P03 Probability Distributions and Bayesian Classification

Luis Sa-Couto[1] and Andreas Wichert[2]

INESC-ID, Instituto Superior Tecnico, Universidade de Lisboa
{luis.sa.couto,andreas.wichert}@tecnico.ulisboa.pt

1 Probability and Distributions

1) Consider the following registry where an experiment is repeated six times and four events (A, B, C and D) are detected. Considering frequentist estimates for probabilities, compute:

	D	C	B	A
1	0	0	1	1
2	0	1	1	1
3	1	0	0	0
4	1	0	0	0
5	0	0	0	0
6	0	0	0	0

- $p(A)$
- $p(A, B)$
- $p(B \mid A)$
- $p(A, B, C)$
- $p(A \mid B, C)$
- $p(A, B, C, D)$
- $p(D \mid A, B, C)$

Solution:

According to the question, what we want are frequentist estimates for probabilities. So, let us count occurrences of specific events and their combinations.

First and foremost let us define $\#\{X_1, \ldots, X_n\}$ as the number of experiment runs where events $X_1, \ldots, X_n$ appeared together. Given this definition, for example $\#\{D\} = 2$ since out of all experiment runs there are two where event D occurred. Out of the six runs, the frequencies of our measured events are:

$$\#\{D\} = 2 \; \#\{C\} = 1 \; \#\{B\} = 2 \; \#\{A\} = 2$$

So, we can get the marginal probabilities by normalizing by the total number of runs:

$$p(D) = \frac{2}{6}\ p(C) = \frac{1}{6}\ p(B) = \frac{2}{6}\ p(A) = \frac{2}{6}$$

We can repeat the same reasoning for pairs of events:

$$\#\{A,B\} = 2\ \#\{A,D\} = 0\ \#\{B,D\} = 0$$
$$\#\{A,C\} = 1\ \#\{B,C\} = 1\ \#\{C,D\} = 0$$

So, we can get the joint probabilities for the pairs by normalizing by the total number of runs:

$$p(A,B) = \frac{2}{6}\ p(A,D) = 0\ p(B,D) = 0$$
$$p(A,C) = \frac{1}{6}\ p(B,C) = \frac{1}{6}\ p(C,D) = 0$$

Yet again, we repeat the process for triples:

$$\#\{A,B,C\} = 1\ \#\{A,B,D\} = 0\ \#\{A,C,D\} = 0\ \#\{B,C,D\} = 0$$

$$p(A,B,C) = \frac{1}{6}\ p(A,B,D) = 0\ p(A,C,D) = 0\ p(B,C,D) = 0$$

Finally, for all events:

$$\#\{A,B,C,D\} = 0\ p(A,B,C,D) = 0$$

Now, we can use the rules of probability to compute everything:

- $p(A) = \frac{2}{6}$
- $p(A,B) = \frac{2}{6}$
- $p(B \mid A) = \frac{p(A,B)}{p(A)} = \frac{\frac{2}{6}}{\frac{2}{6}} = 1$
- $p(A,B,C) = \frac{1}{6}$
- $p(A \mid B,C) = \frac{p(A,B,C)}{p(B,C)} = \frac{\frac{1}{6}}{\frac{1}{6}} = 1$
- $p(A,B,C,D) = 0$
- $p(D \mid A,B,C) = \frac{p(A,B,C,D)}{p(A,B,C)} = \frac{0}{\frac{1}{6}} = 0$

---

2) Consider the following set of height measures in centimeters of a group of people:

$$X \big| 180\ 160\ 200\ 171\ 159\ 150$$

What are the maximum likelihood parameters of a gaussian distribution for this set of points? Plot it approximately.

Solution:

The maximum likelihood gaussian is defined by the sample mean and standard deviation. Let us compute them:

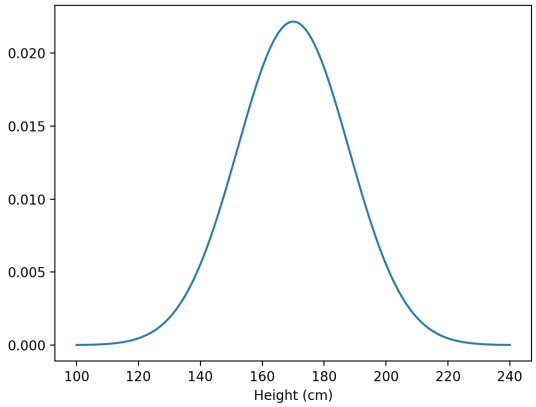
$$\mu = \frac{180 + 160 + 200 + 171 + 159 + 150}{6} = 170$$

$$\sigma = \frac{1}{6-1}((180-170)^2 + (160-170)^2 + \\ + (200-170)^2 + (171-170)^2 + (159-170)^2 + (150-170)^2)^{\frac{1}{2}} = 18.0111$$

Having the parameters, we can write the expression:

$$N(x \mid \mu, \sigma) = \frac{1}{18.0111\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-170}{18.0111}\right)^2\right)$$

This gaussian can be plotted as a function of x as follows:



---

3) Consider the following set of two dimensional measures:

X_1	−2	−1	0	−2
X_2	2	3	1	1

What are the maximum likelihood parameters of a Gaussian distribution for this set of points? What is the shape of the Gaussian? Draw it approximately using a contour map.

