Solution approach and numerical results

Florian Thaler

Graz, 2024

Contents

1 Bayes classification

Let $X = (X_1, \dots, X_d)$ be a $\mathbb{R}^d$-valued discrete random vector and let Y be a real-valued discrete random variable taking values in C. Given a realisation $x = (x_1, \dots, x_d) \in \mathbb{R}^d$, we aim to classify x w.r.t. the classes $y \in C$ according to the policy

$$\arg\max_{y \in C} P(Y = y | X_1 = x_1, \dots, X_n = x_n). \tag{1}$$

To simplify this decision rule, we make the following assumptions

Assumption 1.1 *For all $j \in \{1, \dots, d\}$ the random variables $X_j, \dots, X_d$ are conditional independent given Y, i.e. for all sets $B_j, \dots, B_d \subset \mathbb{R}$ and $D \subset C$ it holds*

$$\mathcal{P}(X_j \in B_j | X_{j+1} \in B_{j+1}, \dots, X_n \in B_n, Y \in D) = \mathcal{P}(X_j \in B_j | Y \in D).$$

Assumption 1.2 *For all sets $B_1, \dots, B_d \subset \mathbb{R}$, and all $D \subset C$ the joint probability $\mathcal{P}(X_1 \in B_1, \dots, X_d \in B_d, Y \in D)$ is larger than zero.*

Let $B_1, \dots, B_d \subset \mathbb{R}$, and let $D \subset C$. According to Bayes' Theorem we get

$$\begin{aligned}
&\mathcal{P}(Y \in D, X_1 \in B_1, \dots, X_d \in B_d) \\
&\quad = \mathcal{P}(X_1 \in B_1 | X_2 \in B_2, \dots, X_d \in B_d, Y \in D) \\
&\qquad \cdot \mathcal{P}(X_2 \in B_2, \dots, X_d \in B_d, Y \in D) \\
&\quad = \mathcal{P}(X_1 \in B_1 | X_2 \in B_2, \dots, X_d \in B_d, Y \in D) \\
&\qquad \cdot \mathcal{P}(X_2 \in B_2 | X_3 \in B_3, \dots, X_d \in B_d, Y \in D) \\
&\qquad \cdot \mathcal{P}(X_3 \in B_3, \dots, X_d \in B_d, Y \in D).
\end{aligned}$$

After finitely many steps we obtain

$$\begin{aligned}
&\mathcal{P}(Y \in D, X_1 \in B_1, \dots, X_d \in B_d) \\
&\quad = (\prod_{j=1}^{d-1} \mathcal{P}(X_j \in B_j | X_{j+1} \in B_{j+1}, \dots, X_d \in B_d, Y \in D)) \\
&\qquad \cdot \mathcal{P}(X_d \in B_d | Y \in D) \cdot \mathcal{P}(Y \in D).
\end{aligned} \tag{2}$$

Using Assumption 1.1, equation (2) reads as follows

$$\begin{aligned}
&\mathcal{P}(Y \in D, X_1 \in B_1, \dots, X_n \in B_d) \\
&\quad = \mathcal{P}(Y \in D) \prod_{j=1}^{d} \mathcal{P}(X_j \in B_j | Y \in D).
\end{aligned}$$

Let $\alpha(B_1, \dots, B_d, D) = \mathcal{P}(X_1 \in B_1, \dots, X_d \in B_n, Y \in D)$. Using Assumption 1.2 we get

$$\begin{aligned}
&\mathcal{P}(Y \in D | X_1 \in B_1, \dots, X_d \in B_d) \\
&\quad = \alpha(B_1, \dots, B_d, C)^{-1} \mathcal{P}(Y \in D) \prod_{j=1}^{d} \mathcal{P}(X_j \in B_j | Y \in D).
\end{aligned}$$

Consequently we can rewrite the decision rule (1) as follows

$$\arg\max_{y \in C} \mathcal{P}(Y = y) \prod_{j=1}^{d} \mathcal{P}(X_j = x_j | Y = y). \tag{3}$$

1.1 Bernoulli naive Bayes

Let us assume that for all $1 \leq j \leq d$, all $y \in C$ and all $t \in \{0, 1\}$ it holds

$$\mathcal{P}(X_j = t | Y = y) = p_{y,j}^{t} (1 - p_{y,t})^{1-t},$$

where $p_{y,j} \in [0, 1]$ refers to the probability of the conditional event $X_j = t$, given $Y = y$. Then we have

$$\begin{aligned} &\mathcal{P}(Y = y) \prod_{j=1}^{d} \mathcal{P}(X_j = x_j | Y = y) \\ &\quad = \mathcal{P}(Y = y) \prod_{j=1}^{d} p_{y,j}^{x_j} (1 - p_{y,t})^{1-x_j}. \end{aligned}$$

The decision rule (3) reads then

$$\arg\max_{y \in C} \mathcal{P}(Y = y) \prod_{j=1}^{d} p_{y,j}^{x_j} (1 - p_{y,t})^{1-x_j},$$

or equivalently

$$\arg\max_{y \in C} \log(\mathcal{P}(Y = y)) + \sum_{j=1}^{d} x_j \log(p_{y,j}) + (1 - x_j) \log(1 - p_{y,t}). \tag{4}$$

1.2 Towards a numerical method

To use decision rule (4) in a concrete scenario we need to compute or approximate for every y the probabilities $\mathcal{P}(Y = y)$ and for every j and every y the success probabilities $p_{y,j}$. Assuming that we have realisations of iid random variables $X^{(1)}, \ldots, X^{(n)} \sim \mathcal{P}_X$ we approximate these terms by means of appropriate relative frequencies. To avoid the so called zero probability problem, we apply Laplace smoothing to obtain approximations of $p_{y,j}$.

