis bedeuten. (Umdrehen von x durch $x_r = \max(x) - x$)
4. Standardisieren Sie die Daten. Warum erscheint das hier sinnvoll?
5. Berechnen Sie eine PCA auf den standardisierten Ergebnissen der 7 Disziplinen. Verwenden Sie die Funktionen `sklearn.decomposition.PCA` und `PCA.fit`. Interpretieren Sie die Ladungen der ersten drei Hauptkomponenten.
6. Versuchen Sie eine geeignete Anzahl an Hauptkomponenten zu wählen, sodass Sie einen guten Kompromiss haben aus erklärter Streuung und Anzahl an Hauptkomponenten; (Sie können auch eine Faustregel wie 'mindestens 90% der Varianz in den Daten soll durch die gewählten Hauptkomponenten erklärt werden.' verwenden.)
7. Berechnen Sie die mit der PCA transformierten Daten für die ersten drei Hauptkomponenten.
8. **Bonusaufgabe:** Visualisieren Sie die Ergebnisse der PCA geeignet in einem biplot.