Solution:

The maximum likelihood gaussian is defined by the sample mean vector and the covariance matrix. Let us compute them:

$$\mu = \frac{1}{4}\left(\begin{bmatrix}-2\\2\end{bmatrix} + \begin{bmatrix}-1\\3\end{bmatrix} + \begin{bmatrix}0\\1\end{bmatrix} + \begin{bmatrix}-2\\1\end{bmatrix}\right) = \begin{bmatrix}-1.25\\1.75\end{bmatrix}$$

$$\begin{aligned}\Sigma_{00} = \frac{1}{4-1}[&(-2+1.25)(-2+1.25) + (-1+1.25)(-1+1.25) + \\ &+ (0+1.25)(0+1.25) + (-2+1.25)(-2+1.25)] \approx 0.9167\end{aligned}$$

$$\begin{aligned}\Sigma_{01} = \frac{1}{4-1}[&(-2+1.25)(2-1.75) + (-1+1.25)(3-1.75) + \\ &+ (0+1.25)(1-1.75) + (-2+1.25)(1-1.75)] \approx -0.0833\end{aligned}$$

$$\begin{aligned}\Sigma_{10} = \frac{1}{4-1}[&(2-1.75)(-2+1.25) + (3-1.75)(-1+1.25) + \\ &+ (1-1.75)(0+1.25) + (1-1.75)(-2+1.25)] \approx -0.0833\end{aligned}$$

$$\begin{aligned}\Sigma_{11} = \frac{1}{4-1}[&(2-1.75)(2-1.75) + (3-1.75)(3-1.75) + \\ &+ (1-1.75)(1-1.75) + (1-1.75)(1-1.75)] \approx 0.9167\end{aligned}$$

To compute the expression for the multivariate gaussian we need to compute the determinant of Σ and its inverse:

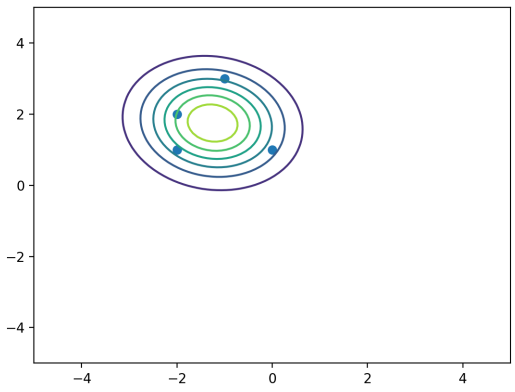
$$det(\Sigma) = det\begin{pmatrix}0.9167 & -0.0833\\-0.0833 & 0.9167\end{pmatrix} = (0.9167 \cdot 0.9167) - (-0.0833 \cdot -0.0833) = 0.8333$$

$$\Sigma^{-1} = \frac{1}{0.8333}\begin{bmatrix}0.9167 & 0.0833\\0.0833 & 0.9167\end{bmatrix} = \begin{bmatrix}1.1 & 0.1\\0.1 & 1.1\end{bmatrix}$$

So, we can write the expression for a two dimensional input $\mathbf{x} = \begin{bmatrix}x_0 & x_1\end{bmatrix}^T$ as follows.

$$N(\mathbf{x} \mid \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{2}{2}}\sqrt{0.8333}} \exp\left(-\frac{1}{2}\left(\begin{bmatrix}x_0\\x_1\end{bmatrix} - \begin{bmatrix}-1.25\\1.75\end{bmatrix}\right)^T \begin{bmatrix}1.1 & 0.1\\0.1 & 1.1\end{bmatrix}\left(\begin{bmatrix}x_0\\x_1\end{bmatrix} - \begin{bmatrix}-1.25\\1.75\end{bmatrix}\right)\right)$$

Looking at the covariance matrix, we can see that the shape must be an ellipse, which can be plotted as follows.



---

4) Consider the following set of two dimensional measures:

X_1	2	1	0	2
X_2	−2	3	−1	1

What are the maximum likelihood parameters of a Gaussian distribution for this set of points? What is the shape of the Gaussian? Draw it approximately using a contour map.

