Basic validation lab

Basic validation methodologies and hyperparameter optimization

Bruno Álvarez
Jan Izquierdo
Sergi Ocaña
20/10/2024

Abstract

The methods described are to differentiate between prostate cancer and control groups using mass spectrometry data. In the analysis is used k-Nearest Neighbors (k-NN), using a dataset from patients with prostate cancer and a control group and benign prostatic hyperplasia. Preprocessing steps were done like averaging replicate spectra and log transformation of intensity values. The model was trained on those three classes, prostate cancer and control and benign prostatic hyperplasia. The model has demonstrated effective classification rates.

Introduction

We base our code on a k-Nearest neighbors model that uses data that contains 327 subjects and 654 mass spectra (2 per subject). The subjects can belong to 3 groups: patients with prostate cancer, patients with prostatic hyperplasia and a control group.

The motivation for using k-NN in this context is its simplicity and effectiveness in handling mass spectrometry sets. By using the k-NN algorithm, we can classify subjects making it easier to differentiate between cancerous and non-cancerous conditions

Dataset description and methods

We use the str() method to display the structure of the created datasets, and in medical_cond we also use the levels function to see which discrete variables it contains.

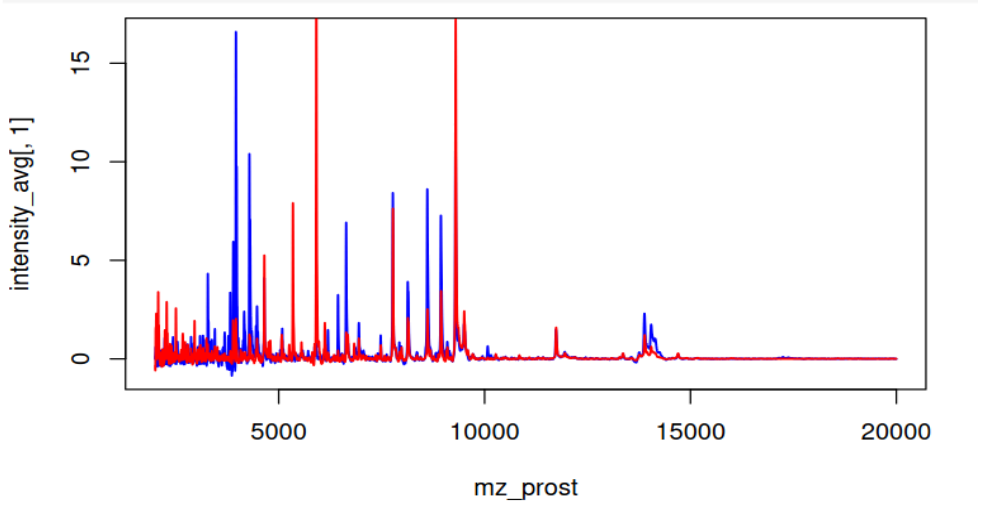
We use the following methods to find out.

```
str(mz_prost)
## num [1:10523] 2000 2001 2002 2003 2003 ...
str(intensity_with_replicates)
## num [1:10523, 1:654] 0.2607 0.0335 0.0617 0.0796 0.0547 ...
## - attr(*, "dimnames")=List of 2
## ..$ : chr [1:10523] "4902" "4903" "4904" "4905" ...
## ..$ : chr [1:654] "X22451A" "X22451B" "X22705A" "X22705B" ...
str(medical_cond)
## Factor w/ 3 levels "bph","control",..: 2 2 2 2 2 2 2 2 2 2 ...
levels(medical_cond)
## [1] "bph" "control" "pca"
table(medical_cond)
## medical_cond
## bph control pca
## 156 162 336
```

We can see that mz_prost is a list of continuous values, that intensity_with_replicates is a data frame containing continuous variables that associate chromosomes and medical_cond contains the class of the data, which we can see that it can be bph, control or pca. From the medial_cond table we can also see that the dataset is unbalanced, as pca is represented in a 2X factor compared to the rest.

