Classification

Machine Learning 2022-2023 - UMONS
Souhaib Ben Taieb

1 Exercise 1

Suppose we collect data for a group of students in a statistics class with variables:

- X_1 = hours studied.
- X_2 = undergrad GPA.
- Y = receive an A.

We fit a logistic regression and produce estimated coefficients:

- $\hat{\beta}_0 = -6$
- $\hat{\beta}_1 = 0.05$
- $\hat{\beta}_2 = 1$

a) How would the model write and how do you interpret its coefficients ?
Solution:

$$p(y|x;\boldsymbol{\beta}) = \begin{cases} \frac{e^{\hat{\beta}_0+\hat{\beta}_1 x_1+\hat{\beta}_2 x_2}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 x_1+\hat{\beta}_2 x_2}} & \text{if } y=1 \\ 1-\frac{e^{\hat{\beta}_0+\hat{\beta}_1 x_1+\hat{\beta}_2 x_2}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 x_1+\hat{\beta}_2 x_2}} & \text{if } y=0 \end{cases}$$
$$= \begin{cases} \frac{e^{-6+0.05x_1+x_2}}{1+e^{-6+0.05x_1+x_2}} & \text{if } y=1 \\ 1-\frac{e^{-6+0.05x_1+x_2}}{1+e^{-6+0.05x_1+x_2}} & \text{if } y=0 \end{cases}$$

The log-odds are linear in the input x for the logistic regression model:

$$\log\left(\frac{p(y=1|x;\boldsymbol{\beta})}{1-p(y=1|x;\boldsymbol{\beta}}\right) = \hat{\beta}_0+\hat{\beta}_1 x_1+\hat{\beta}_2 x_2$$
$$= -6+0.05x_1+x_2.$$

When everything else is held constant, a unit increase in hours studied (X_1) increases the log-odds of a student getting an A by 0.05. In terms of odds, by noting $\text{odds}_1 = e^{\hat{\beta}_0+\hat{\beta}_1 x_1+\hat{\beta}_2 x_2}$ and $\text{odds}_2 = e^{\hat{\beta}_0+\hat{\beta}_1(x_1+1)+\hat{\beta}_2 x_2}$, i.e. when x_1 has been increased by one unit, we find:

$$\begin{aligned} \text{odds}_2 &= e^{\hat{\beta}_0+\hat{\beta}_1(x_1+1)+\hat{\beta}_2 x_2} \\ &= e^{\hat{\beta}_0+\hat{\beta}_1 x_1+\hat{\beta}_2 x_2} e^{\hat{\beta}_1} \\ &= \text{odds}_1 e^{\hat{\beta}_1}, \end{aligned}$$

which yields:

$$\frac{\text{odds}_2-\text{odds}_1}{\text{odds}_1} = e^{\hat{\beta}_1}-1 = 0.051.$$

Per extra hours studied, a student increase his odds of getting an A by $e^{0.05} = 1.051$, i.e. his odds of getting an A are about $100\times(e^{\hat{\beta}_1}-1) = 5\%$ higher per extra hour studied. Similarly, when everything else is held constant, a unit increase in GPA icreases the log-odds of a student getting a A by 1, which is equivalent to increase his odds of getting an A by $e^1 = 2.718$, i.e. his odds of getting an A are about 172% higher per extra GPA score.

b) Estimate the probability that a student who studies for 40h and has an undergrad GPA of 3.5 obtains an A in the class.

Solution :

Using the definition of a logistic regression model, and from the coefficients' estimates, we obtain :

$$\begin{aligned}
p(y=1|x;\boldsymbol{\beta}) &= \frac{e^{\hat{\beta}_0+\hat{\beta}_1 x_1+\hat{\beta}_2 x_2}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 x_1+\hat{\beta}_2 x_2}} \\
&= \frac{e^{-6+0.05\times 40+1\times 3.5}}{1+e^{-6+0.05\times 40+1\times 3.5}} \\
&= \frac{e^{-0.5}}{1+e^{-0.5}} \\
&\simeq 0.378
\end{aligned}$$

b) How many hours would the above student need to study to have a 50% chance of getting an A in the class ?

