Machine Learning	Prof. Dr. Bernt Schiele
	Dr. Paul Swoboda & Dr. Gerard Pons-Moll
Exercise Sheet 2 - 29.10.2018	due: 5.11.2018, 14:15

Exercise 4 - Bayes Error

We have a binary classification problem, $\mathcal{Y} = \{-1, 1\}$, with the following distribution on $\mathcal{X} = [0, 1]$,

$$\mathrm{P}(Y = 1|X = x) = \begin{cases} 0.1, & \text{if } 0 \leq x \leq \frac{1}{8}, \\ 0.9, & \text{if } \frac{1}{8} \leq x \leq \frac{7}{8}, \\ 0.1, & \text{if } \frac{7}{8} \leq x \leq 1, \end{cases}$$

a. **(1 Points)** What is the Bayes optimal error of this problem?

b. **(1 Points)** Determine the parameter(s) (w^*, b^*) and the error of the classifier(s) $f_{(w^*,b^*)}$,

$$f_{(w,b)} = \mathrm{sign}(wx + b), \qquad w, b \in \mathbb{R}.$$

with the smallest error.

Hint:

- You don't need to give a derivation in b)- just write down the optimal parameters and the corresponding error. If there is more than one optimal set of parameters, then provide all possible optimal parameters.

Solution:

a. The Bayes error is,

$$\begin{aligned} R^* &= \int_{\mathcal{X}} \min\left\{\mathrm{P}(Y = 1|X = x), \mathrm{P}(Y = -1|X = x)\right\} p(x)dx \\ &= 0.1\frac{1}{8} + 0.1\frac{6}{8} + 0.1\frac{1}{8} = 0.1 \end{aligned}$$

b. The best linear classifiers $f_{(w^*,b^*)}$ are,

1. $w^* > 0$ and $b^* = -\frac{1}{8}w^*$, then $w^*x + b^* = w^*(x - \frac{1}{8})$,
2. $w^* < 0$ and $b^* = -\frac{7}{8}w^*$, then $w^*x + b^* = w^*(x - \frac{7}{8})$,

Note, that for this we have to assume that $p(X = x) = 1$ for $x \in [0, 1]$.

In both cases one negative part is wrongly classified. The error of both optimal linear classifiers is $\frac{1}{8}\frac{9}{10} + \frac{6}{8}\frac{1}{10} + \frac{1}{8}\frac{1}{10} = \frac{2}{10}$.

Exercise 5 - Loss functions and Bayes optimal functions

a. **(3 Points)** Let $\mathcal{Y} = \{-1, 1\}$ (binary classification). Show that the Bayes optimal function, $f^*(x) = \arg\min_{c \in \mathbb{R}} \mathbb{E}[L(Y, c)|X = x]$, for the least squares loss, $L(y, f(x)) = (y - f(x))^2$, is $f^*(x) = \mathbb{E}[Y|X = x]$ and deduce that the least squares loss is classification calibrated.

b. **(2 Points)** Let $\mathcal{Y} = \mathbb{R}_+ = \{x \in \mathbb{R} | x \geq 0\}$ (regression with output on the positive part of $\mathbb{R}$) and suppose that $\mathbb{E}[Y|X = x] > 0$. Show that the Bayes optimal function, $f^*(x) = \arg\min_{c \in \mathbb{R}} \mathbb{E}[L(Y, c)|X = x]$ for the loss function $L(y, f(x)) = \log(f(x)) + \frac{y}{f(x)}$, is given by $f^*(x) = \mathbb{E}[Y|X = x]$. Discuss the properties of this loss function compared to least square loss. For what kind of noise model do you think is this loss function useful (note that the target space is the set of non-negative reals) ?

