EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Mathematisch-Naturwissenschaftliche Fakultät

Fachbereich Mathematik

Prof. Dr. Andreas Prohl
Armin Beck

Statistical Learning 1

Summer semester 2024 **Tübingen, 29.04.2024**

Assignment 3

Problem 1
Let the data set $D_n = \{(X_i, Y_i)\}_{i=1}^n$ of i.i.d. random variables be given.
We consider the local averaging estimators

$$m_n(x) = \sum_{i=1}^n \alpha_{n,i}(x) Y_i$$

introduced in the lecture. Specifically, we look at the kernel estimator, i.e., we choose

$$\alpha_{n,i}(x) = \frac{K(\frac{x-X_i}{h})}{\sum_{j=1}^n K(\frac{x-X_j}{h})}$$

as weights, where $K : \mathbb{R}^d \to \mathbb{R}$ is the kernel function and $h > 0$ is the bandwidth.
Show that the kernel estimator

$$m_n(x) = \frac{\sum_{i=1}^n \mathbb{1}_{\{\|\frac{x-X_i}{h}\| \le 1\}} Y_i}{\sum_{i=1}^n \mathbb{1}_{\{\|\frac{x-X_i}{h}\| \le 1\}}} \tag{1}$$

with the naive kernel $K(x) = \mathbb{1}_{\{\|x\| \le 1\}}$ satisfies

$$m_n(x) = \mathrm{argmin}_{c \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^n K\Big(\frac{x - X_i}{h}\Big) |Y_i - c|^2.$$

Hint: Show that for any $c \in \mathbb{R}$,

$$\sum_{i=1}^n K\Big(\frac{x-X_i}{h}\Big)|Y_i - c|^2 = \sum K\Big(\frac{x-X_i}{h}\Big)|Y_i - m_n(x)|^2 + K\Big(\frac{x-X_i}{h}\Big)|m_n(x) - c|^2.$$

Problem 2
Let $\mathcal{P} = \bigcup_{j \in \mathbb{N}} \mathcal{A}_{n,j}$ be a partitioning of $\mathbb{R}^d$ and $D_n = \{(X_i, Y_i)\}_{i=1}^n$ a data set. We are again looking at a local averaging estimator (1). The partitioning estimator uses the weights

$$\alpha_{n,i}(x) = \frac{\mathbb{1}_{\{X_i \in \mathcal{A}_{n,j}(x)\}} Y_i}{\sum_{j=1}^n \mathbb{1}_{\{X_j \in \mathcal{A}_{n,j}(x)\}}},$$

where $\mathcal{A}n, j(x)$ denotes the set $\mathcal{A}n, j$ that contains x.

Show that the partitioning estimator defined by

$$m_n(x) = \frac{\sum_{i=1}^n \mathbb{1}_{\{X_i \in \mathcal{A}_{n,j}(x)\}} Y_i}{\sum_{i=1}^n \mathbb{1}_{\{X_i \in \mathcal{A}_{n,j}(x)\}}} \quad \forall x \in \mathbb{R}^d \tag{2}$$

satisfies

$$m_n = \operatorname{argmin}_{f \in \mathcal{F}_{\mathcal{P}}} \frac{1}{n} \sum_{i=1}^n |f(X_i) - Y_i|^2,$$

where $\mathcal{F}_{\mathcal{P}} = \{\sum_{j \in \mathbb{N}} a_j \mathbb{1}_{\mathcal{A}_j} : a_j \in \mathbb{R}\}$ denotes the space of all piecewiese constant functions on $\mathcal{P}$.
Hint: Let m_n be the partitioning estimate. Show by the aid of **Problem 1** in **Assignment 2** that

$$\sum_{i=1}^n |f(X_i) - Y_i|^2 = \sum_{i=1}^n |f(X_i) - m_n(X_i)|^2 + \sum_{i=1}^n |m_n(X_i) - Y_i|^2 \quad \forall f \in \mathcal{F}_{\mathcal{P}}.$$

Problem 3
Let $Z \sim B(n,p)$ be binomially distributed with parameters $n \in \mathbb{N}$ and $p > 0$. Then

a) $\mathbb{E}\big[\frac{1}{1+Z}\big] \leq \frac{1}{(n+1)p}$,

b) $\mathbb{E}\big[\frac{1}{Z}\mathbb{1}_{Z>0}\big] \leq \frac{2}{(n+1)p}$.

Problem 4
Let the random variable $Z \sim B(n, P)$ be binomially distributed with parameters $n \in \mathbb{N}$ and $p > 0$. Then

$$\mathbb{E}\Big[\frac{1}{Z}\Big] \geq \frac{1}{np}(1 - (1-p)^n)^2.$$

Date of Submission: 06.05.2024 in the mailbox at 12 noon.