Solution :

$$\begin{aligned}
p(x) &= \frac{e^{-6+0.05\times x_1+1\times 3.5}}{1+e^{-6+0.05x_1+1\times 3.5}} \\
&= \frac{e^{0.05x_1-2.5}}{1+e^{0.05x_1-2.5}} \\
&= 0.5
\end{aligned}$$

$$\begin{aligned}
&\Rightarrow e^{0.05x_1-2.5} = 0.5+0.5e^{0.05x_1-2.5} \\
&\Rightarrow e^{0.05x_1-2.5} = 1 \\
&\Rightarrow x_1 = \frac{\log(1)+2.5}{0.05} = 50
\end{aligned}$$

Exercise 2

Consider the following dataset with $n = 8$ observations, three binary input features and a binary response.

X_1	X_2	X_3	Y
1	0	1	1
1	1	1	1
0	1	1	0
1	1	0	0
1	0	1	0
0	0	0	1
0	0	0	1
0	0	1	0

Assume we are using a naive Bayes classifier to predict the value of Y from the values of the other variables.

- a) What is $P\Big(Y = 1|X_1 = 1, X_2 = 1, X_3 = 0\Big)$?

Solution :

You've seen in the lecture that in a Naïve Bayes classifier, we make the assumption that the covariance matrix is diagonal, i.e. if $p = 2$, $\Sigma = \begin{pmatrix} \sigma_{11}^2 & 0 \\ 0 & \sigma_{22}^2 \end{pmatrix}$, which implies $\sigma_{12}^2 = \sigma_{21}^2 = \mathrm{Cov}(X_1, X_2) = 0$. In fact, this property results from an even stronger assumption : the variables X_i are mutually **conditionally independent** given Y.

Under this assumption, we have $P\Big(X_1 = x_1, X_2 = x_2|Y = y\Big) = P\Big(X_1 = x_1|Y = y\Big)P\Big(X_2 = x_2|Y = y\Big)$

$$
\begin{aligned}
&P\Big(Y = 1|X_1 = 1, X_2 = 1, X_3 = 0\Big) \\
&= \frac{P\Big(X_1 = 1, X_2 = 1, X_3 = 0|Y = 1\Big)P\Big(Y = 1\Big)}{P\Big(X_1 = 1, X_2 = 1, X_3 = 0\Big)} \\
&= \frac{P\Big(X_1 = 1|Y = 1\Big)P\Big(X_2 = 1|Y = 1\Big)P\Big(X_3 = 0|Y = 1\Big)P\Big(Y = 1\Big)}{P\Big(X_1 = 1, X_2 = 1, X_3 = 0|Y = 0\Big)P\Big(Y = 0\Big) + P\Big(X_1 = 1, X_2 = 1, X_3 = 0|Y = 1\Big)P\Big(Y = 1\Big)} \\
&= \frac{P\Big(X_1 = 1|Y = 1\Big)P\Big(X_2 = 1|Y = 1\Big)P\Big(X_3 = 0|Y = 1\Big)P\Big(Y = 1\Big)}{P\Big(X_1 = 1|Y = 0\Big)P\Big(X_2 = 1|Y = 0\Big)P\Big(X_3 = 0|Y = 0\Big)P\Big(Y = 0\Big) + P\Big(X_1 = 1|Y = 1\Big)P\Big(X_2 = 1|Y + 1\Big)P\Big(X_3 = 0|Y = 1\Big)P\Big(Y = 1\Big)} \\
&= \frac{0.5 \times 0.25 \times 0.5 \times 0.5}{0.5 \times 0.5 \times 0.25 \times 0.5 + 0.5 \times 0.25 \times 0.5 \times 0.5} \\
&= 0.5
\end{aligned}
$$

- b) What is $P\Big(Y = 0|X_1 = 1, X_2 = 1\Big)$?