2 Problem solution

To tackle the classification problem specified in the task description, we use the approach discussed in Section 1. To apply this methodology we discretise the problem as follows:

1. Let $m, n \in \mathbb{N}$ and let $0 = t_1 < t_2 < \ldots < t_m = 1$, $0 = s_1 < s_2 < \ldots < s_n = 1$ be partitions of $[0,1]$. For every $1 \leq i \leq m-1$, $1 \leq j \leq n-1$ define

$$Q_{i,j} = \begin{cases} [t_i, t_{i+1}) \times [s_j, s_{j+1}) & \text{if } 1 \leq i < m-1, 1 \leq j < n-1, \\ [t_i, t_{i+1}] \times [s_j, s_{j+1}] & \text{else} \end{cases}.$$

2. Let $I = [0,1]^2$, let $x \in I$ and let $N = m \cdot n$. Let us denote by $e_k \in \mathbb{R}^N$ the k-th canonical basis vector in $\mathbb{R}^N$. Let $\varphi : I \to \mathbb{N}$ defined via

$$\varphi(x) = i \cdot j,$$

where i, j are such that $x \in Q_{i.j}$.

Let $L \in \mathbb{N}$ and let $S \subset I$. Let further $\{(x_l, y_l) \in I \times C\}_{l=1}^L$ be the set of labeled data as in the task description, where $C = \{0, 1\}$ and the labels y_l are assigned according to the rule

$$y_l = \begin{cases} 1 & \text{if } x_l \in S, \\ 0 & \text{else} \end{cases}.$$

Then we apply the Bernoulli naive Bayes classification scheme onto the set $\{(e_{\varphi(x_l)}, y_l) \in \mathbb{R}^N \times C\}_{l=1}^L$ to solve the addressed problem.

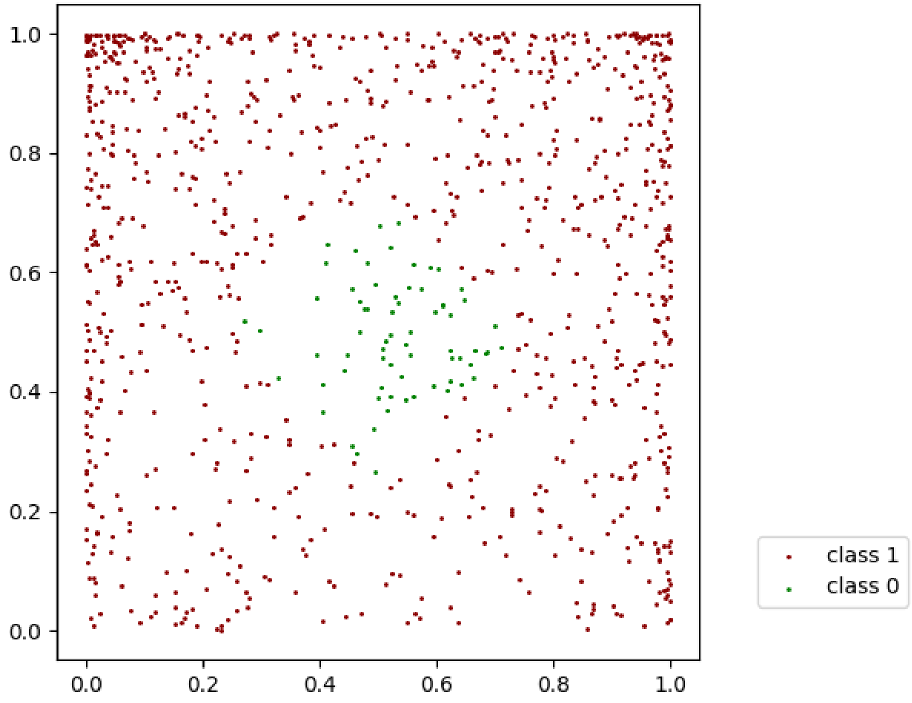

Figure 1: Training dataset consisting of 1000 samples.

3 Numerical results

Let $U_1, U_2 \sim \mathcal{U}(0,1)$ be independently distributed random variables and let $f_1, f_2 : \diagdown\Game\lessdot\sim \to \mathbb{R}$ be measurable functions. Let $\mathcal{Q}$ the distribution of $(f_1(X_1), f_2(X_2))$ on $[0,1]^2$.

Let us consider the functions $f_1(x) = (1 + \cos(\pi x))/2$, $f_2(x) = \sin(\pi x)$, $x \in \mathbb{R}$. Let S be the set defined as

$$S = \{(x_1, x_2) \in \mathbb{R}^2 : \|(x_1 - \frac{1}{2}, x_2 - \frac{1}{2})\|_1 \leq \frac{1}{4}\}.$$

Now, for training consider a randomly sampled training set of size $N = 1000$, and for $n = m = 6$, consider equidistant partitions of the intervals $[0,1]$. We evaluate the performance of the trained classifier using the metrics *accuracy (acc)*, *true positive rate (tpr)* and *false negative rate (fnr)*. Using a randomly sampled evaluation dataset consisting of 1500 samples we get

$$acc = 0.956, \;\; tpr = 0.9835, \;\; fpr = 0.0165.$$