Solution:

The maximum likelihood gaussian is defined by the sample mean vector and the covariance matrix. Let us compute them:

$$\mu = \frac{1}{4}\left(\begin{bmatrix}2\\-2\end{bmatrix} + \begin{bmatrix}1\\3\end{bmatrix} + \begin{bmatrix}0\\-1\end{bmatrix} + \begin{bmatrix}2\\1\end{bmatrix}\right) = \begin{bmatrix}1.25\\0.25\end{bmatrix}$$

$$\Sigma_{00} = \frac{1}{4-1}[(2-1.25)(2-1.25) + (1-1.25)(1-1.25) + \\ + (0-1.25)(0-1.25) + (2-1.25)(2-1.25)] \approx 0.9167$$

$$\Sigma_{01} = \frac{1}{4-1}[(2-1.25)(-2-0.25) + (1-1.25)(3-0.25) + \\ + (0-1.25)(-1-0.25) + (2-1.25)(1-0.25)] \approx -0.0833$$

$$\Sigma_{10} = \frac{1}{4-1}[(-2-0.25)(2-1.25)+(3-0.25)(1-1.25)+ \\ +(-1-0.25)(0-1.25)+(1-0.25)(2-1.25)] \approx -0.0833$$

$$\Sigma_{11} = \frac{1}{4-1}[(-2-0.25)(-2-0.25)+(3-0.25)(3-0.25)+ \\ +(-1-0.25)(-1-0.25)+(1-0.25)(1-0.25)] \approx 4.9167$$

To compute the expression for the multivariate gaussian we need to compute the determinant of Σ and its inverse:

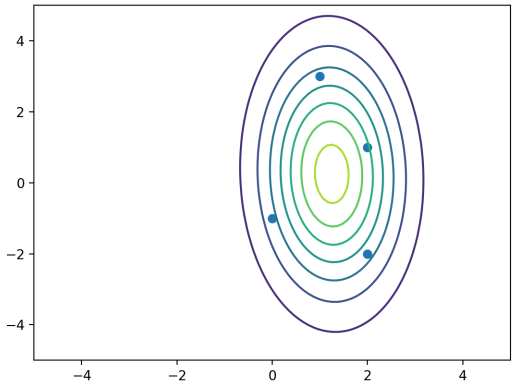
$$det(\Sigma) = det\begin{pmatrix} 0.9167 & -0.0833 \\ -0.0833 & 4.9167 \end{pmatrix} = (0.9167 \cdot 4.9167) - (-0.0833 \cdot -0.0833) = 4.5$$

$$\Sigma^{-1} = \frac{1}{4.5}\begin{bmatrix} 4.9167 & 0.0833 \\ 0.0833 & 0.9167 \end{bmatrix} = \begin{bmatrix} 1.0926 & 0.0185 \\ 0.0185 & 0.2037 \end{bmatrix}$$

So, we can write the expression for a two dimensional input $\mathbf{x} = \begin{bmatrix} x_0 & x_1 \end{bmatrix}^T$ as follows.

$$N(\mathbf{x} \mid \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{2}{2}}\sqrt{4.5}} \exp\left(-\frac{1}{2}\left(\begin{bmatrix} x_0 \\ x_1 \end{bmatrix} - \begin{bmatrix} 1.25 \\ 0.25 \end{bmatrix}\right)^T \begin{bmatrix} 1.0926 & 0.0185 \\ 0.0185 & 0.2037 \end{bmatrix} \left(\begin{bmatrix} x_0 \\ x_1 \end{bmatrix} - \begin{bmatrix} 1.25 \\ 0.25 \end{bmatrix}\right)\right)$$

Looking at the covariance matrix, we can see that the shape must be an ellipse, which can be plotted as follows.



2 Simple Bayesian Learning

Bayes rule tells us how we can compute the probability of a given hypothesis H given a set of observed data D:

$$p(H \mid D) = \frac{p(D \mid H)\, p(H)}{p(D)}$$

This formula is so widely used that all its terms have well established names:

- $p(H)$ is called the prior and denotes the a priori probability of a given hypothesis i.e. how likely a given explanation is before seeing any data;
- $p(D \mid H)$ is called the likelihood and measures how likely is the observed data if we assume that H is true;
- $p(D)$ is usually called the evidence and determines how likely the observed data is under all possible hypotheses;
- $p(H \mid D)$ is named the posterior to refer to the probability of the hypothesis after taking the observed data into account.

Classification with this rule amounts to choosing the most probable class given a specific set of features. For instance, if in our learning problem there are two features X_1 and X_2 to predict a class C, a Bayesian classifier outputs the class with the highest predictive posterior probability $p(C \mid X_1, X_2)$.