Solution :

$$
\begin{aligned}
&P\Big(Y = 0|X_1 = 1, X_2 = 1\Big) \\
&= \frac{P\Big(X_1 = 1|Y = 0\Big)P\Big(X_2 = 1|Y = 0\Big)P\Big(Y = 0\Big)}{P\Big(X_1 = 1|Y = 0\Big)P\Big(X_2 = 1|Y = 0\Big)P\Big(Y = 0\Big) + P\Big(X_1 = 1|Y = 1\Big)P\Big(X_2 = 1|Y = 1\Big)P\Big(Y = 1\Big)} \\
&= \frac{0.5 \times 0.5 \times 0.5}{0.5 \times 0.5 \times 0.5 + 0.5 \times 0.25 \times 0.5} \\
&= 2/3
\end{aligned}
$$

Now, suppose that we are using a joint Bayes classifier to predict the value of Y from the values of the other variables.

- c) What is $P\Big(Y=1|X_1=1,X_2=1,X_3=0\Big)$?

Solution :

In a joint Bayes classifier, we do not make the above assumption of conditional independence, meaning that $P\Big(X_1=x_1,X_2=x_2|Y=y\Big)\neq P\Big(X_1=x_1|Y=y\Big)P\Big(X_2=x_2|Y=y\Big)$.

$$\begin{aligned}
&P\Big(Y=1|X_1=1,X_2=1,X_3=0\Big)\\
&=\frac{P\Big(X_1=1,X_2=1,X_3=0|Y=1\Big)P\Big(Y=1\Big)}{P\Big(X_1=1,X_2=1,X_3=0\Big)}\\
&=\frac{0\times 0.5}{0.125}=0
\end{aligned}$$

As $P\Big(X_1=1,X_2=1,X_3=0|Y=1\Big)=0\neq\frac{1}{16}=P\Big(X_1=1|Y=1\Big)P\Big(X_2=1|Y=1\Big)P\Big(X_3=0|Y=1\Big)$, the variables X_1,X_2 and X_3 are not mutually conditionally independent given Y, which means that the assumption that we made when using Naïve Bayes is in reality not valid.

- d) What is $P\Big(Y=0|X_1=1,X_2=1\Big)$?

Solution :

$$\begin{aligned}
&P\Big(Y=0|X_1=1,X_2=1\Big)\\
&=\frac{P\Big(X_1=1,X_2=1|Y=0\Big)P\Big(Y=0\Big)}{P\Big(X_1=1,X_2=1\Big)}\\
&=\frac{0.25\times 0.5}{0.25}\\
&=0.5
\end{aligned}$$

Exercise 3

This problem relates to the QDA model, in which the observations within each class are drawn from a normal distribution with a class specific mean vector and a class specific covariance matrix. We consider the simple case where $p = 1$; i.e. there is only one feature.

Suppose that we have K classes, and that if an observation belongs to the k^{th} class, then X comes from a one-dimensional normal distribution, $X \sim \mathcal{N}(\mu_k, \sigma_k^2)$. Prove that, in that case, the Bayes' classifier is not linear. Argue that it is in fact quadratic.

Solution :

For a QDA model, we don't make the assumption of equal covariance matrices (or equal variances here as $p = 1$) across the classes. Therefore, we have that $\sigma_1^2 \neq \sigma_2^2 \neq ... \neq \sigma_K^2$, and thus :

$$f_k(x) = \frac{1}{\sigma_k\sqrt{2\pi}}\exp\Big(-\frac{1}{2\sigma_k^2}(x-\mu_k)^2\Big)$$

And therefore :

$$\begin{aligned}
p_k(x) &= \frac{\pi_k f_k(x)}{\sum_l^K \pi_l f_l(x)} \\
&= \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma_k}\exp\Big(-\frac{1}{2\sigma_k^2}(x-\mu_k)^2\Big)}{\sum_l^K \pi_l \frac{1}{\sqrt{2\pi}\sigma_l}\exp\Big(-\frac{1}{2\sigma_l^2}(x-\mu_l)^2\Big)} \\
&= \frac{\frac{\pi_k}{\sigma_k}e^{\gamma_k(x)}}{\sum_l^K \frac{\pi_l}{\sigma_l}e^{\gamma_l(x)}} \qquad \text{By posing : } \gamma_l(x) = -\frac{1}{2\sigma_l^2}(x-\mu_l)^2
\end{aligned}$$

