Naive Bayes Classifier

Guillaume Wisniewski
guillaume.wisniewski@u-paris.fr

September 2021

1 Goal

We are aiming to train a classifier, implementing the Naive Bayes approach, to predict whether a mail is a spam or a ham. To train and evaluate our classifier we will consider the `ling-spam` dataset that contains spam messages and messages from the Linguist list.[1]

2 Dataset

1. Write a function that takes as parameter the name of the directory containing the `ling-spam` and returns a list of examples. Each examples is made of an observation (the list of words extracted from the mail) and a label (either `ham` or `spam`).
2. How many examples are there in the corpus? Represent the label distribution.
3. Write a function that split the corpus into a train and a test sets. The function will take as parameters the list of examples and the proportion of examples that must be included into the train set.

3 Naive Bayes classifier

In the Naive Bayes classifier the conditional distribution over the class variable y is estimated by:

$$p(y|\mathbf{x}) \propto p(y) \cdot \prod_{i=1}^{k} p(x_i|y) \tag{1}$$

[1] See this article for further details.

where $\propto$ denotes proportionality and $\mathbf{x} = (x_i)_{i=1}^{k}$ is an observation represented by a vector of k components.[2]

The computation of $p(y|\mathbf{x})$ involves the multiplication of a large number of tiny numbers (there as many terms in the multiplication as words in an observation, each of the terms is a probability and, consequently, smaller than 1), which can result in numerical stability issues (computers are not able to represent small real numbers). That is why, you have to manipulate log-probability in our program rather than probabilities.

4. Use Equation 1 to express $\log p(y|x)$ as a function of $p(x_i|y)$ and $p(y)$. Why can we use $\log p(y|x)$ in the MAP decision rule instead of $p(y|x)$.

5. Write a function that takes as input a list of examples a returns a dictionary of dictionaries that maps every pairs (`word`, `label`) to the number of occurrences of `word` in mails labeled by `label`

6. Write down a class[3] with:
   - a `fit` method that will estimate the parameters of the Naive Bayes classifier;
   - a `predict` method that will take as input an observation (a list of words) and return the predicted label (either `spam` or `ham`);
   - a `score` method that will take as input a list of annotated examples and return the proportion of correctly predicted labels. This corresponds to the classifier *accuracy*.

7. What is the accuracy of a naive Bayes classifier when the train set is made of 80% of all the examples. How is the accuracy related to the 0/1 loss?

4 Evaluation

8. Compute the *confusion matrix* of the classifier trained in the previous section. A confusion matrix is a $n \times n$ matrix, where n is the number of possible label. An entry $m_{i,j}$ of the confusion matrix indicates the number of times an examples with gold label i has been classified as j. What is a confusion matrix useful for? Interpret.

9. Compute the accuracy of the classifier for different size of the training set (e.g. starting with a 80:20 train-test split, consider 10%, 20%, 30%, ... of the training data to estimate the classifier parameters and estimate the classifier performance on the test set). Why do we have to consider a single test set (rather than, for instance, consider all the remaining data to evaluate the classifier performance). Plot the classifier performance with respect to the size to the train set. Interpret.

[2] In the next classes we will see that these components are usually called 'features'.

[3] Defining a class is a simple way to 'link' the different methods together. If you are not at ease with OOP, you can write two separated methods.

10. Generate 100 train-test splits (all with a 80:20 proportion), train and evaluate a classifier on each of this split. What is the smallest and largest accuracies achieved? Plot the resulting accuracy distribution. Interpret.

11. Find, for each label y, the ten word with the largest probability $p(w|y)$. Interpret.