To compute posteriors using Bayes's rule, we need to compute likelihoods $p(X_1, X_2 \mid C)$. Often, these likelihoods are complex conditional joint distributions that are hard to estimate from limited data. In such cases, assumptions must be made. A well-known, rather useful one is the Naive Bayes assumption where features X_1 and X_2 are treated as conditionally independent given the class. Specifically, this allows us to factorize the conditional joint distributions into a product of simpler ones which are easier to estimate from limited data $p(X_1, X_2 \mid C) = p(X_1 \mid C)\, p(X_2 \mid C)$.

1) Assuming that 1 means $True$ and 0 means $False$, consider the following features and class:

- X_1: "Fast processing"
- X_2: "Decent Battery"
- X_3: "Good Camera"
- X_4: "Good Look and Feel"
- X_5: "Easiness of Use"
- $Class$: "iPhone"

You are given the following training set:

X_1	X_2	X_3	X_4	X_5	$Class$
1	1	0	1	0	1
1	1	1	0	0	0
0	1	1	1	0	0
0	0	0	1	1	0
1	0	1	1	1	1
0	0	1	0	0	1
0	0	0	0	1	1

And the query vector $\mathbf{x} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \end{bmatrix}^T$.

a) Using Bayes' rule, without making any assumptions, compute the class for the query vector.

Solution:

In this exercise we will see that a small training sample will not be enough to make a decision. If we apply the same methodology of estimating probabilities by counting and normalizing, we will get lots of zeros in the likelihood distributions which will lead to posteriors that cannot be computed.

Let us apply the methodology and verify just that. So, to compute the class we must choose the one that yields the maximum posterior probability. To compute the posterior, we need the likelihoods and the priors:

$$p\left(C = 0\right) = \frac{3}{7}$$

$$p\left(C = 1\right) = \frac{4}{7}$$

X_1	X_2	X_3	X_4	X_5	$p(X_1, X_2, X_3, X_4, X_5 \mid C = 0)$	$p(X_1, X_2, X_3, X_4, X_5 \mid C = 1)$
0	0	0	0	0	0	0
0	0	0	0	1	0	$\frac{1}{4}$
0	0	0	1	0	0	0
0	0	0	1	1	$\frac{1}{3}$	0
0	0	1	0	0	0	$\frac{1}{4}$
0	0	1	0	1	0	0
0	0	1	1	0	0	0
0	0	1	1	1	0	0
0	1	0	0	0	0	0
0	1	0	0	1	0	0
0	1	0	1	0	0	0
0	1	0	1	1	0	0
0	1	1	0	0	0	0
0	1	1	0	1	0	0
0	1	1	1	0	$\frac{1}{3}$	0
0	1	1	1	1	0	0
1	0	0	0	0	0	0
1	0	0	0	1	0	0
1	0	0	1	0	0	0
1	0	0	1	1	0	0
1	0	1	0	0	0	0
1	0	1	0	1	0	0
1	0	1	1	0	0	0
1	0	1	1	1	0	$\frac{1}{4}$
1	1	0	0	0	0	0
1	1	0	0	1	0	0
1	1	0	1	0	0	$\frac{1}{4}$
1	1	0	1	1	0	0
1	1	1	0	0	$\frac{1}{3}$	0
1	1	1	0	1	0	0
1	1	1	1	0	0	0
1	1	1	1	1	0	0

Now, we would use Bayes's rule to get a posterior for each class:

$$p(C = 0 \mid X_1 = 1, X_2 = 1, X_3 = 1, X_4 = 1, X_5 = 1) = \frac{p(C = 0)\, p(X_1 = 1, X_2 = 1, X_3 = 1, X_4 = 1, X_5 = 1 \mid C = 0)}{p(X_1 = 1, X_2 = 1, X_3 = 1, X_4 = 1, X_5 = 1)}$$

$$p(C = 1 \mid X_1 = 1, X_2 = 1, X_3 = 1, X_4 = 1, X_5 = 1) = \frac{p(C = 1)\, p(X_1 = 1, X_2 = 1, X_3 = 1, X_4 = 1, X_5 = 1 \mid C = 1)}{p(X_1 = 1, X_2 = 1, X_3 = 1, X_4 = 1, X_5 = 1)}$$

However, according to our estimated likelihoods, the denominators are equal to zero:

$$\begin{aligned} p(X_1 = 1, X_2 = 1, X_3 = 1, X_4 = 1, X_5 = 1) = \\ p(X_1 = 1, X_2 = 1, X_3 = 1, X_4 = 1, X_5 = 1, C = 0) + \\ + p(X_1 = 1, X_2 = 1, X_3 = 1, X_4 = 1, X_5 = 1, C = 1) = \\ p(C = 0)\, p(X_1 = 1, X_2 = 1, X_3 = 1, X_4 = 1, X_5 = 1 \mid C = 0) + \\ + p(C = 1)\, p(X_1 = 1, X_2 = 1, X_3 = 1, X_4 = 1, X_5 = 1 \mid C = 1) = \\ \frac{3}{7}0 + \frac{4}{7}0 = \\ 0 \end{aligned}$$

So, both posteriors are not defined and, thus, we cannot classify the input.