In QDA, we want to find the value k that maximizes $p_k(x)$, i.e. we want to solve the following problem :

$$\begin{aligned}
\underset{k}{\text{argmax}}\ p_k(x) &= \underset{k}{\text{argmax}}\ \frac{\frac{\pi_k}{\sigma_k}e^{\gamma_k(x)}}{\sum_l^K \frac{\pi_l}{\sigma_l}e^{\gamma_l(x)}} \\
&= \underset{k}{\text{argmax}}\ \log\Bigg(\frac{\frac{\pi_k}{\sigma_k}e^{\gamma_k(x)}}{\sum_l^K \frac{\pi_l}{\sigma_l}e^{\gamma_l(x)}}\Bigg) \\
&= \underset{k}{\text{argmax}}\ \log\Big(\frac{\pi_k}{\sigma_k}e^{\gamma_k(x)}\Big) - \log\Big(\sum_l^K \pi_l e^{\gamma_l(x)}\Big) \\
&= \underset{k}{\text{argmax}}\ \log(\pi_k) + \gamma_k(x) - \log(\sigma_k) - \log\Big(\sum_l^K \pi_l e^{\gamma_l(x)}\Big) \\
&= \underset{k}{\text{argmax}}\ \log(\pi_k) + \gamma_k(x) - \log(\sigma_k) \qquad \text{As } \sum_l^K \pi_l e^{\gamma_l(x)} \text{ is constant } \forall k \\
&= \underset{k}{\text{argmax}}\ \log(\pi_k) + \frac{1}{2\sigma_k^2}(x-\mu_k)^2 - \log(\sigma_k) \\
&= \underset{k}{\text{argmax}}\ \log(\pi_k) + \frac{(x^2+\mu_k^2-2\mu_k x}{\sigma_k^2} - \log(\sigma_k) \\
&= \underset{k}{\text{argmax}} - \frac{1}{2\sigma_k^2}x^2 + \frac{\mu_k}{\sigma_k^2}x + (\log(\pi_k) - \log(\sigma_k) - \frac{\mu_k^2}{2\sigma_k}
\end{aligned}$$

Which is quadratic in x, hence the name *Quadratic Discriminant Analysis*.

Exercise 4

Bob is playing a bar game, for which the principle is the following: While being blindfolded, Bob has to throw a dart at random on a target that only contains number between 0 and 1. Once he has thrown, he can take off the blindfold, and look at the target value x he got. Based on this value, the bartender secretly pours a beer with probability $0.2+0.4x$, a mojito with probability $0.6-0.4x$, and a glass of wine with probability 0.2. If Bob correctly guesses the beverage that has been served, he gets it for free, otherwise he is obliged to pay for it.

1. Depending on the target value obtained, what could be the optimal prediction that Bob could make and what would be the name of such a classifier ? You will need to derive the boundary decisions of the classifier.

2. What would be the misclassification error rate of this classifier ? Your answer should be a scalar.

Solution

(1)

Given $\mathscr{Y} = \{b,m,w\}$, where b,m,w stand for "beer", "mojito", and "wine" respectively, the optimal prediction that Bob could make is by creating a Bayes optimal classifier, which is defined as:

$$f_{BOC} = \underset{y\in\mathscr{Y}}{\text{argmax}}\ P(Y=y|x)$$

Let's note $P(Y=b|x)$, $P(Y=m|x)$ and $P(Y=w|x)$ the probabilities to have a beer, a mojito, and a glass of wine respectively. The data generating process described above can be resumed as such:

$$X \sim \mathrm{U}[0,1] \qquad Y|X=x \sim \begin{cases} b, & \text{with probability } 0.2+0.4x \\ m, & \text{with probability } 0.6-0.4x \\ w, & \text{with probability } 0.2, \end{cases}$$

with $\mathrm{U}[0,1]$ being the uniform distribution defined in the interval [0,1]. We have:

$$\begin{cases} P(Y=b|x) = 0.2+0.4x \\ P(Y=m|x) = 06.-0.4x \\ P(Y=w|x) = 0.2. \end{cases} \tag{1}$$

