Statistical Learning, Tutorato #3

Veronica Vinciotti, Marco Chierici

March 31, 2023

Exercise 1

This is an exercise on Bayesian inference, specifically an application to the Beta-Binomial derivation discussed at lectures.

A seller of video games is trying to decide how many copies of a particular game to have on hand for the upcoming holiday season. To get an idea of what the demand might be, they collect data on 40 potential customers and find that 2 would buy a copy. It is natural to assume a Binomial distribution for X: number of customers willing to buy the game, i.e. $X \sim \text{Binomial}(40, \theta)$, and the interest is in estimating θ.

Bayesian inference needs a specification of a prior distribution. Below are two options:

- The seller has prior information from a different video game. In particular, they decide on a prior distribution $\theta \sim \text{Beta}(2, 4)$. On the basis of this and of the data collected, what is the posterior distribution of θ? And what is the posterior estimate of θ?
- Assume that the prior information that the seller has is that on average 4% of customers may buy this product. Noting that this value corresponds to the mean of a $\text{Beta}(\alpha, \beta)$ distribution, i.e. $\frac{\alpha}{\alpha+\beta}$, one could reach a value for α and β by fixing the denominato $\alpha + \beta$. This value may be thought of as representing the "total number of customers" in the data that one attributes to the prior information (in a way, how confident are we about our prior information?). Let us say that the seller assigns 6 imaginary customers to the prior information. What would be the prior distribution then? And how would this change the posterior distribution compared to the earlier result?

Exercise 2

In this exercise, we explore another conjugate combination, that of a Poisson distribution with a Gamma prior on the λ parameter.

- Suppose that $X_1, \ldots, X_n$ are a random sample from a Poisson distribution with unknown parameter $\theta > 0$, so $f(x, \theta) = e^{-\theta}\frac{\theta^x}{x!}$, for $x = 0, 1, \ldots$. We assume that θ has a $\text{Gamma}(r, \lambda)$ prior distribution, i.e. $h(\theta) = \frac{\lambda^r}{\Gamma(r)}\theta^{r-1}e^{-\lambda\theta}$, for $\theta > 0$. Find the posterior distribution of θ.
- Suppose that we observe the number of customers arriving at a store in a specific one-hour period. The following data is collected on four days: $X_1 = 2$, $X_2 = 5$, $X_3 = 4$, $X_4 = 6$. Assuming a Poisson distribution for the data and a Gamma(3,2) distribution for the parameter θ of the Poisson distribution, use the first part to find the posterior distribution of θ.
- Use the posterior distribution of part 2 to derive an estimate of θ from the data. How close is this estimate to the maximum likelihood estimate?

Hints

- First calculate the likelihood and multiply by the prior
- In order to recognize the posterior distribution, you can ignore all normalizing constants and only concentrate on the terms in the "prior $\times$ likelihood" that depend on θ - can you spot what distribution it is? (The normalizing constants are just there to make sure that the distribution integrates to 1)
- In order to calculate the posterior estimate, you can use the fact that: if $Y \sim \text{Gamma}(r, \lambda)$, $E[Y] = \frac{r}{\lambda}$.

Exercise 3

This is a theoretical exercise on Naïve Bayes, with an application in R.

Consider the following vector of binary attributes: shortbread, lager, whiskey, porridge, football. An observation $\mathbf{x}$=(1,0,1,1,0) would describe that a person likes shortbread, does not like lager, drinks whiskey, eats porridge and has not watched England play football. Together with each vector, there is a label describing the nationality of the person (English or Scottish). The following data is collected on 13 people:

shortbread	0	1	1	1	0	0	1	1	1	1	1	1	1
lager	0	0	1	1	1	0	0	1	1	1	1	0	1
whiskey	1	1	0	0	0	0	0	0	1	0	0	1	1
porridge	1	1	0	0	0	1	1	0	1	1	1	1	0
football	1	0	1	0	1	0	1	1	0	0	1	0	0
nationality	E	E	E	E	E	E	S	S	S	S	S	S	S

- Using a Naïve Bayes classifier, estimate the probability that someone is Scottish based on the observation x=(1,0,1,1,0). Thus classify this person to one of the two classes, by comparing this probability with the default 0.5 threshold.
- Confirm your calculations by repeating the analysis in the previous question using the function `naiveBayes` from the R package `e1071`.

Hints

In R, the Naïve Bayes (NB) classifier is included in the package `e1071`. The syntax is `naiveBayes(formula, data)` for model fitting, and the usual `predict(fit, newdata)` for predicting on new data. A fitted `naiveBayes` object stores the conditional probabilities for each feature, together with the *a priori* probabilities. To compute the posterior probabilities, you call `predict()` with the argument `type="raw"`.

Exercise 4

In this exercise, you will use a Naïve Bayes classifier on the classical "Titanic" dataset (`data(Titanic)`), to predict whether a passenger survived or not. The dataset contains information about 2,200 RMS Titanic passengers summarised according to four factors:

- economic status (1st class, 2nd class, 3rd class, crew);
- gender (male, female);
- age category (child, adult);
- survived (yes, no).

```
##   Class    Sex   Age Survived Freq
## 1   1st   Male Child       No    0
## 2   2nd   Male Child       No    0
## 3   3rd   Male Child       No   35
## 4  Crew   Male Child       No    0
## 5   1st Female Child       No    0
## 6   2nd Female Child       No    0
```

Notice how the data are already summarized, so your first task will be to expand it based on the counts (`Freq`).

After you prepared the data for modeling, split into train/test, then fit a Naïve Bayes classifier to predict `Survived` from the other predictors. Evaluate the confusion matrix, the accuracy (global and per class), and the posterior probabilities.