---

b) What is the problem of working with this data set without assumptions?

Solution:
Nearly all entries are zero. There is not enough data to construct a meaningful joint distribution.

---

c) Compute the class for the same query vector under the Naive Bayes assumption.

Solution:
The Naive Bayes assumption of conditional independence posits that the conditional joint distribution that constitutes the likelihood can be written as a product of conditional distributions for each feature. So, for this exercise, instead of estimating a big table for all combinations of features, we need to estimate one table per feature.

X_1	$p(X_1 \mid C = 0)$	$p(X_1 \mid C = 1)$
0	$\frac{2}{3}$	$\frac{2}{4}$
1	$\frac{1}{3}$	$\frac{2}{4}$

X_2	$p(X_2 \mid C = 0)$	$p(X_2 \mid C = 1)$
0	$\frac{1}{3}$	$\frac{3}{4}$
1	$\frac{2}{3}$	$\frac{1}{4}$

X_3	$p(X_3 \mid C = 0)$	$p(X_3 \mid C = 1)$
0	$\frac{1}{3}$	$\frac{2}{4}$
1	$\frac{2}{3}$	$\frac{2}{4}$

X_4	$p(X_4 \mid C=0)$	$p(X_4 \mid C=1)$
0	$\frac{1}{3}$	$\frac{2}{4}$
1	$\frac{2}{3}$	$\frac{2}{4}$

X_5	$p(X_5 \mid C=0)$	$p(X_5 \mid C=1)$
0	$\frac{2}{3}$	$\frac{2}{4}$
1	$\frac{1}{3}$	$\frac{2}{4}$

Just like in the previous exercise we compute the posteriors:

$$p(C=0 \mid X_1=1, X_2=1, X_3=1, X_4=1, X_5=1) = \frac{p(C=0)\,p(X_1=1, X_2=1, X_3=1, X_4=1, X_5=1 \mid C=0)}{p(X_1=1, X_2=1, X_3=1, X_4=1, X_5=1)} = \frac{p(C=0)\,p(X_1=1 \mid C=0)\,p(X_2=1 \mid C=0)\,p(X_3=1 \mid C=0)\,p(X_4=1 \mid C=0)\,p(X_5=1 \mid C=0)}{p(X_1=1, X_2=1, X_3=1, X_4=1, X_5=1)} = \frac{\frac{3}{7}\frac{1}{3}\frac{2}{3}\frac{2}{3}\frac{2}{3}\frac{1}{3}}{p(X_1=1, X_2=1, X_3=1, X_4=1, X_5=1)}$$

$$p(C=1 \mid X_1=1, X_2=1, X_3=1, X_4=1, X_5=1) = \frac{p(C=1)\,p(X_1=1, X_2=1, X_3=1, X_4=1, X_5=1 \mid C=1)}{p(X_1=1, X_2=1, X_3=1, X_4=1, X_5=1)} = \frac{p(C=1)\,p(X_1=1 \mid C=1)\,p(X_2=1 \mid C=1)\,p(X_3=1 \mid C=1)\,p(X_4=1 \mid C=1)\,p(X_5=1 \mid C=1)}{p(X_1=1, X_2=1, X_3=1, X_4=1, X_5=1)} = \frac{\frac{4}{7}\frac{2}{4}\frac{1}{4}\frac{2}{4}\frac{2}{4}\frac{1}{4}}{p(X_1=1, X_2=1, X_3=1, X_4=1, X_5=1)}$$

Since the denominator is the same for both expressions, to classify, we just need to choose the larger numerator:

$$\frac{3}{7}\frac{1}{3}\frac{2}{3}\frac{2}{3}\frac{2}{3}\frac{1}{3} = 0.0141 > 0.0090 = \frac{4}{7}\frac{2}{4}\frac{1}{4}\frac{2}{4}\frac{2}{4}\frac{2}{4}$$

So the classifier outputs label $C=0$ which means not an iPhone.

---

2) From the following training set:

X_1	X_2	X_3	X_4	X_5	C
1	1	0	1	0	a
1	0	0	1	1	a
1	0	0	1	1	a
1	1	1	0	1	b
0	0	1	1	1	b
1	0	0	0	0	c

a) Compute the class for the pattern $\mathbf{x} = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 \end{bmatrix}^T$ under the Naive Bayes assumption.

