# Lab 15

Machine Learning 2021-2022 - UMONS
Souhaib Ben Taieb

## 1

You observe a dataset $\mathscr{D} = \{(\boldsymbol{x}_i, y_i\}_{i=1}^n$ where $\boldsymbol{x}_i \in \mathbb{R}^d$ and $y \in \mathbb{R}$. Assume that your model for the data is

$$y_i \sim \text{Laplace}(\boldsymbol{x}_i^\intercal \boldsymbol{\beta}, 1),$$

where $\boldsymbol{\beta} \in \mathbb{R}^d$ are the parameters of your model, and Laplace$(\mu, b)$ is the Laplace distribution with mean $\mu$ and scale $b$. Its probability density function is given by

$$f(y; \mu, b) = \frac{1}{2b} \exp\left(\frac{-|y-\mu|}{b}\right).$$

Write down the formula for the (conditional) log-likelihood as a function of the observed data and the (unknown) parameters $\boldsymbol{\beta}$. Explain your derivations.

$L(\boldsymbol{\beta}) = f(y_1, y_2, ..., y_n | \boldsymbol{x}_1, ..., \boldsymbol{x}_2; \boldsymbol{\beta})$

**Solution :**

By definition, the (conditional) likelihood of the observed data under the model is defined as:

$$\begin{aligned} L(\boldsymbol{\beta}) &= f(y_1, y_2, ..., y_n | \boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n; \boldsymbol{\beta}) \\ &= f(y_1|\boldsymbol{x}_1; \boldsymbol{\beta}) f(y_2|\boldsymbol{x}_2; \boldsymbol{\beta}) ... f(y_n|\boldsymbol{x}_n; \boldsymbol{\beta}) \qquad y_i \text{ are i.i.d. and } f(y_i|\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n; \boldsymbol{\beta}) = f(y_i|\boldsymbol{x}_i; \boldsymbol{\beta}) \\ &= \Pi_{i=1}^n f(y_i|\boldsymbol{x}_i; \boldsymbol{\beta}), \end{aligned}$$

With $f(y_i|\boldsymbol{x}_i; \boldsymbol{\beta}) = \frac{1}{2} \exp\Big(-|y_i - \boldsymbol{x}_i^\intercal \boldsymbol{\beta}|\Big)$. The (conditional) log-likelihood is obtained as:

$$\begin{aligned} \log L(\boldsymbol{\beta}) &= \log \left(\Pi_{i=1}^n f(y_i|\boldsymbol{x}_i; \boldsymbol{\beta})\right) \\ &= \sum_{i=1}^n \log f(y_i|\boldsymbol{x}_i; \boldsymbol{\beta}) \\ &= \sum_{i=1}^n \log \left(\frac{1}{2} \exp\Big(-|y_i - \boldsymbol{x}_i^\intercal \boldsymbol{\beta}|\Big)\right) \\ &= n \log \frac{1}{2} + \sum_{i=1}^n -|y_i - \boldsymbol{x}_i^\intercal \boldsymbol{\beta}| \end{aligned}$$

## 2

Consider a three-class classification problem where $X \in [0,1]$ and $Y \in \{0,1,2\}$, with the following data generating process:

$$X \sim U(0,1) \text{ and } Y|X=x \sim \begin{cases} 0, & \text{with probability } 0.2 \\ 1, & \text{with probability } 0.5x \\ 2, & \text{with probability } 0.8-0.5x \end{cases}$$

where $U(a,b)$ is a uniform random variable on the interval $[a,b]$.

(a) What is the expression of the Bayes optimal classifier ?

(b) What is the misclassification error rate of the Bayes optimal classifier ? Your answer should be a scalar.

$f_{BOC}(x) = \underset{k\in\{0,1,2\}}{\operatorname{argmax}}\, p(Y=k|X=x)$

**Solution :** (a) The Bayes optimal classifier is defined as:

$$f_{BOC}(x) = \underset{k\in\{0,1,2\}}{\operatorname{argmax}}\ p(Y=k|x),$$

where $p(Y=k|x)$ are the true conditional probabilities that generated the data, i.e.:

$$\begin{cases} p(Y=0|x) = 0.2 \\ p(Y=1|x) = 0.5x \\ p(Y=2|x) = 0.8-0.5x \end{cases}.$$

We have that:

$$\begin{aligned} p(Y=0|x) > p(Y=1|x) &\iff 0.2 > 0.5x \\ &\iff x < 0.4 \end{aligned}$$

$$\begin{aligned} p(Y=0|x) > p(Y=2|x) &\iff 0.2 > 0.8-0.5x \\ &\iff x > 1.2 \end{aligned}$$

$$\begin{aligned} p(Y=1|x) > p(Y=2|x) &\iff 0.5x > 0.8-0.5x \\ &\iff x > 0.8, \end{aligned}$$

which leads to:

$$f_{BOC}(x) = \begin{cases} 1 & \text{if } x \in\ ]0.8,1] \\ 2 & \text{if } x \in [0,0.8[ \end{cases}$$

(b)

The definition of the Bayes Error Rate is:

$$
\begin{aligned}
\text{BER} &= \mathbb{E}_x\Big[1 - \max_{k\in\{0,1,2\}} p(Y=k|x)\Big] \\
&= 1 - \int\limits_{x\in\mathscr{X}} \max_{k\in\{0,1,2\}} p(Y=k|x)f(x)dx \\
&= 1 - \int_0^1 \max_{k\in\{0,1,2\}} p(Y=k|x)dx \qquad \text{if } X\sim U[0,1], \text{ then } f(x)=1 \\
&= 1 - \int_0^{0.8}(0.8-0.5x)dx - \int_{0.8}^1 0.5xdx \\
&= 1 - [0.8x - 0.25x^2]_0^{0.8} - [0.25x^2]_{0.8}^1 \\
&= 1 - 0.64 + 0.16 - 0.25 + 0.16 \\
&= 0.43
\end{aligned}
$$

