Classification

Bernhard Stankewitz
bernhard.stankewitz@posteo.de

March 5, 2019

Abstract

Summary of elemetary results for classification.

Contents

1 Elementary definitions and the Bayes classifier

Definition 1.1 (Classifier).

(a) For i.i.d. training data $(X_1, Y_1), \dots, (X_n, Y_n) \in \mathbb{R}^d \times \{1, \dots, K\}$, a classifier is a measurable function $C : \mathbb{R}^d \to \{1, \dots, K\}$. The classification error is given by

$$R(C) := \mathbb{P}\{C(X) \neq Y\} = \mathbb{E}\mathbf{1}_{C(X) \neq Y}.$$

(b) In case that the labels are given by $\{0, 1\}$, the classification error

$$R(C) = \mathbb{E}(Y - C(X))^2 = \int \mathbf{1}_{y \neq C(x)} \, \mathbb{P}^{X,Y}(d(x, y)).$$

In case all theoretical quantities are known, a classification problem has an optimal solution.

Proposition 1.2 (Bayes-Classifier).

(i) In the situation of Definition 1.1, the classification error is minimised by the Bayes classifier

$$C^{Bayes}(x) := \underset{k=1,\dots,K}{\operatorname{argmax}} \, \mathbb{P}\{Y = k | X = x\}.$$

(ii) If the labels are given by $\{0, 1\}$, we have

$$C^{Bayes}(x) = \mathbf{1}\{\eta(x) \geq 1/2\} \qquad \textit{with} \qquad \eta(x) := \mathbb{P}\{Y = 1 | X = x\}.$$

Proof. For any classifier C, we have

$$R(C) = 1 - \mathbb{E}\mathbb{E}\mathbf{1}_{C=Y}|X = 1 - \mathbb{E}\sum_{k=1}^{K} \mathbb{E}(\mathbf{1}_{C=k}\mathbf{1}_{Y=k}|X) = 1 - \mathbb{E}\sum_{k=1}^{K} \mathbf{1}_{C=k}\mathbb{E}(\mathbf{1}_{Y=k}|X).$$

□

2 The KNN-classifier

Definition 2.1 (KNN-classifier).

(a) Let $(X_1, Y_1), \dots, (X_n, Y_n) \in \mathbb{R}^d \times \{1, \dots, J\}$ be a training sample and $K \in \mathbb{N}$. For $x \in \mathbb{R}^d$, let $N_K(x)$ be the set of the K nearest neighbours of x with respect to the euclidean distance. Then, the KNN-classifier is given by

$$\hat{C}^{\text{KNN}}(x) := \underset{j=1,\dots,J}{\operatorname{argmax}} \frac{1}{K} \sum_{X_i \in N_K(x)} \mathbf{1}_{Y_i = j}.$$

(b) In case, the labels are given by $\{0, 1\}$, we have

$$\hat{C}^{\text{KNN}}(x) = \mathbf{1}\{\hat{\eta}(x) \geq 1/2\} \qquad \text{with} \qquad \hat{\eta}(x) := \frac{1}{K} \sum_{X_i \in N_K(x)} \mathbf{1}_{Y_i = 1} =: \sum_{i=1}^{n} w_i(x) Y_i,$$

where $w_i := \mathbf{1}_{X_i \in N_K(x)}/K$ with $\sum_{i=1}^n w_i = 1$.

Lemma 2.2 (Reduction to the regression function). *In the situation of Definition 2.1 (b), we have that*

$$|\mathbb{E}_{\leq n} R(\hat{C}^{KNN}) - R(C^{Bayes})| \leq \sqrt{\mathbb{E}_{n+1} |\hat{\eta}(X) - \eta(X)|^2}$$

Proof. For any classifier C, we have

$$\mathbb{P}\{C(X) = Y | X\} = \mathbf{1}_{C=1}\eta + \mathbf{1}_{C=0}(1 - \eta) = \eta + \mathbf{1}_{C=0}(1 - 2\eta).$$

