Machine Learning @ Data science 2019: Set 2

Jordi Morera Serra

February 15, 2019

Exercise 5:

Compute $\eta(x)$

Given the $F_{X|Y=0} = \mathbf{P}(X \leq x|Y = 0)$ and $F_{X|Y=1} = \mathbf{P}(X \leq x|Y = 1)$, we can easily compute $f_{X|Y=0} = dF_{X|Y=0}$ and $f_{X|Y=1} = dF_{X|Y=1}$. Applying Bayes law we get $\eta(x) = \mathbf{P}(Y = 1|X = x)$:

$$\eta(x) = \mathbf{P}(Y = 1|X = x) = \frac{f_{X|Y=1}\mathbf{P}(Y = 1)}{f_X} = \frac{f_{X|Y=1}\mathbf{P}(Y = 1)}{f_{X|Y=1}\mathbf{P}(Y = 1) + f_{X|Y=0}\mathbf{P}(Y = 0)}$$

As stated, $\mathbf{P}(Y = 1) = \mathbf{P}(Y = 0)$, hence $\eta(x)$ ends up being quotient of conditional probability density functions. We obtain:

$$\eta(x) = \begin{cases} 0 & x \leq 1 \\ \frac{1}{1+x} & 1 < x \leq 2 \\ 1 & 2 < x \leq 3 \end{cases} \qquad f_x = \begin{cases} \frac{x}{4} & 0 < x \leq 1 \\ \frac{x+1}{4} & 1 < x \leq 2 \\ \frac{1}{4} & 2 < x \leq 3 \end{cases} \tag{1}$$

Note that X is only defined from $0 < x \leq 3$. In figure 1 we can see this functions plotted for valid x.

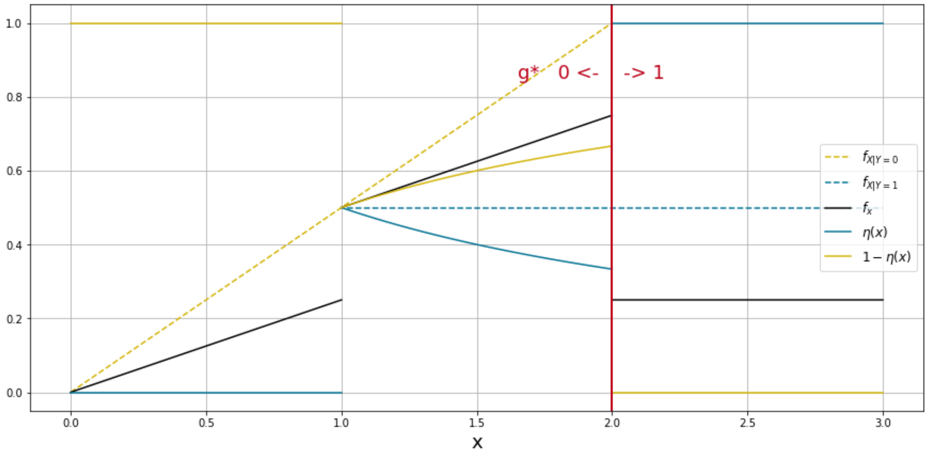


Figure 1: Conditional probability density functions

Bayes classifier and Bayes Risk

We define the loss function $l : l(g(x), Y) = \mathbb{1}_{g(x)\neq Y}$ where $g(x) \in \{0, 1\}$.Given l, one can show $R^* = \mathbb{E}\big[min\big(\eta(x), 1 - \eta(x)\big)\big]$.

If we analyze $\eta(x)$, for all possible values of x, we get:

$$R^* = \begin{cases} \mathbb{E}\big[\eta(x)\big] = 0 & x \leq 1 \\ \mathbb{E}\big[min\big(\eta(x), 1 - \eta(x)\big)\big] = \mathbb{E}\big[min\big(\frac{1}{1+x}, \frac{x}{1+x}\big)\big] = \mathbb{E}\big[\big(\frac{1}{1+x}\big)\big] = \frac{1}{4} & 1 < x \leq 2 \\ \mathbb{E}\big[1 - \eta(x)\big] = 0 & 2 < x \leq 3 \end{cases}$$

For $x \leq 1$ and $2 < x \leq 3$, the two conditional distributions $f_{X|Y=1}$ and $f_{X|Y=0}$ are not overlapping and hence the risk of misclassification is 0. For $1 < x \leq 2$ the conditional distributions are overlapped and the conditional probability of predicting $Y = 1$ is higher than $Y = 0$. Therefore, the risk of misclassification will be minimized when Y is classified as 0.

