TÉCNICO LISBOA

Machine Learning Report

Homework II - Probability Distributions and Bayesian Classification

ist1114964 - Axel Carapinha
ist1106565 - Martim Gordino

October 6, 2024

Contents

1 Bayesian Classifier

1.1 Exercise a)

We will first estimate the priors.

$$p(Class = A) = \frac{4}{8} = \frac{1}{2}$$
$$p(Class = B) = \frac{4}{8} = \frac{1}{2}$$

Now, to estimate the likelihoods, we must consider each distribution as independent (Naive Bayes's assumption), so it will be:

	$p(x_1 \mid \text{Class} = A)$	$p(x_1 \mid \text{Class} = B)$
μ	1.25	1.2
σ	0.5508	0.6055

Now, for the query vector:

$$\begin{aligned} p(\text{Class} = A \mid x_1 = 1, x_2 = 2) &= \frac{p(\text{Class} = A) \cdot p(x_1 = 1, x_2 = 2 \mid \text{Class} = A)}{p(x_1 = 1, x_2 = 2)} \\ &= \frac{p(\text{Class} = A) \cdot p(x_1 = 1 \mid \text{Class} = A) \cdot p(x_2 = 2 \mid \text{Class} = A)}{p(x_1 = 1, x_2 = 2)} \\ &= \frac{\frac{1}{2} \cdot N(1 \mid \mu = 1.25, \sigma = 0.5508) \cdot N(2 \mid \mu = 1.2, \sigma = 0.6055)}{p(x_1 = 1, x_2 = 2)} \\ &= \frac{0.0899}{p(x_1 = 1, x_2 = 2)} \end{aligned}$$

$$\begin{aligned} p(\text{Class} = B \mid x_1 = 1, x_2 = 2) &= \frac{p(\text{Class} = B) \cdot p(x_1 = 1, x_2 = 2 \mid \text{Class} = B)}{p(x_1 = 1, x_2 = 2)} \\ &= \frac{p(\text{Class} = B) \cdot p(x_1 = 1 \mid \text{Class} = B) \cdot p(x_2 = 2 \mid \text{Class} = B)}{p(x_1 = 1, x_2 = 2)} \\ &= \frac{\frac{1}{2} \cdot N(1 \mid \mu = 2.7500 \sigma = 0.9574) \cdot N(2 \mid \mu = 0.5500, \sigma = 0.6403)}{p(x_1 = 1, x_2 = 2)} \\ &= \frac{0.0019}{p(x_1 = 1, x_2 = 2)} \end{aligned}$$

By comparing the numerators, we conclude that:

$$p(\text{Class} = A \mid x_1 = 1, x_2 = 2) > p(\text{Class} = B \mid x_1 = 1, x_2 = 2)$$

Hence, the most probable class for the query vector

$$x = (1, 2)^{\mathrm{T}}$$

is class A.

1.2 Exercise b)

As before, the priors have the following probabilities:

$$p(Class = A) = \frac{1}{2}$$
$$p(Class = B) = \frac{1}{2}$$

Now, to find the parameters of the two class conditional 2-d Gaussians that model the likelihoods:

	$p(x_1, x_2 \mid \text{Class} = A)$	$p(x_1, x_2 \mid \text{Class} = B)$
μ	$\begin{pmatrix} 1.25 \\ 1.20 \end{pmatrix}$	$\begin{pmatrix} 2.75 \\ 0.55 \end{pmatrix}$
$\sum$	$\begin{pmatrix} 0.3033 & 0.3267 \\ 0.3267 & 0.36667 \end{pmatrix}$	$\begin{pmatrix} 0.9166 & 0.2500 \\ 0.2500 & 0.4100 \end{pmatrix}$