Solution:

Just like in the previous exercise, we have to estimate the priors and the likelihoods.

$$p(C = a) = \frac{3}{6}$$

$$p(C = b) = \frac{2}{6}$$

$$p(C = c) = \frac{1}{6}$$

X_1	$p(X_1 \mid C = a)$	$p(X_1 \mid C = b)$	$p(X_1 \mid C = c)$
0	$\frac{0}{3}$	$\frac{1}{2}$	$\frac{0}{1}$
1	$\frac{3}{3}$	$\frac{1}{2}$	$\frac{1}{1}$

X_2	$p(X_2 \mid C = a)$	$p(X_2 \mid C = b)$	$p(X_2 \mid C = c)$
0	$\frac{2}{3}$	$\frac{1}{2}$	$\frac{1}{1}$
1	$\frac{1}{3}$	$\frac{1}{2}$	$\frac{0}{1}$

X_3	$p(X_3 \mid C = a)$	$p(X_3 \mid C = b)$	$p(X_3 \mid C = c)$
0	$\frac{3}{3}$	$\frac{0}{2}$	$\frac{1}{1}$
1	$\frac{0}{3}$	$\frac{2}{2}$	$\frac{0}{1}$

X_4	$p(X_4 \mid C = a)$	$p(X_4 \mid C = b)$	$p(X_4 \mid C = c)$
0	$\frac{0}{3}$	$\frac{1}{2}$	$\frac{1}{1}$
1	$\frac{3}{3}$	$\frac{1}{2}$	$\frac{0}{1}$

X_5	$p(X_5 \mid C = a)$	$p(X_5 \mid C = b)$	$p(X_5 \mid C = c)$
0	$\frac{1}{3}$	$\frac{0}{2}$	$\frac{1}{1}$
1	$\frac{2}{3}$	$\frac{2}{2}$	$\frac{0}{1}$

Having estimated the required parameters, we can compute the posteriors:

$$p(C = a \mid X_1 = 1, X_2 = 0, X_3 = 1, X_4 = 0, X_5 = 1) =$$

$$\frac{p(C = a)\, p(X_1 = 1, X_2 = 0, X_3 = 1, X_4 = 0, X_5 = 1 \mid C = a)}{p(X_1 = 1, X_2 = 0, X_3 = 1, X_4 = 0, X_5 = 1)} =$$

$$\frac{p(C = a)\, p(X_1 = 1 \mid C = a)\, p(X_2 = 0 \mid C = a)\, p(X_3 = 1 \mid C = a)\, p(X_4 = 0 \mid C = a)\, p(X_5 = 1 \mid C = a)}{p(X_1 = 1, X_2 = 0, X_3 = 1, X_4 = 0, X_5 = 1)} =$$

$$\frac{\frac{3}{6}\frac{3}{3}\frac{2}{3}\frac{0}{3}\frac{0}{3}\frac{2}{3}}{p(X_1 = 1, X_2 = 0, X_3 = 1, X_4 = 0, X_5 = 1)} =$$

$$0$$

$$p\left(C=b \mid X_1=1, X_2=0, X_3=1, X_4=0, X_5=1\right) =$$
$$\frac{p\left(C=b\right) p\left(X_1=1, X_2=0, X_3=1, X_4=0, X_5=1 \mid C=b\right)}{p\left(X_1=1, X_2=0, X_3=1, X_4=0, X_5=1\right)} =$$
$$\frac{p\left(C=b\right) p\left(X_1=1 \mid C=b\right) p\left(X_2=0 \mid C=b\right) p\left(X_3=1 \mid C=b\right) p\left(X_4=0 \mid C=b\right) p\left(X_5=1 \mid C=b\right)}{p\left(X_1=1, X_2=0, X_3=1, X_4=0, X_5=1\right)} =$$
$$\frac{\frac{2}{6}\frac{1}{2}\frac{1}{2}\frac{2}{2}\frac{1}{2}\frac{2}{2}}{p\left(X_1=1, X_2=0, X_3=1, X_4=0, X_5=1\right)}$$

$$p\left(C=c \mid X_1=1, X_2=0, X_3=1, X_4=0, X_5=1\right) =$$
$$\frac{p\left(C=c\right) p\left(X_1=1, X_2=0, X_3=1, X_4=0, X_5=1 \mid C=c\right)}{p\left(X_1=1, X_2=0, X_3=1, X_4=0, X_5=1\right)} =$$
$$\frac{p\left(C=c\right) p\left(X_1=1 \mid C=c\right) p\left(X_2=0 \mid C=c\right) p\left(X_3=1 \mid C=c\right) p\left(X_4=0 \mid C=c\right) p\left(X_5=1 \mid C=c\right)}{p\left(X_1=1, X_2=0, X_3=1, X_4=0, X_5=1\right)} =$$
$$\frac{\frac{1}{6}\frac{1}{1}\frac{1}{1}\frac{0}{1}\frac{1}{1}\frac{0}{1}}{p\left(X_1=1, X_2=0, X_3=1, X_4=0, X_5=1\right)} =$$
$$0$$

The only non zero probability occurs for class $C=b$. So, that is the classifier's output.