The Bayes classifier g^* that minimize the risk R is:

$$g^* = \begin{cases} 0 & x \leq 2 \\ 1 & x > 2 \end{cases}$$

Graphically, this can be observed in figure 1.

Asymptotic risk R_{1-NN}

The asymptotic risk for 1- nearest neighbor is

$$\lim_{n\to\infty} \mathbb{E}[R(g_n) = 2\mathbb{E}[\eta(x)(1-\eta(x))] = 2\int_1^2 \frac{x}{(1+x)^2}\frac{1+x}{4}\,\mathrm{d}x = \frac{1}{2}\int_1^2 \frac{x}{1+x}\,\mathrm{d}x = \frac{1}{2}(1 + ln(\frac{2}{3}))$$

Where g_n is the classifier depending on the data. Note that for $x \leq 1$ and $2 < x \leq 3$ the limit is 0.

Exercise 6:

We define the loss function $l : l(p(X, X'), Y, Y')$ which takes one if $\{p\,(X, X') = 1 \cap Y < Y'\} \cup \{p\,(X, X') = 0 \cap Y' < Y\}$ and 0 otherwise.

The optimal risk is

$$R^*(l) = \mathbb{E}\big[l(p(X, X'), Y, Y')\big] = \mathbf{P}(p(X, X') = 1, Y < Y') + \mathbf{P}(p(X, X') = 0, Y > Y') \tag{2}$$

Applying the law of total expectations to (2),

$$\begin{aligned} &= \mathbb{E}\big[\mathbf{P}(p(X, X') = 1, Y < Y', |X, X') + \mathbf{P}(p(X, X') = 0, Y > Y'|X, X')\big] \\ &= \mathbb{E}\big[\mathbb{1}_{p(X,X')=1}\mathbf{P}(Y < Y', |X, X') + (1 - \mathbb{1}_{p(X,X')=1})\mathbf{P}(Y > Y'|X, X')\big] \end{aligned} \tag{3}$$

Let's now analyze $\mathbf{P}(Y < Y', |X, X')$ and $\mathbf{P}(Y > Y'|X, X')$. These two probabilities are the probability of two independent events happening at the same time. Therefore,

$$\begin{aligned} \mathbf{P}(Y < Y', |X, X') &= \mathbf{P}(Y = 0|X)\mathbf{P}(Y' = 1|X') = (1 - \eta(x))\overset{\text{eta(x')}}{\eta'(x)} \\ \mathbf{P}(Y > Y', |X, X') &= \mathbf{P}(Y = 1|X)\mathbf{P}(Y' = 0|X') = \eta(x)(1 - \overset{\text{eta(x')}}{\eta'(x)}) \end{aligned}$$

i.e. $\mathbf{P}(Y < Y', |X, X')$,$\mathbf{P}(Y > Y'|X, X') \sim Bernouilli(\eta(x))$. Plugging in these results back to (3), we get

$$\mathbb{E}\big[\mathbb{1}_{p(X,X')=1}(1 - \eta(x))\overset{\text{eta(x')}}{\eta'(x)} + (1 - \mathbb{1}_{p(X,X')=1})\eta(x)(1 - \overset{\text{eta(x')}}{\eta'(x)})\big] \tag{4}$$

Note that the expression is the sum of two positive terms with functions bounded by $[0, 1]$. Hence, the only way to minimize it is that one of the two is 0. To achieve it, we define $p^*(X, X')$ as follows:

$$p^*(X, X') = \begin{cases} 0 & Y > Y' \ \Rightarrow \ \eta(x) > \eta'(x) \\ 1 & Y < Y' \ \Rightarrow \ \eta(x) < \eta'(x) \end{cases}$$

Finally, from (4), as only one of the two terms can be non zero, we obtain:

$$R^*(l) = \mathbb{E}\big[min\{(1 - \eta(x))\overset{\text{eta(x')}}{\eta'(x)}, (1 - \overset{\text{eta(x')}}{\eta'(x)})\eta(x)\}\big] \cancel{= \mathbb{E}\big[(1 - \eta(x))\overset{\text{eta(x')}}{\eta'(x)}\big]}$$

~~As $\eta(x)$ and $\eta'(x)$~~ are two independent observations of the same distribution, we conclude:

$$\cancel{R^*(l) = \mathbb{E}\big[(1 - \eta(x))\eta'(x)\big] = \mathbb{E}\big[\eta'(x) - \eta'(x)\eta(x)\big] = \mathbb{E}\eta(x) - \mathbb{E}^2\eta(x)}$$

