Classification

Machine Learning 2021-2022 - UMONS
Souhaib Ben Taieb

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

Solution :

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

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

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

Solution :

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

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

2

Suppose that we wish to predict whether a given stock will issue a dividend this year ("Yes" or "No") based on X, last year's percent profit. We examine a large number of companies and discover that the mean value of X for companies that issued a dividend was $X = 10$, while the mean for those that didn't was $X = 0$. In addition, the variance of X for these two sets of companies was $\sigma^2 = 36$. Finally, 80% of companies issued dividends. Assuming that X follows a normal distribution, predict the probability that a company will issue a dividend this year given that its percentage profit was $X = 4$ last year.

Hint: Recall that the density function for a normal random variable is :

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{1}{2\sigma^2}(x-\mu)^2\Big)$$

You will need to use Bayes' theorem.

Solution :

Let $p_k(x)$ be the probability that a company will ($k = 1$) or will not ($k = 0$) issue a dividend this year given that its percentage profit was x last year. Using Bayes' theorem and since we assume that if X belongs to the k^{th} class, then X follows a normal distribution with density $f_k(x)$, we can write:

$$\begin{aligned} p_k(x) &= \frac{\pi_k f_k(x)}{\sum_l^K f_l(x)\pi_l} \qquad k = 1,2 \\ &= \frac{\pi_k \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\Big(-\frac{1}{2\sigma_k^2}(x-\mu_k)^2\Big)}{\sum_l^K \pi_l \frac{1}{\sqrt{2\pi\sigma_l^2}} \exp\Big(-\frac{1}{2\sigma_l^2}(x-\mu_l)^2\Big)} \\ &= \frac{\pi_k \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{1}{2\sigma^2}(x-\mu_k)^2\Big)}{\sum_l^K \pi_l \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{1}{2\sigma^2}(x-\mu_l)^2\Big)} \qquad \sigma_1 = \sigma_2 = \sigma \end{aligned}$$

We know that $\pi_1 = 0.8$, $\sigma = 6$, $\mu_1 = 10$, $\mu_2 = 0$, and thus :

$$p_1(x) = \frac{0.8 * \exp\Big(-\frac{1}{2*36}(x-10)^2\Big)}{0.8 * \exp\Big(-\frac{1}{2*36}(x-10)^2\Big) + 0.2 * \exp\Big(-\frac{1}{2*36}x^2\Big)}$$

Finally, for $X = 4$:

$$p_1(4) \simeq 0.75$$

3

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.

- 3.1) 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 $\sigma_{12}^2 = \sigma_{21}^2 = \text{Cov}(X_1, X_2) = 0$. 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 $P\Big(X_1 = x_1, X_2 = x_2|Y = y\Big) = P\Big(X_1 = x_1|Y = y\Big)P\Big(X_2 = x_2|Y = y\Big)$

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

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

Solution :

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

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

- 3.3) 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 $P\Big(X_1=x_1,X_2=x_2|Y=y\Big)\neq P\Big(X_1=x_1|Y=y\Big)P\Big(X_2=x_2|Y=y\Big)$.

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

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.

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

Solution :

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

4

Suppose that we take a data set, divide it into equally-sized training and test sets, and then try out two different classification procedures. First we use logistic regression and get an error rate of 20% on the training data and 30% on the test data. Next we use 1-nearest neighbors (i.e. K = 1) and get an average error rate (averaged over both test and training data sets) of 18%. Based on these results, which method should we prefer to use for classification of new observations? Why?

Solution :

For 1-nearest neighbors, we have $E_{train} = 0$ since in the training set the closest neighbor of each data point is itself. In other words, we have $E_{test} = 0.36$ for 1-nearest neighbors, which is higher than 0.30. Thus, we prefer logistic regression.

5

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 :

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

And therefore :

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

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

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

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