---

b) What is the posterior probability $p\left(b \mid \mathbf{x}\right)$?

Solution:

To get the final values for all posterior elements, we need to get rid of the unknown quantity in the denominator $p\left(X_1=1, X_2=0, X_3=1, X_4=0, X_5=1\right)$:

$$\sum_{y \in \{a,b,c\}} p\left(C=y \mid X_1=1, X_2=0, X_3=1, X_4=0, X_5=1\right) = 1$$

$$\sum_{y \in \{a,b,c\}} p\left(C=y \mid X_1=1, X_2=0, X_3=1, X_4=0, X_5=1\right) = 1$$
$$\frac{\frac{2}{6}\frac{1}{2}\frac{1}{2}\frac{2}{2}\frac{1}{2}\frac{2}{2}}{p\left(X_1=1, X_2=0, X_3=1, X_4=0, X_5=1\right)} = 1$$
$$p\left(X_1=1, X_2=0, X_3=1, X_4=0, X_5=1\right) = \frac{2}{6}\frac{1}{2}\frac{1}{2}\frac{2}{2}\frac{1}{2}\frac{2}{2}$$

So, the required posterior is $p\left(C=b \mid X_1=1, X_2=0, X_3=1, X_4=0, X_5=1\right) = 1$.

c) What do you do if we have missing features? More specifically, under the Naive Bayes assumption, to what class does $\mathbf{x_{missing}} = \begin{bmatrix} 1 & ? & 1 & ? & 1 \end{bmatrix}^T$ belong to?

Solution:
We compute the posteriors for the features we have:

$$\begin{aligned} p(C = a \mid X_1 = 1, X_3 = 1, X_5 = 1) = \\ \frac{p(C = a)\, p(X_1 = 1, X_3 = 1, X_5 = 1 \mid C = a)}{p(X_1 = 1, X_3 = 1, X_5 = 1)} = \\ \frac{p(C = a)\, p(X_1 = 1 \mid C = a)\, p(X_3 = 1 \mid C = a)\, p(X_5 = 1 \mid C = a)}{p(X_1 = 1, X_3 = 1, X_5 = 1)} = \\ \frac{\frac{3}{6}\frac{3}{3}\frac{0}{3}\frac{2}{3}}{p(X_1 = 1, X_3 = 1, X_5 = 1)} = \\ 0 \end{aligned}$$

$$\begin{aligned} p(C = b \mid X_1 = 1, X_3 = 1, X_5 = 1) = \\ \frac{p(C = b)\, p(X_1 = 1, X_3 = 1, X_5 = 1 \mid C = b)}{p(X_1 = 1, X_3 = 1, X_5 = 1)} = \\ \frac{p(C = b)\, p(X_1 = 1 \mid C = b)\, p(X_3 = 1 \mid C = b)\, p(X_5 = 1 \mid C = b)}{p(X_1 = 1, X_3 = 1, X_5 = 1)} = \\ \frac{\frac{2}{6}\frac{1}{2}\frac{2}{2}\frac{2}{2}}{p(X_1 = 1, X_3 = 1, X_5 = 1)} \end{aligned}$$

$$\begin{aligned} p(C = c \mid X_1 = 1, X_3 = 1, X_5 = 1) = \\ \frac{p(C = c)\, p(X_1 = 1, X_3 = 1, X_5 = 1 \mid C = c)}{p(X_1 = 1, X_3 = 1, X_5 = 1)} = \\ \frac{p(C = c)\, p(X_1 = 1 \mid C = c)\, p(X_3 = 1 \mid C = c)\, p(X_5 = 1 \mid C = c)}{p(X_1 = 1, X_3 = 1, X_5 = 1)} = \\ \frac{\frac{1}{6}\frac{1}{1}\frac{0}{1}\frac{0}{1}}{p(X_1 = 1, X_2 = 0, X_3 = 1, X_4 = 0, X_5 = 1)} = \\ 0 \end{aligned}$$

Again, class b is the only one with non zero probability.

3) So far we have been dealing always with discrete feature domains. In this exercise, we will work with continuous values for features like Height and Weight. Assuming that 1means $True$ and 0means $False$, consider the following features and class:

- X_1: "Weight (Kg)"
- X_2: "Height (Cm)"
- $Class$: "NBA Player"

You are given the following training set:

X_1	X_2	$Class$
170	160	0
80	220	1
90	200	1
60	160	0
50	150	0
70	190	1

And the query vector $\mathbf{x} = \begin{bmatrix} 100 & 225 \end{bmatrix}^T$.

a) Compute the most probable class for the query vector assuming that the likelihoods are 2-dimensional Gaussians.