Now, calculating the posteriors:

$$\begin{aligned} p(\text{Class} = A \mid x_1 = 1, x_2 = 2) &= \frac{p(\text{Class} = A) \cdot p(x_1 = 1, x_2 = 2 \mid \text{Class} = A)}{p(x_1 = 1, x_2 = 2)} \\ &= \frac{p(\text{Class} = A) \cdot p(x_1 = 1 \mid \text{Class} = A) \cdot p(x_2 = 2 \mid \text{Class} = A)}{p(x_1 = 1, x_2 = 2)} \\ &= \frac{\frac{1}{2} \cdot N(\begin{pmatrix} 1 \\ 2 \end{pmatrix} \mid \mu = \begin{pmatrix} 1.25 \\ 1.20 \end{pmatrix}, \sum = \begin{pmatrix} 0.3033 & 0.3267 \\ 0.3267 & 0.36667 \end{pmatrix})}{p(x_1 = 1, x_2 = 2)} \\ &= \frac{4.3346 \times 10^{-17}}{p(x_1 = 1, x_2 = 2)} \end{aligned}$$

$$\begin{aligned} p(\text{Class} = B \mid x_1 = 1, x_2 = 2) &= \frac{p(\text{Class} = B) \cdot p(x_1 = 1, x_2 = 2 \mid \text{Class} = B)}{p(x_1 = 1, x_2 = 2)} \\ &= \frac{p(\text{Class} = B) \cdot p(x_1 = 1 \mid \text{Class} = B) \cdot p(x_2 = 2 \mid \text{Class} = B)}{p(x_1 = 1, x_2 = 2)} \\ &= \frac{\frac{1}{2} \cdot N(\begin{pmatrix} 1 \\ 2 \end{pmatrix} \mid \mu = \begin{pmatrix} 2.75 \\ 0.55 \end{pmatrix}, \sum = \begin{pmatrix} 0.9167 & 0.2500 \\ 0.2500 & 0.4100 \end{pmatrix})}{p(x_1 = 1, x_2 = 2)} \\ &= \frac{2.3373 \times 10^{-4}}{p(x_1 = 1, x_2 = 2)} \end{aligned}$$

By comparing the numerators, we conclude that:

$$p(\text{Class} = A \mid x_1 = 1, x_2 = 2) < p(\text{Class} = B \mid x_1 = 1, x_2 = 2)$$

Hence, the most probable class for the query vector

$$x = (1, 2)^{\mathrm{T}}$$

is class B. The predicted class is not the same, which may indicate that the parameters x_1 and x_2 are not independent, making a Naive Gaussian distribution inadequate for this situation. The parameters x_1 and x_2 are likely not independent.
A more complex model that accounts for the joint distribution of the features could provide better predictive accuracy.

1.3 Exercise c)

In this exercise, we want to determine which class (A or B) appears the most times when $x_3 = 1$.

To compute that, we use the Bayes' rule of interpretation:

$$P(C \mid E) = \frac{P(E \mid C) \cdot P(C)}{P(E)}$$

Where:

- $P(C \mid E)$ is the posterior probability of the class C, given the evidence E (in this case, $x_3 = 1$).
- $P(E \mid C)$ is the likelihood of observing $x_3 = 1$ given class C.
- $P(C)$ is the prior probability of class C (the general probability of class C).
- $P(E)$ is the total probability of observing $x_3 = 1$ (the normalizing factor).

Given the following probabilities:

$$P(A) = \frac{1}{2},\ P(B) = \frac{1}{2},\ P(E \mid A) = \frac{1}{2},\ P(E \mid B) = \frac{3}{4},\ P(E) = \frac{5}{8}$$

We can now calculate both final probabilities:

$$P(A \mid E) = \frac{\frac{1}{2} \times \frac{1}{2}}{\frac{5}{8}} = \frac{2}{5}$$

$$P(B \mid E) = \frac{\frac{1}{2} \times \frac{3}{4}}{\frac{5}{8}} = \frac{3}{5}$$

So, with the calculations, we conclude that the most probable class is B.

1.4 Exercise d)

We will need to calculate two probabilities, defined as follows:

$$p(A, x_{\text{query}}) = p((1,2)|A) \cdot p(1|A) \cdot p(A)$$
$$p(B, x_{\text{query}}) = p((1,2)|B) \cdot p(1|B) \cdot p(B)$$

The higher probability between those will indicate the most probable class for the query vector $x_{\text{query}} = (1, 2)$. For simplicity, its probability will be referred as $p((1, 2))$, instead of $p(x_1 = 1, x_2 = 2)$. Noteworthy, this calculations will be of 3 main parts.