Exercise 7

Compute R^*, R^{1-NN}, and R^{3-NN}

Bayes risk We define the loss function $l : l(g(x), Y) = \mathbb{1}_{g(x) \neq Y}$ where $g(x) \in \{0, 1\}$.Given l, one can show $R^* = \mathbb{E}\big[min\big(\eta(x), 1 - \eta(x)\big)\big]$ where $\mathbf{P}(Y = 1|X = x) = x^{(1)}$.

$$g^* = \begin{cases} 1 & \eta(x) > 1/2 \\ 0 & otherwise \end{cases}$$

$$R^* = \mathbb{E}\big[min\big(\eta(x), 1 - \eta(x)\big)\big] = \int_0^{1/2} \frac{x^{(1)}}{2}\, dx = \frac{1}{4}$$

Asymptotic risk 1-NN One can show that the asymptotic risk for 1-nearest-neigbour classifier is $R^{1-NN} = \lim_{n\to\infty} \mathbb{E}[R(g_n) = 2\mathbb{E}[\eta(x)(1 - \eta(x))]$.

$$R^{3-NN} = \lim_{n\to\infty} \mathbb{E}[R(g_n)] = \mathbb{E}[\eta(x)(1 - \eta(x))] = 2\int_0^1 x^{(1)}(1 - x^{(1)})\, dx = \frac{1}{3}$$

Asymptotic risk 3-NN One can show that the asymptotic risk for 3-nearest-neigbour classifier is $R^{3-NN} = \lim_{n\to\infty} \mathbb{E}[R(g_n)] = \mathbb{E}[\eta(x)(1 - \eta(x))] + 4\mathbb{E}[\eta^2(x)(1 - \eta(x))^2]$.

$$\begin{aligned} R^{3-NN} &= \lim_{n\to\infty} \mathbb{E}[R(g_n)] = \mathbb{E}[\eta(x)(1 - \eta(x))] + 4\mathbb{E}[\eta^2(x)(1 - \eta(x))^2] \\ &= \int_0^1 x^{(1)}(1 - x^{(1)})\, dx + 4\int_0^1 (x^{(1)})^2(1 - x^{(1)})^2\, dx = \frac{3}{10} \end{aligned}$$

We showed that non of the three values depend on d.

Simulations

In figure 2 we can see the evolution of the empirical risk of our K-NN classifiers as a function of dimension d compared to Bayes classifier.

Remarks

- As d increases, our empirical risk worsens in all cases. Adding more dimensions to our vectors add noise to our classifiying problem. Our classifiers compute its nearest neighbours taking into account all the dimensions instead of only the first component, which is the critical one in our classification problem.
- Note that, as expected, for n large enough and $d = 1$, risk tends to the asymptotic risk computed in the previous sections.
- As computed in the previous section, bayes risk is invariant to d. Note also that the variance also decreases when n is higher.
- The more neighbours we consider for classification, the more we reduce the risk.

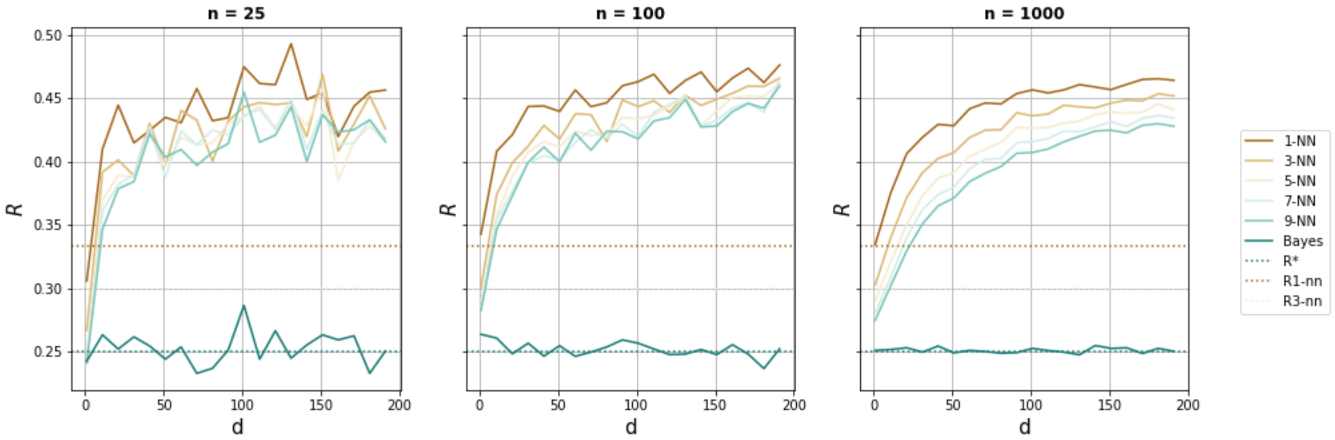


Figure 2: Bayes and K-NN classifiers risk

Error probability of empirical risk minimization rule

In this section we estimate the probability of choosing the wrong classifier using empirical risk minimization as a function of the number of test observations m. The main target is to define a threshold m' over which the probability of chosing the right classifier is optimal.

