Übungen zur Vorlesung „Maschinelles Lernen und Künstliche Intelligenz aus Sicht der Stochastik“

Sommersemester 2017, Blatt 5

Abgabetermin: 13.06.2017, spätestens zu Beginn der Vorlesung

Aufgabe 1 (4 Punkte)

Sei $y_i = f(X_i) + \varepsilon_i$ mit unabhängig identisch normalverteilten Fehlern ε_i und $\hat{y}_i$ die Vorhersage der linearer Regression. Zeigen Sie, dass

$$\frac{1}{\sigma_\varepsilon^2} \sum_{i=1}^{N} \mathbb{COV}(\hat{y}_i, y_i)$$

der Anzahl der Parameter (Freiheitsgrade) bei der linearen Regression entspricht.

Aufgabe 2 (4 Punkte)

a) Generieren Sie je 100 Daten von zehn unabhängig identisch normalverteilten Prädiktoren $X_1, \ldots, X_{10}$.

b) Erzeugen Sie 10 Mal aufs Neue 100 Beobachtungen einer normalverteilten Zielvariable y, die unabhängig von allen Prädiktoren ist.

c) Passen Sie Regressionsbäume mit verschieden vielen Endknoten, an diese Daten an und schätzen Sie die effektiven Freiheitsgrade eines Regressionsbaumes in Abhangigkeit zur Anzahl seiner Endknoten mit Hilfe der Formel aus Aufgabe 1.

d) Erstellen Sie eine Grafik die Freiheitsgrade und Endknoten gegenüberstellt.

Aufgabe 3 (4 Punkte)

Zeigen Sie

$$f^\star(x) = \text{argmin}_{f(x)} \mathbb{E}_{Y|x}[e^{-Yf(x)}] = \frac{1}{2} \log \frac{\mathbb{P}(Y=1|x)}{\mathbb{P}(Y=-1|x)}$$

Aufgabe 4 (4 Punkte)

Sei X eine kategoriale Prädiktorvariable mit q Ausprägungen und y eine quantitative Zielvariable. Wir wollen mittels einer einzigen binären Aufteilung dieser Ausprägungen von X eine gute Vorhersage für y treffen.

a) Wie viele Möglichkeiten für eine binäre Aufteilung gibt es insgesamt?

b) Zeigen Sie: Die optimale Aufteilung bei quadratischem Fehler oder auch Gini-Index lässt sich immer dadurch finden, dass die q Klassen nach ihrem mittleren Response $\bar{y}_k$ für $k = 1, \ldots, q$ geordnet werden und der Schnitt entlang dieser Reihe gesetzt wird. Auf wie viele mögliche Schnitte wird dadurch die Auswahl beschränkt?