To analyze what the data looks like

```
plot(mz_prost,intensity_avg[,1], type="l", lty="solid", lwd=1.5,
col="blue")
lines(mz_prost,intensity_avg[,100],
type="l",lty="solid",lwd=1.5,col="red")
```



Results

The analysis has revealed several things, so let's break it in sections to progressively understand it. First of all, the dataset has a total of 327 subjects with a total of 654 mass spectra, so we have 2 spectra for each subject. Each subject falls into one of the following categories: prostate cancer, prostatic hyperplasia and the control group. When we take a look at the distribution of the different classes we see that 156 of them are of prostatic hyperplasia, 162 are the control group and 336 are prostate cancer. We can see we have a rather unbalanced distribution since prostate cancer is over represented, occuring twice the times as the other categories. Also, the information we have about each case is “mz_prost” which contains the mass-to-charge ratio, “intensity_with_replicates” which contains the values of intensity for each mass spectrum, and last “medical_cond” which indicate the type of condition (bhp, control or pca). The mass spectra was visualized using line plots but beforehand we transformed the data logarithmically. Then we compared two different samples. Then we took a look at the distribution of the intensity, to do so we made use of a histogram making a logarithmic transformation of the intensities in order to be able to compare the values across samples. Taking into account the previous we used it as a base to design the k-NN model for classification. So if we get inside the model design, it uses 5-fold cross-validation to assess its performance and different values of k to see which one worked better (1, 3, 5, 7, 9).

The accuracy was the main thing taken into account to determine if the model was performing accurately or not. For each fold this accuracy was camputed and averaged between the 5 folds:

k = 1: Mean accuracy of 0.65

k = 3: Mean accuracy of 0.65

k = 7: Mean accuracy of 0.71

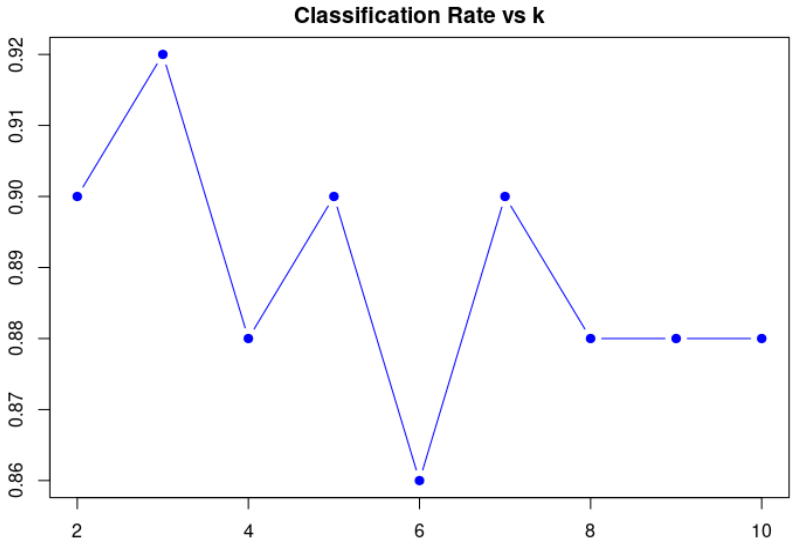
k = 9: Mean accuracy of 0.70

k = 5: Mean accuracy of 0.74

From all the tested values the one better performing was k = 5, which had an accuracy of 0.74. So the performance plateaued at k = 5 which means increasing it would make no sense, however the model overall had a consistent performance in the different values of k, with all configurations achieving the 74% of accuracy.

Further questions and exercises Build a model without benign prostatic hypertrophy and characterize the performance with 2 classes only. Now see what happens when you try to classify these patients that are not present in the training set.

From the original dataset we excluded benign prostatic hyperplasia (BPH) cases to see the performance in this case. After training and testing the model, the confusion matrix shows very accurate results, with 33 out of 34 "pca" cases and 14 out of 16 "control" cases correctly classified, a classification rate of 88% for k= 10. Even though the good results we look for the optimal k, which resulted to be 3 as we see in the plot. We may consider taking k = 5 in order to not get biased by only 3 neighbors.



Classification rate in function of k in the k-Nearest Neighbors (k-NN) model

Conclusion

What we can conclude from the model is that since it achieves a 95% of accuracy in all folds and its plateaus from k=5 is robust enough and the classes are separated correctly in the feature space.

The consistency on the performance from k = 5 to bigger values could mean that the model is not too sensitive to the choice of k in this range.

Based on the mass spectrometry data, these findings show that the k-NN model was very successful in differentiating between benign prostatic hyperplasia, prostate cancer, and control samples. The model appears to have good generalization capabilities based on its ability to maintain high accuracy across a range of k values.