Solution:

a. We have

$$\mathbb{E}[L(Y, c)|X = x] = (1 - c)^2 \mathrm{P}(Y = 1|X = x) + (-1 - c)^2 \mathrm{P}(Y = -1|X = x)$$

This function is continuously differentiable in c. Taking the derivative with respect to c yields

$$\frac{\partial}{\partial c} \mathbb{E}[L(Y, c)|X = x] = -2(1 - c)\mathrm{P}(Y = 1|X = x) + 2(1 + c)\mathrm{P}(Y = -1|X = x)$$

We compute the unique critical point as $c^* = \mathrm{P}(Y = 1|X = x) - \mathrm{P}(Y = -1|X = x) = \mathbb{E}[Y|X = x]$. As there are no other critical points and $\lim_{c \to \infty} \mathbb{E}[L(Y, c)|X = x] = \lim_{c \to -\infty} \mathbb{E}[L(Y, c)|X = x] = \infty$, this has to be the global minimum. Alternatively, we could have argued that the function is convex and thus every critical point is a global minimum. Obviously, $f^*(x) = c^* = \mathbb{E}[Y|X = x]$ is classification calibrated as $\mathrm{sign}(f^*(x))$ is equal to the Bayes classifier.

b. We have

$$\begin{aligned}\mathbb{E}[L(Y, c)|X = x] &= \int_0^\infty \Big[\log(c) + \frac{y}{c}\Big] p(y|x)\, dy = \log(c) \int_0^\infty p(y|x)\, dy + \frac{1}{c} \int_0^\infty y\, p(y|x)\, dy \\ &= \log(c) + \frac{\mathbb{E}[Y|X = x]}{c}\end{aligned}$$

The function is continuously differentiable on $(0, \infty)$. We have

$$\frac{\partial}{\partial c} \mathbb{E}[L(Y, c)|X = x] = \frac{1}{c} - \frac{\mathbb{E}[Y|X = x]}{c^2}$$

By assumption $\mathbb{E}[Y|X = x] > 0$. Under this condition the unique minimizer is given by $c^* = \mathbb{E}[Y|X = x]$ as $\lim_{c \to 0_+} \mathbb{E}[L(Y, c)|X = x] = \lim_{c \to \infty} \mathbb{E}[L(Y, c)|X = x] = \infty$. Note, that $\mathbb{E}[L(Y, c)|X = x]$ is not a convex function (as the second derivative becomes negative when $c > 2\mathbb{E}[Y|X = x]$), thus we cannot use the convexity argument in order to argue about the global minimizer.

Interestingly, the Bayes optimal function is the same as for the squared loss, even though the loss function are quite different. This kind of loss makes sense if one has multiplicative noise, that is $Y = (1 + \alpha)\, f(X)$, where $\alpha \in [-1, \infty]$ is the multiplicative noise level as it can happen e.g. in image processing (speckle noise). Note that the noise level is proportional to the actual function value and thus a penalization in the loss of **relative deviations** to the prediction via the ratio $\frac{y}{f(x)}$ makes sense as we will just penalize the noise term but not the prediction $f(x)$ (note that this is not completely true as we have also the term $\log(f(x))$ but this has only minor influence). This is quite different to least squares loss where the underlying noise model is additive, $Y = f(X) + \varepsilon$, and ε is Gaussian distributed and thus we penalize **absolute deviations** from the prediction $f(x)$, $L(y, f(x)) = (y - f(x))^2$ and in the case of additive noise, we penalize the noise, $L(y, f(x)) = \varepsilon^2$, but not the prediction $f(x)$.

Exercise 6 - Maximum Likelihood and Maximum A Posteriori Estimation

We have as likelihood function,

$$p(x \,|\, \theta) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{(x-\theta)^2}{2}},$$

which is a Gaussian density with mean θ and unit variance.

a. **(3 Points)** Derive the maximum-likelihood estimate of θ given n independent samples $x_1, \ldots, x_n$.

b. **(3 Points)** We have now an additional prior for the value of θ given as

$$p(\theta) = \begin{cases} \gamma\, e^{-\gamma\theta}, & \theta \geq 0, \\ 0, & \theta < 0. \end{cases},$$

for some fixed value of $\gamma > 0$. Derive the MAP estimator of θ.

c. **(1 Point)** Is the prior $p(\theta)$ reasonable ? Suppose that the data is sampled from a Gaussian distribution with unit variance. Does the MAP estimator converge to the true mean parameter of the Gaussian as $n \to \infty$?

