Mathematik des maschinellen Lernens (WiSe 2021/22)

6. Übung

17. Januar 2022

1. Aufgabe$^{(*)}$
Zeigen Sie, dass eine Stichprobe S, in der alle Merkmale die gleiche Markierung ($+1$ oder -1) haben, stets durch eine Hyperebene linear getrennt werden kann.

2. Aufgabe$^{(*)}$
Vergleichen Sie die vier linearen Methoden zur binären Klassifikation tabellarisch bzgl. ihrer Voraussetzungen, ihrer Hypothesenklassen, der zu minimierenden Risiko- bzw. Zielfunktion, ggf. äquivalente Optimierungsaufgaben, verwendete Verlustfunktion und Eindeutigkeit der Lösung.

3. Programmieraufgabe$^{(*)}$ (Klassifikation von Herzerkrankungen)

a) Laden Sie von

http://archive.ics.uci.edu/ml/datasets/Statlog+%28Heart%29

den Datensatz zur Klassifikation von Patienten mit Herzkrankheiten herunter und laden Sie den Datensatz in MATLAB (*Hinweis:* z. B. mit `readtable()`). Ziehen Sie aus der Datentabelle die sechs reellwertigen Merkmale (siehe Datensatzbeschreibung) sowie die Markierungen (letzte Spalte) heraus.

Hinweis: Der Befehl `T = T{:,:}` überträgt in MATLAB eine Datentabelle in eine reellwertige Matrix.

b) Teilen Sie den Datensatz in zwei Teile: einen *Trainingsdatensatz* (70% der Daten) und einen *Testdatensatz* (30% der Daten).

Hinweis: Sie können dazu `randperm` und `setdiff` nutzen.

c) Nutzen Sie die Trainingsdaten um per *logistischer Regression* eine Hypothese aus $\mathcal{L}_{d,\text{sig}}$ zu lernen. Wie viele der Trainigsdaten werden falsch klassifiziert? Schätzen Sie mittels der Testdaten auch das resultierende erwartete Risikio bzgl. des 0-1-Verlustes.

Ist die Trainingsstichprobe linear trennbar oder nicht? Woran können Sie das ablesen?

d) Lernen Sie per *weicher SVM-Regel* nun auch eine Hypothese $h_S \in \mathcal{L}_d$ aus den Trainingsdaten. Berechnen Sie erneut, wieviel der Trainingsdaten und wieviel der Testdaten falsch klassifiziert werden. Geben Sie eine Schätzung des erwarteten Risikos bzgl. des 0-1-Verlustes an.

e) Variieren Sie für die weiche SVM-Regel den Steuerparameter λ im Bereich 10^{-3} bis 10^{1}. Was stellen Sie fest? Welche der erlernten Hypothese würden Sie für Anwendungsdaten benutzen und warum? Lassen Sie sich zudem den Gewichtsvektor ausgeben und vergleichen Sie diesen mit dem der logistischen Regression.