3. Übungsblatt

(Hauptachsen-Transformation)

Nachdem bisher mit artifiziellen Datensätzen gearbeitet wurde, wenden wir uns jetzt realen Daten zu. Dazu verwenden wir den MNIST-Datensatz der Grauwert-Bilder handgeschriebener Ziffern enthält. Die 28×28 Pixel großen Bilder können als 784-dimensionale Merkmalsvektoren aufgefasst werden. Um den Trainingsaufwand gering zu halten, werden im Rahmen dieser Übung lediglich 500 zufällig ausgewählte aus den ingesamt zur Verfügung stehenden 6000 Merkmalsvektoren pro Klasse verwendet.

Folgen Sie den Kommentaren im Quelltext der Module *blatt3.aufg05* und *blatt3.aufg06*.

5. Aufgabe: Einen Klassifikator für 784-dimensionale Merkmalsvektoren zu erstellen erfordert erheblich mehr Trainingsdaten als für die bisher verwendeten zweidimensionalen Beispiele. Insbesonders ist zu erwarten, dass viele der 784 Dimensionen keine relevante Information enthalten oder dass die unterschiedlichen Merkmalsvektorkomponenten stark korrelliert oder sogar linear abhängig sind.

Ein Verfahren zur Dimensionsreduktion ist die Hauptachsen-Transformation oder *Principal Component Analysis* (PCA).

1. Zum besseren Verständnis der Hauptachsen-Transformation schauen Sie sich zunächst ein Spielbeispiel an. An diesem lassen sich die Konzepte einfacher visualisieren und nachvollziehen. Verstehen und erarbeiten Sie dazu den Code der Methode *pca_example*.
2. Laden Sie nun die Daten des MNIST-Datensatzes. Visualisieren Sie einige Ziffernabbilder.
3. Transformieren Sie die 784-dimensionalen Daten des MNIST-Datensatzes in einen geeignet gewählten niedriger-dimensionalen Merkmalsraum. Begründen Sie die gewählte Größe. Den Code, den Sie bei dem Spielbeispiel kennengelernt haben, können Sie wiederverwenden.

   Hinweis: Die Unterraumdimension lässt sich mit der möglichen Rekonstruktionsqualität verknüpfen. Woran lässt sich das messen? *Optional:* Implementieren Sie eine passende Visualisierung.
4. *(optional)* Transformieren Sie die MNIST-Daten in einen zweidimensionalen Unterraum. Visualisieren Sie diese Daten in einem 2D Plot und färben Sie die Datenpunkte nach Klassenzugehörigkeit.

6. Aufgabe: Nun sollen die Daten des MNIST-Datensatzes mit den vorher erstellten Klassifikatoren klassifiziert werden. Trainieren Sie dazu jeweils mit den Trainigsdaten einen Klassifikator und berechnen Sie den sich ergebenden Klassifikationsfehler. Variieren Sie die Parametrisierung der Klassifikatoren, um einen möglichst geringen Klassifikationsfehler zu erreichen. *Optional:* Verwenden Sie dabei eine Kreuzvalidierung.

1. Trainieren Sie einen Mischverteilungsklassifikator für verschiedene mit PCA dimensionsreduzierte Versionen des MNIST-Datensatzes und vergleichen Sie die erzielten Ergebnisse.
2. Trainieren Sie einen k-NN-Klassifikator für verschiedene mit PCA dimensionsreduzierte Versionen des MNIST-Datensatzes und vergleichen Sie die erzielten Ergebnisse.

Wichtig: Beachten Sie die Probleme bei der Auswertung und Bewertung von Verteilungsmodellen (siehe Übungsblatt 2)!