Dimensionsreduktion

In diesem Dokument soll beschrieben und gezeigt werden wie ein Datenset verändert werden kann, dass nur noch relevante,unabhängige Variablen und Parameter enthalten sind. Dies dient dazu hochdimensionale Daten besser visualisieren zu können und unnötige Daten (Schuhgröße des Hausmeisters usw.) zu exkludieren.

Feature Selection

`Selection` **bedeutet, dass wir entscheiden welche Features Wichtigsten sind. Entweder manuell suchen, oder suchen lassen**

Feature Extraction

`Extraction` **bedeutet dass neue Features gefunden / erstellt werden. Die Daten werden vom hochdimensionalen Raum einen niedrigdimensioanlen Raum transformiert**

Principal Component Analyse (Dimensionsreduktion) - UNSUPERVISED

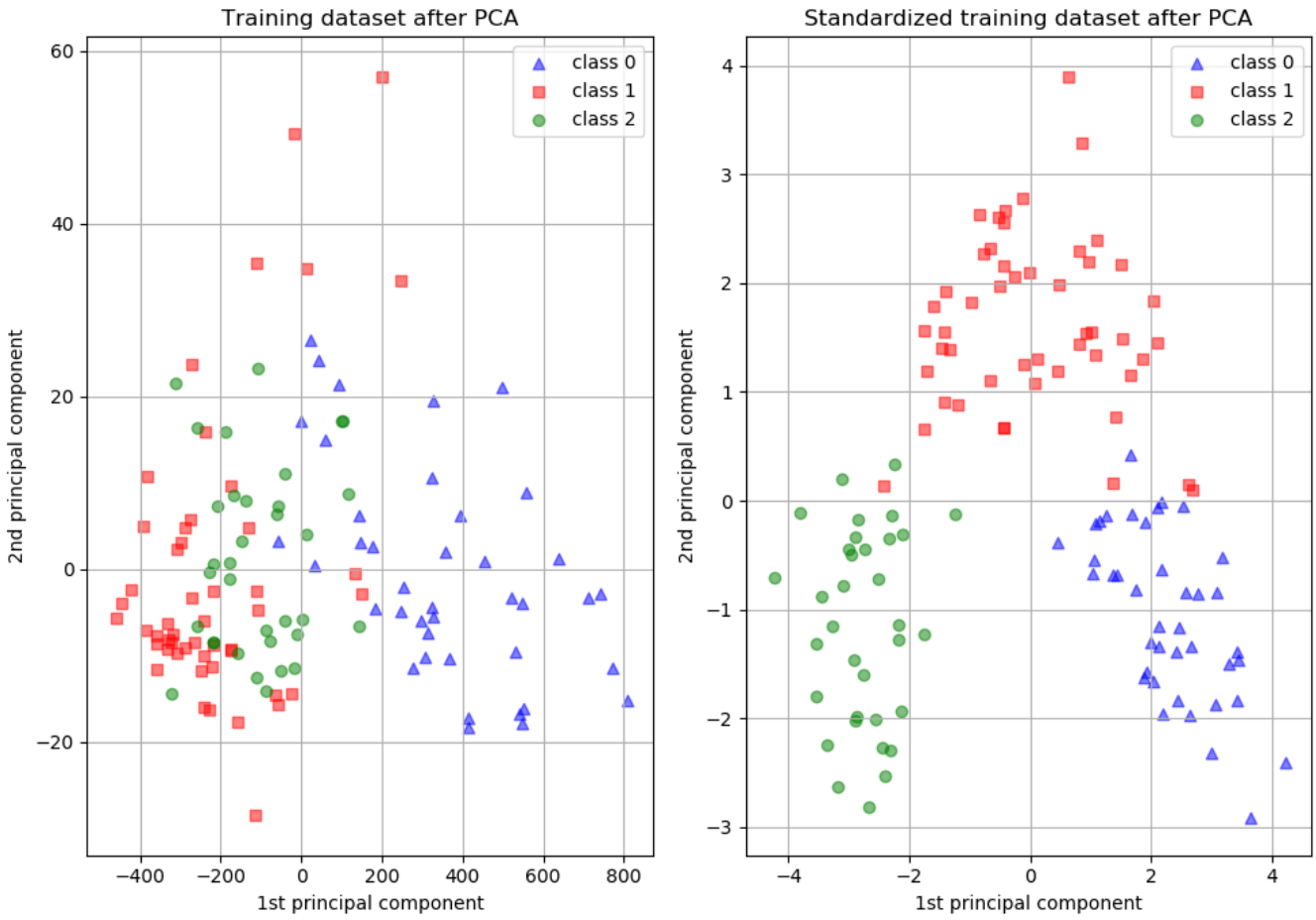
Nehmen wir das Beispiel Nahrungsmittel: Kohlenhydrate, Eiweiß, Fett usw.
Um zu entscheiden ob man wirklich alle Eigenschaften / Dimensionen braucht, macht man eine PCA.
Brennwert wird zum Beispiel aus mehreren der anderen Variablen berechnet. D.h. herrscht eine Abhängigkeit und man kann die entsprechende Dimension weglassen.

Wichtig ist auch die Daten zu skalieren!

