HAW HAMBURG

Praktikum 1

Aufgabe 1 (Spam-Datensatz):

In der Datei `spamData.mat` befindet sich der klassische Spam-Datensatz. Eine nähere Beschreibung findet man auf https://archive.ics.uci.edu/ml/datasets/Spambase. Die Daten sind schon in Trainings- und Testdaten aufgeteilt.

(a) Spalten Sie den Trainingsdatensatz zufällig im Verhältnis 4:1 in einen Trainings- und einen Validierungsteil auf.

(b) Analysieren Sie den Datensatz mit dem Naive-Bayes-Algorithmus (Matlabs `fitcnb`) mit der Fenstermethode zur Modellierung der class-conditionals.

(i) Für welchen Wert der Fensterbreite erhalten Sie die beste Fehlerrate? Plotten Sie dazu Trainings- und Validierungsfehler in Abhängigkeit der Fensterbreite für Werte zwischen 0 und 1. In welchen Bereichen hat man Über- bzw. Unteranpassung?

(ii) Verifizieren Sie Ihre optimale Klassifizierunsgleistung mit dem Testdatensatz.

(iii) Normieren Sie nun die Merkmale der Trainings-, Validierungs- und Testdaten, d.h. normieren Sie die ensprechenden Datenmatrizen spaltenweise und wiederholen Sie die Analyse. Welche Fehlerraten erreichen Sie jetzt?

(c) Analysieren Sie den Datensatz analog zu Aufgabenteil (b) mit dem kNN-Algorithmus.

Hinweise:

- Benutzen Sie zur Normierung der Validierungs- und Testdaten die Mittelwerte und Standardabweichungen der Trainingsdaten!
- Die Schleife über die Fensterbreite beim Naive-Bayes-Algorithmus kann etwas dauern: Lassen Sie Ihre Schleife daher zuerst für nur über wenige Werte der Fensterbreite laufen. Erst wenn alles richtig ist, erhöhen Sie die Anzahl der Schleifendurchläufe.

Aufgabe 2 (Cars-Datensatz):

In dieser Regressionsaufgabe geht es darum, den Verbrauch von Autos anhand einiger Merkmale wie z.B. Gewicht, Leistung und Hubraum zu bestimmen. Die Datei `cars.csv` enthält die Daten für 392 Automodelle. Die Merkmale in den Spalten der Datei sind:

Spalte	Merkmals-Name	Bedeutung
1	`mpg`	Verbrauch in miles per gallon
2	`cylinders`	Anzahl Zylinder
3	`displacement`	Hubraum in Kubik-Inch
4	`horsepower`	Leistung in PS
5	`weight`	Gewicht in Pfund
6	`acceleration`	Zeit (Sek), um das Auto auf 60 mph (miles per hour) zu beschleunigen
7	`year`	Erscheinungsjahr des Modells
8	`origin`	Herkunft (1. Amerikanisch, 2. Europäisch, 3. Japanisch)
9	`name`	Name des Modells

Merkmal 1 `mpg` ist im Folgenden die Zielgröße und die Variable 9 `name` wird nicht weiter verwendet.

(a) Vorbereitung der Daten:

(i) Laden Sie die Daten mit der Funktion `readTable`.

(ii) Benutzen Sie Funktion `onehotencode`, um aus dem Merkmal `origin` eine dreispaltige Indikatormatrix zu berechnen (siehe Abschnitt A.2.1 im Skript).

(iii) Benutzen Sie die Funktion `table2array`, um aus den Merkmalen `cylinders` bis `year` eine Datenmatrix zu erstellen.

(b) Erstellen Sie aus den Merkmalen 2 bis 8 (für `origin` die Indikatormatrix benutzen) ein lineares Modell der Form $z(x) = b + \sum w_p x_p$, um den Verbrauch `mpg` vorherzusagen. Berechnen Sie die Größe RMSE aus den Trainingsdaten und plotten Sie die Zielgröße `mpg` gegen die vom Modell vorhergesagten Werte z. Plotten Sie außerdem die Abweichungen z - `mpg` gegen `mpg`.

Hinweis: Bei der Lösung der Normalengleichung sehen Sie eine Warnmeldung zu einer fast singulären Matrix. Addieren Sie einen quadratischen Regularisierungsterm mit $\lambda = 0.001$, um diese Warnung zu beseitigen.

(c) (optional) Passen Sie jetzt ein quadratisches Modell an, also ein Modell, das neben den Termen erster Ordnung wie in Teil (b) auch alle quadratischen Terme der Form $x_p \cdot x_q$ enthält. Welchen RMSE-Wert können Sie jetzt erreichen? Plotten Sie auch die Abweichungen z - `mpg` gegen `mpg`.

Hinweis: Schreiben Sie eine Schleife, um die Designmatrix zu füllen.