This yields

$$\begin{aligned}
|\mathbb{P}\{\hat{C}^{\text{KNN}}(X) \neq Y | X\} - \mathbb{P}\{\hat{C}^{\text{Bayes}}(X) \neq Y | X\}| &= |\mathbb{P}\{\hat{C}^{\text{KNN}}(X) = Y | X\} - \mathbb{P}\{\hat{C}^{\text{Bayes}}(X) = Y | X\}| \\
&= |\hat{\eta} + \mathbf{1}_{\hat{C}^{\text{KNN}}=0}(1 - 2\hat{\eta}) - \eta + \mathbf{1}_{C^{\text{Bayes}}=0}(1 - 2\eta)| \\
&= \begin{cases} |\hat{\eta} - \eta|, & \hat{C}^{\text{KNN}} = C^{\text{Bayes}}, \\ |1 - \hat{\eta} - \eta|, & \hat{C}^{\text{KNN}} = C^{\text{Bayes}}. \end{cases} \\
&\leq \begin{cases} |\hat{\eta} - \eta|, & \hat{C}^{\text{KNN}} = C^{\text{Bayes}}, \\ |1/2 - \hat{\eta} \wedge \eta|, & \hat{C}^{\text{KNN}} = C^{\text{Bayes}}. \end{cases} \leq |\hat{\eta} - \eta|.
\end{aligned}$$

For the second to last step, use that one of η and $\hat{\eta}$ has to be above and one below $1/2$. By conditioning on X and Jensen's inequality, this gives

$$|\mathbb{E}_{\leq n} R(\hat{C}^{\text{KNN}}) - R(C^{\text{Bayes}})|^2 = |\mathbb{E}_{\leq n+1}(\mathbf{1}_{\hat{C}^{\text{KNN}} \neq Y} - \mathbf{1}_{C^{\text{Bayes}} \neq Y})|^2 \leq \mathbb{E}_{\leq n+1} |\hat{\eta} - \eta|^2.$$

$\square$

Theorem 2.3 (Consistency of KNN). *In the situation of Definition 2.1 (b), let $k \to \infty$, $k/n \to 0$ and let $x \mapsto \eta(x)$ be uniformly continuous. Then, the KNN-classifier $\hat{C}^{KNN}$ is consistent, i.e.*

$$|\mathbb{E}_{\leq n} R(\hat{C}^{KNN}) - R(C^{Bayes})| \xrightarrow{n \to \infty} 0.$$

Proof. From Lemma 2.2, we obtain with the triangle inequality

$$\begin{aligned}
|\mathbb{E}_{\leq n} R(\hat{C}^{\text{KNN}}) - R(C^{\text{Bayes}})| &\leq \sqrt{\mathbb{E}_{\leq n+1} |\hat{\eta}(X) - \eta(X)|^2} = \sqrt{\mathbb{E}_{\leq n+1} \Big|\sum_{i=1}^n w_i(X)(Y_i - \eta(X))\Big|^2} \\
&\leq \sqrt{\mathbb{E}_{\leq n+1} \Big|\sum_{i=1}^n w_i(X)(Y_i - \eta(X_i))\Big|^2} + \sqrt{\mathbb{E}_{\leq n+1} \Big|\sum_{i=1}^n w_i(X)(\eta(X_i) - \eta(X))\Big|^2}
\end{aligned}$$

In the following, we consider the two terms separately.

For the first term, we have by independence

$$\begin{aligned}
\mathbb{E}_{\leq n+1} \Big|\sum_{i=1}^n w_i(X)(Y_i - \eta(X_i))\Big|^2 &= \mathbb{E}_{\leq n+1} \sum_{i=1}^n w_i(X)^2 (Y_i - \eta(X_i))^2 \\
&\leq \mathbb{E}_{\leq n+1} \Big(\max_{i \leq n} w_i(X) \sum_{i=1}^n w_i(X) \Big) \leq 1/K \xrightarrow{n \to \infty} 0.
\end{aligned}$$

For the second term, we have for any $\varepsilon > 0$ and corresponding $\delta > 0$, that

$$\begin{aligned}
\mathbb{E}_{\leq n+1}|\sum_{i=1}^{n} w_i(X)(\eta(X_i) - \eta(X))^2| &\leq \mathbb{E}_{\leq n+1}|\sum_{i=1}^{n} w_i(X)\mathbf{1}_{|X_i - X| \geq \delta} + \varepsilon| \\
&\leq \mathbb{E}_{\leq n+1}|K^{-1}\sum_{i=1}^{K} \mathbf{1}_{\sum_{i=1}^{n}\mathbf{1}_{|X_i - X| < \delta} \leq i} + \varepsilon \\
&\leq \mathbb{P}\Big\{\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{|X_i - X| < \delta} \leq K/n\Big\} + \varepsilon,
\end{aligned}$$

since in case that the **i**-th neighbour has a distance of more than δ of X, at most i observations are this close to X. (W.l.o.g. order the (X_i)). That the last term converges to 0 follows from dominated convergence. □