Solution:

We start by estimating the priors:

$$p\left(C=0\right) = \frac{1}{2}$$

$$p\left(C=1\right) = \frac{1}{2}$$

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

$$\begin{array}{cc} & p\left(X_1, X_2 \mid C=0\right) \quad p\left(X_1, X_2 \mid C=1\right) \\ \mu & \begin{bmatrix} 93.3333 \\ 156.6667 \end{bmatrix} \quad \begin{bmatrix} 80 \\ 203.3333 \end{bmatrix} \\ \Sigma & \begin{bmatrix} 4433.3333 & 216.6667 \\ 216.6667 & 33.3333 \end{bmatrix} \quad \begin{bmatrix} 100 & 50 \\ 50 & 233.3333 \end{bmatrix} \end{array}$$

We can thus compute the posteriors:

$$p\left(C=0 \mid X_1=100, X_2=225\right)= \frac{p\left(C=0\right)p\left(X_1=100, X_2=225 \mid C=0\right)}{p\left(X_1=100, X_2=225\right)}= \frac{\frac{1}{2}N\left(\begin{bmatrix}100\\225\end{bmatrix} \mid \mu=\begin{bmatrix}93.3333\\156.6667\end{bmatrix}, \Sigma=\begin{bmatrix}4433.3333 & 216.6667\\216.6667 & 33.3333\end{bmatrix}\right)}{p\left(X_1=100, X_2=225\right)}= \frac{3.4783\times 10^{-48}}{p\left(X_1=100, X_2=225\right)}$$

$$p\left(C=1 \mid X_1=100, X_2=225\right)= \frac{p\left(C=1\right)p\left(X_1=100, X_2=225 \mid C=1\right)}{p\left(X_1=100, X_2=225\right)}= \frac{\frac{1}{2}N\left(\begin{bmatrix}100\\225\end{bmatrix} \mid \mu=\begin{bmatrix}80\\203.3333\end{bmatrix}, \Sigma=\begin{bmatrix}100 & 50\\50 & 233.3333\end{bmatrix}\right)}{p\left(X_1=100, X_2=225\right)}= \frac{0.0001}{p\left(X_1=100, X_2=225\right)}$$

Comparing the numerators, we find that it is classified as an NBA player.

---

b) Compute the most probable class for the query vector, under the Naive Bayes assumption, using 1-dimensional Gaussians to model the likelihoods.

Solution:
We can estimate the priors in the same manner:

$$p\left(C=0\right)=\frac{1}{2}$$

$$p\left(C=1\right)=\frac{1}{2}$$

To estimate the likelihoods we will find the sample mean and standard deviation for each conditional distribution:

	$p\left(X_1 \mid C=0\right)$	$p\left(X_1 \mid C=1\right)$
μ	93.3333	80
σ	66.5832	10

	$p\left(X_2 \mid C=0\right)$	$p\left(X_2 \mid C=1\right)$
μ	156.6667	203.3333
σ	5.7735	15.2753

Now, for the query vector we have:

$$
\begin{aligned}
&p\left(C=0 \mid X_1=100, X_2=225\right)= \\
&\frac{p\left(C=0\right) p\left(X_1=100, X_2=225 \mid C=0\right)}{p\left(X_1=100, X_2=225\right)}= \\
&\frac{p\left(C=0\right) p\left(X_1=100 \mid C=0\right) p\left(X_2=225 \mid C=0\right)}{p\left(X_1=100, X_2=225\right)}= \\
&\frac{\frac{1}{2} N\left(100 \mid \mu=93.3333, \sigma=66.5832\right) N\left(225 \mid \mu=156.6667, \sigma=5.7735\right)}{p\left(X_1=100, X_2=225\right)}= \\
&\frac{7.8542 \times 10^{-35}}{p\left(X_1=100, X_2=225\right)}
\end{aligned}
$$

$$
\begin{aligned}
&p\left(C=1 \mid X_1=100, X_2=225\right)= \\
&\frac{p\left(C=1\right) p\left(X_1=100, X_2=225 \mid C=1\right)}{p\left(X_1=100, X_2=225\right)}= \\
&\frac{p\left(C=1\right) p\left(X_1=100 \mid C=1\right) p\left(X_2=225 \mid C=1\right)}{p\left(X_1=100, X_2=225\right)}= \\
&\frac{\frac{1}{2} N\left(100 \mid \mu=80, \sigma=10\right) N\left(225 \mid \mu=203.3333, \sigma=15.2753\right)}{p\left(X_1=100, X_2=225\right)}= \\
&\frac{2.5783 \times 10^{-5}}{p\left(X_1=100, X_2=225\right)}
\end{aligned}
$$

Comparing the numerators we determine that it is an NBA player.

---

3 Thinking Questions

a) Assuming training examples with d boolean features, how many parameters do you have to estimate if you make no assumptions about how the data is distributed? What about if you make the Naive Bayes assumption?

b) Is Naive Bayes a linear classifier?