3. Übungszettel Mustererkennung SS15

Prof. Raúl Rojas, Daniel Göhring, Fritz Ulbrich
Institut für Informatik, Freie Universität Berlin
Abgabe Online bis Dienstag, 26.05.15, 24 Uhr

1. Aufgabe (3 Punkte): Multivariate Normalverteilung / Bayes-Klassifikator

Laden Sie die Dateien **pendigits-testing.txt** und **pendigits-training.txt** aus dem Resources-Ordner der KVV Seite zur Vorlesung herunter. Jede Zeile dieser Dateien ist ein Datensatz für einen Linienzug einer Ziffer bestehend aus 17 Zahlen, die durch Leerzeichen getrennt sind. Die ersten 16 Zahlen sind 8 X/Y-Koordinatenpaare. Die letzte Zahl ist die Ziffer, die der Linienzug darstellen soll.

Berechnen Sie die multivariate (mehrdimensionale) **Normalverteilung** (Erwartungswert und Kovarianzmatrix) über dem 16-dimensionalen Koordinatenvektor jeweils für alle 10 Ziffern anhand der Werte aus **pendigits-training.txt**. (Diese Werte bitte nicht mit der Lösung abgeben, das sind mir zuviele Zahlen!)

Klassifizieren Sie die Ziffern in **pendigits-testing.txt** anhand der entsprechenden A-posteriori-Wahrscheinlichkeitsdichtefunktionen. Nehmen dabei Sie eine gleichverteilte A-priori-Wahrscheinlichkeit für jede Ziffer an. **Geben Sie die die Konfusionsmatrix und Klassifikationsgüte aus**.

2. Aufgabe (4 Punkte): PCA / Dimensionsreduktion

a. (1 Punkt) Geben sie die erste Hauptkomponente der Daten in **pendigits-training.txt** an.
b. (3 Punkte) Reduzieren Sie die Dimension des pendigits-Datensatzes mittels einer Hauptkomponentennanalyse (PCA) und klassifizieren die Testdaten anhand der Trainingsdaten mit einem Bayes-Klassifikator (wie Aufgabe 1). Geben Sie die Klassifikationsgüte für verschiedene Anzahlen von Dimensionen von 1 bis 15 aus. (Also einmal für 1D, einmal für 2D, usw. bis 15D mit jeweils den “besten” Hauptkomponenten)

3. Aufgabe (3 Punkte): k-Means

Laden Sie die Datei **clusters.txt** aus dem Resources-Ordner der KVV Seite zur Vorlesung herunter. Jede Zeile dieser Datei entspricht einem X/Y-Koordinatenpaar.

Clustern Sie den Datensatz mit dem k-Means-Algorithmus. (kmeans soll selbst implementiert werden, nicht das kmeans aus der Statistic-Toolobox verwenden!)

Visualisieren Sie die Clusterzentren und Zuordnung der Puńkte der ersten 5 Iterationsschritte mit k=3 (Also insgesamt 5 Bilder)