In the first part, and given the probabilities:

$$P(A) = \frac{1}{2},\ P(B) = \frac{1}{2}$$

We can calculate the probability of x_3 being 1, with the class being both A and B:

$$P(x_3 = 1 \mid A) = \frac{card(A.1)}{card(A)} = \frac{1}{2}$$

$$P(x_3 = 1 \mid B) = \frac{card(B.1)}{card(B)} = \frac{3}{4}$$

For the second part, and remembering Bayes' theorem, we can state that:

$$p(A|(1, 2)) = \frac{p((1, 2)|A) \cdot p(A)}{p((1, 2))}$$

From this, we can rearrange to find the likelihood $p((1, 2)|A)$:

$$p((1, 2)|A) = \frac{p(A|(1, 2)) \cdot p((1, 2))}{p(A)}$$

Substituting the known values into the formula:

$$p((1, 2)|A) = \frac{\frac{0.0899 \cdot p((1,2))}{p((1,2))}}{p(A)}$$

Since $p((1, 2))$ cancels out in the numerator and denominator (in 1. a) it also was considered in the denominator without problem of being zero), we simplify to:

$$p((1, 2)|A) = \frac{0.0899}{p(A)} = \frac{0.0899}{0.5} = 0.1798$$

And finally substitute to calculate $p(A, x_{\text{query}})$:

$$p(A, x_{\text{query}}) = p((1, 2)|A) \cdot p(1|A) \cdot p(A) =$$

$$= 0.1798 \cdot 0.5 \cdot 0.5 = 0.04495$$

Now, let's calculate the same for class B. Using the same steps:

From the previous calculation:

$$p(B|(1,2)) = \frac{0.0019}{p((1,2))}$$

Rearranging for $p((1,2)|B)$:

$$p((1,2)|B) = \frac{p(B|(1,2)) \cdot p((1,2))}{p(B)}$$

Substituting known values:

$$p((1,2)|B) = \frac{0.0019}{0.5} = 0.0038$$

Finally, we calculate $p(B, x_{\text{query}})$:

$$p(B, x_{\text{query}}) = p((1,2)|B) \cdot p(1|B) \cdot p(B) =$$

$$= 0.0038 \cdot 0.75 \cdot 0.5 = 0.001425$$

We now have both probabilities:

$$p(A, x_{\text{query}}) = 0.04495$$

$$p(B, x_{\text{query}}) = 0.001425$$

Since $p(A, x_{\text{query}}) > p(B, x_{\text{query}})$, the most probable class for the query vector $x_{\text{query}} = (1,2)$ is class A.

2 Software Experiment

	NIST test	**Wine test**
Train/Test Size	718 / 1079	71 / 107
kNN (k = 3)	0.98	0.69
kNN (k = 30)	0.95	0.67
Bayesian (Gauss)	0.87	0.98

Table 1: Combined results from NIST and Wine experiments.

The kNN method gives better results (greater accuracy) for the NMIST dataset, while the Gaussian Naive Baye's approach (GaussianNB) was better for the wine dataset, and this is caused by two main aspects: the Curse of Dimensionality and the type of distribution.

From one side, kNN performs surprisingly well with the NMIST (digits) dataset, and just a bit worse when the high number of neighbours (K = 30) leads to similar features (a similar input value) being interpreted as the same. Noteworthy, in this case the Curse of Dimensionality did not influence kNN because all attributes (each of the 64 matrix input' cells) are equally important (what is improved by the dimensionality reduction done with the initial input). GaussianNB, however, performs worse on the dataset, because it relates less to a Gaussian distribution.

From the other side, wine's data distribution can be well-approximated by a Gaussian distribution considering the Naive Baye's approach, leading to more accurate results. Meanwhile, kNN's lazy learning here does diminish the accuracy (specially for K = 30, for the same reason as above). Here, kNN's worse accuracy is explained by the due to the instance space being high-dimensional and also not equally relevant, consequently misleading the importance of some features (Curse of dimensionality). For instance, the color of the wine is more important than the amount of alcohol, for example.