EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Mathematisch-Naturwissenschaftliche Fakultät

Fachbereich Mathematik

Prof. Dr. Andreas Prohl
Armin Beck

# Statistical Learning 1

**Summer semester 2024** **Tübingen, 22.04.2024**

## Assignment 2

**Problem 1.** This task generalizes the classification problem from the lecture. In a classification problem, $Y$ can only have a finite number of values in $\mathcal{M} = \{1, ..., M\}$ with $M \geq 2$. Such a classification problem could be, for example, the classification of animals on images (a sequence of pixels). In this case, the values of Y in $\mathcal{M} = \{1, ..., M\}$ could stand for „dog“, „cat“, „mouse“, „elephant“ etc.
In order to classify new data (e.g. images), a function $f^* : \mathbb{R}^d \to \{1, ..., M\}$ is sought such that the error probability is minimum, i.e.,

$$\mathbb{P}[f^*(X) \neq Y] = \min_{f:\mathbb{R}^d \to \{1,...,M\}} \mathbb{P}[f(X) \neq Y]. \tag{1}$$

a) **Show** that (1) is met by Bayes decision

$$f^*(x) = \text{argmax}_{1 \leq k \leq M} \mathbb{P}[Y = k | X = x] \quad \forall x \in \mathbb{R}^d.$$

b) The a posteriori probabilities

$$\mathbb{P}[Y = k | X = x] =: m^{(k)}(x) \quad \text{für } (1 \leq k \leq M).$$

are required for the Bayes estimator. These a posteriori probabilities cannot be calculated directly. Therefore, we approximate the a posteriori probabilities $m^{(k)}$ with the help of an estimator $m_n^{(k)}$, which is based on a data set $D_n = \{(X_j, Y_j)\}_{j=1}^n$. With these estimators, the „plug-in“ estimator is defined as an approximation of Bayes' estimator

$$f^*(x) \approx f_n(x) = \text{argmax}_{1 \leq k \leq M} m_n^{(k)}(x)$$

for all $x \in \mathbb{R}^d$.
**Show** that the error probability of the „plug-in“ estimator satisfies the following estimate

$$\begin{aligned} 0 &\leq \mathbb{P}[f_n(X) \neq Y | D_n] - \mathbb{P}[f^*(X) \neq Y] \\ &\leq \sum_{k=1}^{M} \int_{R^d} |m_n^{(k)}(x) - m^{(k)}(x)| \mu[dx] \\ &\leq \sum_{k=1}^{M} \Big( \int_{\mathbb{R}^d} |m_n^{(k)}(x) - m^{(k)}(x)|^2 \mu[dx] \Big)^{\frac{1}{2}}. \end{aligned}$$

This shows that the error probability of the „plug-in“ estimator is bounded by the errors of the estimators for the a posteriori probabilities.

**Problem 2**. Let $n \in \mathbb{N}$. For $z_1, ..., z_n \in \mathbb{R}$ we consider $\overline{z} := \frac{1}{n}\sum_{i=1}^{n} z_i$.

a) **Show** that

$$\frac{1}{n}\sum_{i=1}^{n} |c - z_i|^2 = |c - \overline{z}|^2 + \frac{1}{n}\sum_{i=1}^{n} |\overline{z} - z_i|^2$$

for all $c \in \mathbb{R}$.

b) **Conclude** from a) that

$$\frac{1}{n}\sum_{i=1}^{n} |\overline{z} - z_i|^2 = \min_{c\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^{n} |c - z_i|^2.$$

**Problem 3**. After looking at the classification problem from a theoretical perspective in the lecture and in problem 1, we here consider a concrete example. Given the data set $D_n = \{(X_i, Y_i)\}_{i=1}^{n}$, where $X_i$ is a $\mathbb{R}^d$-valued random variable and $Y_i$ is a $\mathcal{M}$-valued random variable. In this specific example, $n = 9$, $d = 2$ and $\mathcal{M} = \{0, 1\}$. In the following, we look at a concrete realization $D_n(\tilde{\omega}) = \{(X_i(\tilde{\omega}), Y_i(\tilde{\omega}))\}_{i=1}^{9} \subset \mathbb{R}^2 \times \mathcal{M}$ for a $\tilde{\omega} \in \Omega$ from the underlying probability space

| $X_i(\tilde{\omega})$ | (1,-2) | (2,-2) | $(1,0)$ | $(2,0)$ | $(0,1)$ | $(3,1)$ | $(0,2)$ | $(2,2)$ | $(3,2)$ |
|---|---|---|---|---|---|---|---|---|---|
| $Y_i(\tilde{\omega})$ | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 1 |

.

The Bayes classifier

$$f^*(x) = \begin{cases} 1, & \text{falls } \mathbb{P}[Y = 1|X = x] \geq \frac{1}{2} \\ 0, & \text{sonst} \end{cases}$$

is used to classify further data points. This involves the a posteriori probability, which is not available. Therefore, it will be estimated using the following estimator

$$\mathbb{P}[Y = 1|X = x] \approx m_9(x) = \frac{\sum_{i=1}^{9} \mathbb{1}_{\{\|X_i(\tilde{\omega})-x\|\leq 2\}} Y_i(\tilde{\omega})}{\sum_{i=1}^{n} \mathbb{1}_{\{\|X_i(\tilde{\omega})-x\|\leq 2\}}} \quad \forall x \in \mathbb{R}^2. \tag{2}$$

This yields the „plug-in“ estimator, which provides approximate values for the Bayes' estimator

$$f_9^* = \begin{cases} 1, & \text{if } m_9(x) \geq \frac{1}{2} \\ 0, & \text{else} \end{cases} .$$

a) Using the „plug-in“ estimator $f_9^*$, classify the data points $x = (1{,}5\ ,\ 2{,}5)$ and $(0, 0)$.

b) Draw a sketch showing the areas in which the „plug-in“ estimator estimates the value „$0$“ and in which it estimates the value „$1$ “.

**Date of Submission: 29.04.2024 in the mailbox at 12 noon.**