Solution:

a. Using independence of the sample, we have

$$p(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{n} p(x_i|\theta), \quad \Longrightarrow \quad \log(p(x_1, \ldots, x_n|\theta)) = \sum_{i=1}^{n} \log(p(x_i|\theta)).$$

Thus in order to get the ML estimator, we have to maximize

$$\sum_{i=1}^{n} \log p(x_i \,|\, \theta) = \sum_{i=1}^{n} \Big(-(x_i - \theta)^2 \Big) - \frac{1}{2}\log(2\pi).$$

respectively minimize

$$\sum_{i=1}^{n} (x_i - \theta)^2 + \frac{1}{2}\log(2\pi).$$

The objective is convex in θ (second derivative is strictly positive), thus it suffices to solve for the critical point in order to get the global minimum. We have

$$\frac{d}{d\theta}\Big(-\log p(x \,|\, \theta) \Big) = -2 \sum_{i=1}^{n} (x_i - \theta).$$

This leads to the equation for the maximum likelihood estimator θ^*,

$$n\theta^* - \sum_{i=1}^{n} x_i = 0.$$

This yields finally the ML estimator,

$$\theta^* = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

b. The MAP estimator θ^* can be found by maximizing,

$$p(\theta \,|\, x_1, \ldots, x_n) = \frac{p(x_1, \ldots, x_n|\theta)p(\theta)}{p(x_1, \ldots, x_n)}.$$

As the denominator does not depend on θ, we can eliminate it and and only maximize the enumerator resp. the logarithm of it,

$$\sum_{i=1}^{n} \log p(x_i \,|\, \theta) + \log(p(\theta)).$$

Turning this into a minimization problem, we have to minimize (up to constants which we omit)

$$\min_{\theta \geq 0} \sum_{i=1}^{n} (x_i - \theta)^2 + 2\gamma\theta,$$

where θ is restricted to lie in $[0, \infty)$ as otherwise the prior is infinity. We first ignore the constraint and solve the unconstrained problem noting that

$$\frac{d}{d\theta} - \log p(\theta) = 2\gamma.$$

Using the fact that also $-\log p(\theta)$ is a convex function of θ, we obtain the unconstrained solution θ^* by solving

$$-2\sum_{i=1}^{n} (x_i - \theta) + 2\gamma = 0.$$

which yields

$$\theta^* = \frac{1}{n}\sum_{i=1}^{n} x_i - \frac{\gamma}{n}.$$

This is the solution if the unconstrained solution is feasible, that is $\sum_{i=1}^{n} x_i \geq \gamma$.

Otherwise, $\sum_{i=1}^{n} x_i < \gamma$. The derivative of the objective is

$$-2\sum_{i=1}^{n} (x_i - \theta) + 2\gamma = -2\sum_{i=1}^{n} x_i + 2n\theta + 2\gamma > -2\gamma + 2n\theta + 2\gamma = 2n\theta \geq 0,$$

for $\theta \in [0, \infty)$. Thus the solution is $\theta^* = 0$ as the objective is strictly increasing on $[0, \infty)$. Thus we have in total the solution

$$\theta^* = \begin{cases} \frac{1}{n}\sum_{i=1}^{n} x_i - \frac{\gamma}{n} & \sum_{i=1}^{n} x_i \geq \gamma, \\ 0 & \sum_{i=1}^{n} x_i < \gamma \end{cases}.$$

c. The prior is not very reasonable asymptotically, as it prevents estimates which are non-positive. Thus if the true mean is negative, the MAP estimator will be asymptotically biased (that is even as the number of samples $n \to \infty$ the estimator does not converge to the true mean of the Gaussian).