In theory, one can show that we we can bound the absolute difference between the empirical risk on a testing set $R'_m(g_n^{(i)})$ and the true risk $R(g_n^{(i)})$ for a classifier class $i \in C$ as follows:

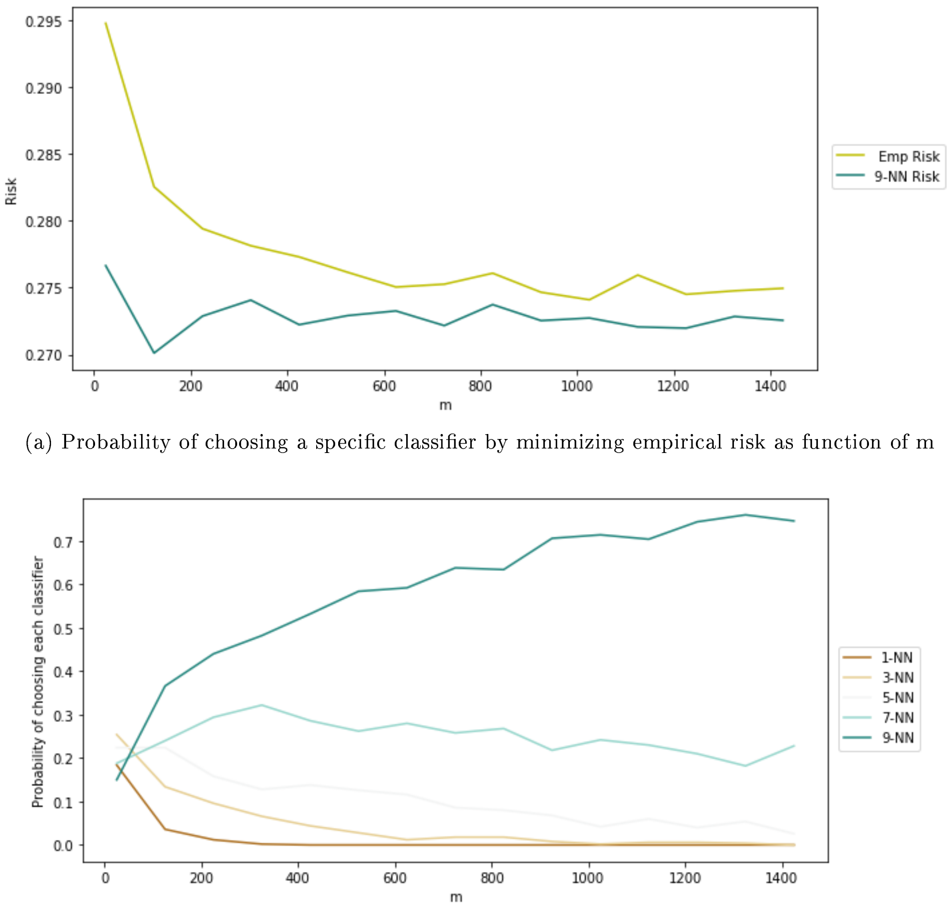
$$\underset{i=1-NN,\dots,9-NN}{max} \left| R'_m\left(g_n^{(i)}\right) - R\left(p_n^{(i)}\right)\right| \leqslant \sqrt{\frac{\log\frac{2N}{\delta}}{2m}} \qquad w.p. 1-\delta$$

where m is the number of data points in the test set D_m and N is the number of classifiers in class C.

In practice, in order to simulate what has been stated above, we are going to reproduce several times the process we would do if we had only one data set. For simplicity, the following process has been done only for 2 data.

1. Train g_n^{k-NN} classifiers for $k = \{1, 3, 5, 7, 9\}$ with a training data setD_n.
2. Choose the classifier g_j^{k-NN} from $\{g_n^{k-NN}\}$ that minimizes the empirical risk $R_m(g_j^{k-NN})$ in a test data set D_m.
3. Repeat the previous two steps several times for m.

The results can be observed in figure 3. From figure 2, we know that the classifier that minimizes the risk for different values of n and d is the 9-NN classifier. The computations show that probability of choosing the best classifier (9-NN) increase with m. For an error of 0.05, m should be around 1000 if we consider a probability of 95% as we can see in figure 3a.



(a) Probability of choosing a specific classifier by minimizing empirical risk as function of m

(b) Evolution of best performer classifier compared to the risk obtained selecting the classifier with lowest empirical risc

Figure 3