## 3

We consider Discriminant Analysis for a one-dimensional two-class classification problem. Let $X \in \mathbb{R}$ be the input variable and $Y \in N, E$, the output. We have the following:

- The prior probabilities are given by $\pi_N = P(Y = N) = \frac{\sqrt{2\pi}}{1+\sqrt{2\pi}}$ and $\pi_E = P(Y = E) = \frac{1}{1+\sqrt{2\pi}}$.
- The distribution of $X$ given $Y = N$ is Gaussian (Normal) with zero mean and variance $\sigma^2$, i.e. $X|Y = N \sim \mathcal{N}(0, \sigma^2)$.
- The distribution of $X$ given $Y = E$ is given by:

$$P(X = x|Y = E) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0. \end{cases}$$

Starting from the posterior probabilities $P(Y = N|X = x)$ and $P(Y = E|X = x)$, derive the decision boundary, i.e. an equation in $x$. Note that only the positive solutions of your equation will be relevant; ignore all $x < 0$.

If $X \sim \mathcal{N}(\mu, \sigma)$, then $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$

**Solution :**

The posterior probabilities are expressed as

$$\begin{aligned} P(Y = E|X = x) &= \frac{P(Y = E, X = x)}{f(x)} \\ &= \frac{P(X = x|Y = E)P(Y = E)}{f(x)} \\ &= \frac{\lambda e^{-\lambda x} \frac{1}{1+\sqrt{2\pi}}}{f(x)} \quad \text{Only positive values of } x \text{ are considered.} \end{aligned}$$

and

$$\begin{aligned} P(Y = N|X = x) &= \frac{P(Y = N, X = x)}{f(x)} \\ &= \frac{P(X = x|Y = N)P(Y = N)}{f(x)} \\ &= \frac{\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}} \frac{\sqrt{2\pi}}{1+\sqrt{2\pi}}}{f(x)} \end{aligned}$$

The decision boundary of Discriminant Analysis verifies:

$$\begin{aligned} & P(Y = N|X = x) = P(Y = E|X = x) \\ \iff & \frac{\lambda e^{-\lambda x} \frac{1}{1+\sqrt{2\pi}}}{f(x)} = \frac{\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}} \frac{\sqrt{2\pi}}{1+\sqrt{2\pi}}}{f(x)} \\ \iff & \lambda e^{-\lambda x} \frac{1}{1+\sqrt{2\pi}} = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}} \frac{\sqrt{2\pi}}{1+\sqrt{2\pi}} \\ \iff & \lambda e^{-\lambda x} = \frac{1}{\sigma} e^{-\frac{x}{2\sigma^2}} \end{aligned}$$

Taking the log of the right-hand and left-hand sides of the equation leads to:

$$\log \lambda + \log \sigma - \lambda x + \frac{x^2}{2\sigma^2} = 0$$

## 4

Below is a Principal Component Analysis (PCA) of a dataset *after centering and scaling* each column:

| | PC1 | PC2 | PC3 | PC4 |
|---|---|---|---|---|
| X1 | -0.5628749 | 0.2324633 | -0.5286078 | 0.5913599 |
| X2 | -0.4214823 | -0.6750169 | 0.5126131 | 0.3223859 |
| X3 | 0.5730054 | 0.2201464 | 0.3024299 | 0.7292026 |
| X4 | -0.4209386 | 0.6647170 | 0.6052584 | -0.1209310 |

| | PC1 | PC2 | PC3 | PC4 |
|---|---|---|---|---|
| Standard deviation | 1.6165 | 0.9985 | 0.50804 | 0.36313 |
| Cumulative Proportion | 0.6533 | ? | ? | ? |

(a) What is the total variance ?

(b) What proportion of the total variance does the second principal component explain ?

(c) Complete the missing values for the cumulative proportions of total variance explained.

(d) How many principal component directions would we need to explain at least 95% of the variance ?

(e) Let $\phi_1 \in \mathbb{R}^4$ and $\phi_2 \in \mathbb{R}^4$ be the first two loading vectors. What is the value of $\phi_1^\intercal \phi_2$ ? Briefly explain your answer.

**Solution :**

(a) The total variance is the sum of the variance of the individual variables, which is equal to the sum of the variance of each individual principal components.

$$\begin{aligned}
\text{TV} &= \sum_{i=1}^{k} \text{Var}(PC_k) \\
&= (1.6165)^2 + (0.9985)^2 + (0.50804)^2 + (0.36313)^2 \\
&= 4
\end{aligned}$$

(b)

$\text{PVE}_2 = \frac{\text{Var}(PC2)}{\text{TV}} = \frac{(0.9985)^2}{4} = 0.25$

(c)

$\text{PVE}_3 = \frac{\text{Var}(PC3)}{\text{TV}} = \frac{(0.50804)^2}{4} = 0.064$

| | PC1 | PC2 | PC3 | PC4 |
|---|---|---|---|---|
| Standard deviation | 1.6165 | 0.9985 | 0.50804 | 0.36313 |
| Cumulative Proportion | 0.6533 | 0.9033 | 0.9673 | 1 |

(d)

As the cumulative proportion of variance explained by the 3 first principal components amounts to 96.73%, we would only need the three first components.

(e)

The loading vectors $\phi_k \in \mathbb{R}^4$ form an orthonormal basis in $\mathbb{R}^4$. Consequently, $\phi_1^\intercal \phi_2 = 0$ as those vectors are orthogonal to one another. Furthermore, we have $\phi_1^\intercal \phi_1 = \phi_2^\intercal \phi_2 = 1$ as the loading vectors are of norm 1.