We have that:

$$\begin{aligned} P(Y=b|x) \geq P(Y=m|x) &\iff 0.2+0.4x \geq 0.6-0.4x && (2)\\ &\iff 0.8x \geq 0.4 && (3)\\ &\iff x \geq 0.5 && (4) \end{aligned}$$

$$\begin{aligned} P(Y=b|x) \geq P(Y=w|x) &\iff 0.2+0.4x \geq 0.2 \\ &\iff 0.6x \geq 0 \\ &\iff x \geq 0 \end{aligned}$$

$$\begin{aligned} P(Y=m|x) \geq P(Y=w|x) &\iff 0.6-0.4x \geq 0.2 \\ &\iff 0.4x \leq 0.4 \\ &\iff x \leq 1, \end{aligned}$$

which leads to:

$$f_{BOC}(x) = \begin{cases} b & x \in [0.5; 1] \\ m & x \in [0; 0.5] \end{cases} \tag{5}$$

(2)

The Bayes error rate is given by:

$$\begin{aligned}
\text{BER} &= \mathbb{E}_X \left[1 - \max_{k \in \mathscr{Y}} P(Y = k|x) \right] \\
&= 1 - \int_0^1 \max_{k \in \mathscr{Y}} P(Y = k|x) f(x) dx \\
&= 1 - \int_0^1 \max_{k \in \mathscr{Y}} P(Y = k|x) dx \\
&= 1 - \int_0^{0.5} P(Y = m|x) dx - \int_{0.5}^1 P(Y = b|x) dx \\
&= 1 - \int_0^{0.5} (0.6 - 0.4x) dx - \int_{0.5}^1 (0.2 + 0.4x) dx \\
&= 1 - \left[0.6x - 0.2x^2 \right]_0^{0.5} - \left[0.2x + 0.2x^2 \right]_{0.5}^1 \\
&= 0.5
\end{aligned}$$

Exercise 5

Suppose that you are given a set of n i.i.d. observations $\mathscr{D} = \{(x_i, y_i)\}_{i=1}^n$ where y_i is a categorical variable belonging to K categories, $\mathscr{Y} = \{C_1, ..., C_K\}$. You wish to fit a multiclass logistic regression model to $\mathscr{D}$, i.e.

$$\mathbb{P}(y_i = C_k | x_i) = p_k(x_i; \boldsymbol{\beta}) = \frac{e^{\beta^{(k)} x_i}}{\sum_{l=1}^K e^{\beta^{(l)} x_i}}$$

Write the expression of the conditional log-likelihood as a function of the data and the unknown coefficients $\boldsymbol{\beta}$.

Solution:

$$\begin{aligned}
\mathscr{L}(\boldsymbol{\beta}; \mathscr{D}) &= p(y_1, ..., y_n | x_i, ..., x_n; \boldsymbol{\beta}) \\
&= \Pi_{i=1}^n p(y_i | x_i; \boldsymbol{\beta}) \quad \text{The } y_i \text{ are conditionally independent given the } x_i. \\
&= \prod_{i:y_i=C_1} p_1(x_i; \boldsymbol{\beta}) ... \prod_{i:y_i=C_K} p_K(x_i; \boldsymbol{\beta}).
\end{aligned}$$

$$\begin{aligned}
\log \mathscr{L}(\boldsymbol{\beta}; \mathscr{D}) &= \sum_{i:y_i=C_1} \log p_1(x_i; \boldsymbol{\beta}) ... \sum_{i:y_K=C_K} \log p_K(x_i; \boldsymbol{\beta}) \\
&= \sum_{i=1}^n \sum_{k=1}^K \mathbb{1}_{y_i=C_k} \log p_k(x_i; \boldsymbol{\beta}) \\
&= \sum_{i=1}^n \sum_{k=1}^K \mathbb{1}_{y_i=C_k} \log \left(\frac{e^{\beta^{(k)} x_i}}{\sum_{l=1}^K e^{\beta^{(l)} x_i}} \right)
\end{aligned}$$

where $\mathbb{1}_{y_i=C_k}$ is an indicator function that equals 1 when $y_i = C_k$, 0 otherwise. This is the definition of the categorical cross-entropy, sometimes referred to as the ”log-loss”.