Standardisierung der Daten

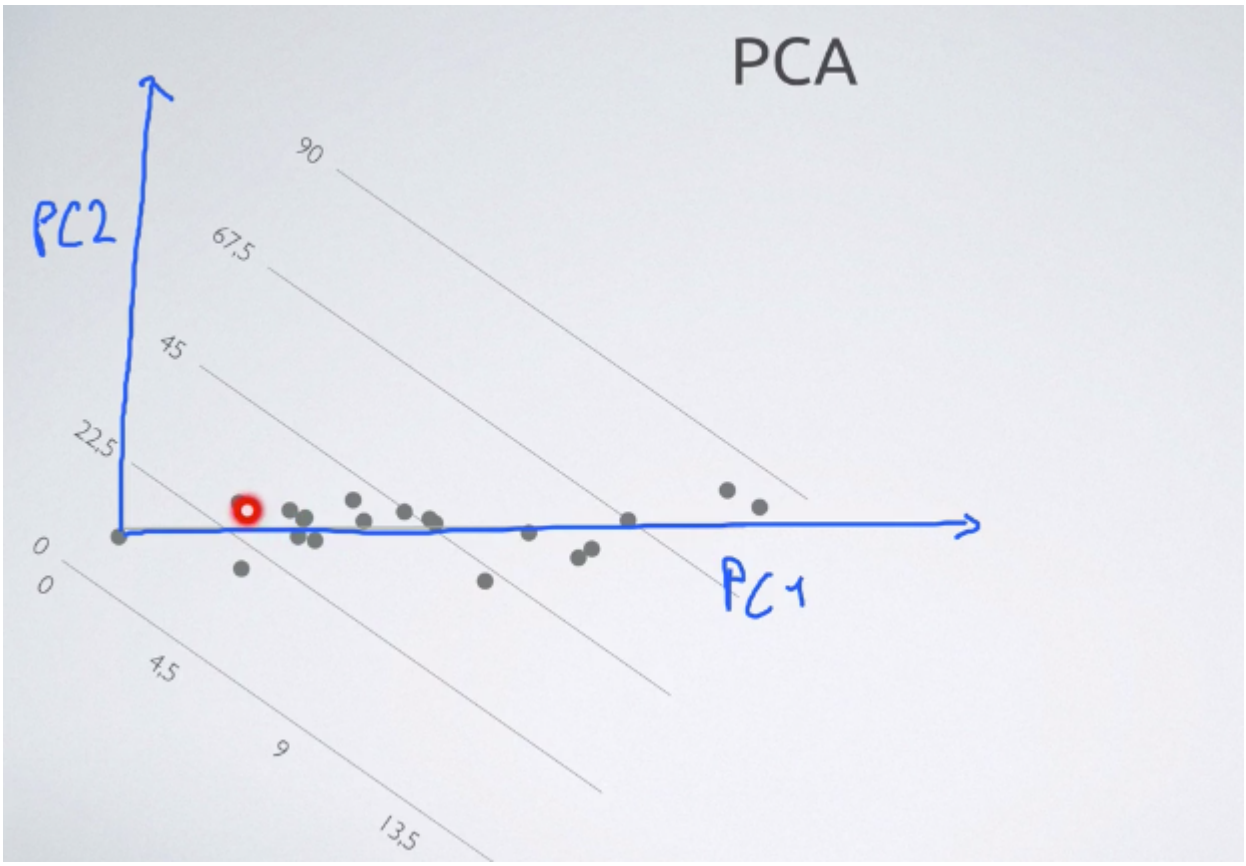
PCA wird durch die Skalierung beeinflusst, so dass Sie die Merkmale in Ihren Daten skalieren müssen, bevor Sie PCA anwenden. Verwenden Sie StandardScaler, um die Funktionen des Datensatzes auf die Einheitsskala zu standardisieren (Mittelwert = 0 und Varianz = 1), was eine Voraussetzung für die optimale Leistung vieler maschineller Lernalgorithmen ist. Wenn Sie den negativen Effekt sehen wollen, dass die Skalierung Ihrer Daten nicht möglich ist, hat scikit-learn einen Abschnitt über die Auswirkungen der Nicht-Standardisierung Ihrer Daten.

Ein gutes Beispiel um zu zeigen wie wichtig Standardidiserung ist, bietet dieses Bild:



Mathematisch: Wir haben k Dimensionen und wollen n<k Dimensionen

Grundsätzlich besitzt PC1 die größte Varianz des gesamten Datensatzes. PC2 die zweitgrößte Varianz etc. Hat man zum Beispiel eine 'normale' Regression zwischen einer X und Y Achse, dreht man einfach die Regressionsgerade zur neuen X-Achse. Somit beschreibt nun die neue X-Achse (PC1) die größte Varianz. Die alte Y-Achse gibt jetzt wieder nur noch die Höhe an, jedoch wird diese nun als PC2 beschrieben. Und wie man in dem Schaubild dann sieht, beschreibt diese Achse eine geringere Varianz (hoch / runter) als die PC1 (links / rechts)



Merksätze:
Hierbei werden neue Dimensionen gefunden, die absteigend möglichst viel Varianz erfassen
Die erste Dimension enthält die meiste Varianz, die zweite dann etwas weniger etc.
Dadurch können wir die Daten "automatisch" komprimieren
Neue Dimensionen werden automatisch gefunden
Können uns angeben lassen, wie viel Varianz enthalten geblieben ist

Achtung: Neue Dimensionen haben keine direkte Bedeutung mehr! PC's sind Kombinationen aus den Daten

PCA un d PCA-Regression

Die PCA braucht im Vergleich zur PCA-Regression eine Vorgabe auf wie viele Hauptkomponenten der Datensatz reduziert werden soll.
Die PCA-Regression dagegen hilft dabei herauszufinden welche Anzahl an Hauptkomponenten am besten geeignet ist.

Hinweis für sich selbst:
Verwenden Sie pca, wenn X korreliert sind. Verwende eine Pipeline mit Gridsearch, um zu sehen, wie sich die Genauigkeit ändert, wenn Sie die Anzahl der Komponenten ändern. Rufen Sie die Ellenbogenmethode auf.

Code

Link zum entsprechenden Notebook Dimensionsreduction in Action