Introduction to Data Science Project

Instructions

This project considers the use of classifiers to classify emails. Answer all questions. Save the R code you have written as a R script. All plots and discussion should be saved separately as a document. Submit all files via email to heng@essec.edu by 10pm on 14 June 2021. Late submissions will not be accepted.

Naive Bayes classifier on Windows dataset

We will fit a naive Bayes classifier to classify emails into two categories, i.e. pertaining to the X Window System ($Y = 1$) that is common in Unix operating systems, or the Microsoft Windows operating system ($Y = 2$). The dataset `windows.RData` can be found in the `Project` folder on Moodle. We adopt a bag of words model to extract features of an email, i.e. $X = (X_1, X_2, \ldots, X_d) \in \{0, 1\}^d$ is a binary vector denoting absence ($X_i = 0$) or presence ($X_i = 1$) of each word in the object `vocab`. These features and the corresponding class labels (last column) are stored in `dataset`.

1 Handling dataset

1. We will use the first $N = 900$ emails as our training dataset and the last $M = 900$ emails as our testing dataset. Implement this split and store the training and testing datasets as `training` and `testing`, respectively. [2 marks]

2. What class of object is `vocab`? [1 mark]

3. How many words are there in the bag of words? [1 mark]

4. Which positions in the bag of words correspond to the words "retrieval" and "subject"? [2 marks]

5. Compute the sample proportions of training emails that pertain to X Window and Microsoft Windows? [2 marks]

6. For training emails that pertain to X Window, compute the sample proportions of these emails that contain the words "retrieval" and "subject". [3 marks]

7. For training emails that pertain to MS Window, compute the sample proportion of these emails that contain the words "retrieval" and "subject". [3 marks]

2 Fitting model

1. Discuss whether the assumption that the features $X = (X_1, X_2, \ldots, X_d)$ are conditionally independent given the class label is sensible? [2 marks]

2. Using the `naiveBayes` function in the `e1071` package, fit a naive Bayes classifier on the training dataset. [2 marks]

3. Using the output of the `naiveBayes` function, what is the estimated prior distribution of the classes? Does this agree with your results in Question 1.5? [2 marks]

4. Using the output of the `naiveBayes` function, what is the estimated class conditional distributions of the features corresponding to the words "retrieval" and "subject"? Does this agree with your results in Questions 1.6–1.7? [3 marks]

5. Given the estimates obtained in Question 2.4, show analytically that the naive Bayes classifier will predict both classes are impossible if the word "retrieval" is present in a testing email. Explain your arguments clearly. [3 marks]

6. Given the estimates obtained in Question 2.4, show analytically that the naive Bayes classifier will predict both classes are impossible if the word "subject" is not present in a testing email. Explain your arguments clearly. [3 marks]

7. The behavior described in Questions 2.5–2.6 is known as overfitting. Explain why the naive Bayes classifier is overfitting the words "retrieval" and "subject"? [2 marks]

8. Investigate whether other words in `vocab` suffer from overfitting. Provide a short summary of your findings. [5 marks]

9. To mitigate the difficulties mentioned in Questions 2.5–2.6, one could consider prediction with modified probabilities, i.e. replace probabilities that are zero with a small positive threshold. Implement this prediction rule by specifying the `threshold` argument in the `predict` function. Compute the misclassification rates for thresholds $10^{-9}, 10^{-6}, 10^{-3}$ on the testing dataset, and explain which threshold we should prefer? [3 marks]

10. A principled approach to deal with overfitting, without modifying probabilities as considered in Question 2.9, is to learn model parameters using Bayesian inference. A specific instance of Bayesian inference in this setting amounts to adding the following four artificial data points to the training dataset:
(i) an email classified as X Window System that contains none of the words in the bag of words;
(ii) an email classified as X Window System that contains all the words in the bag of words;
(iii) an email classified as Microsoft Windows that contains none of the words in the bag of words;
(iv) an email classified as Microsoft Windows that contains all the words in the bag of words;
and fitting a naive Bayes classifier on the augmented dataset. Implement this approach and compute the misclassification rate on the testing dataset. [6 marks]

3 Feature selection

In this section, we will consider a methodology to remove words from the bag of words that are not so relevant for the classification task. The feature selection procedure first measures the relevance of each word by computing the mutual information between each feature $X_i \in \{0, 1\}$ and the class label $Y \in \{1, 2\}$, defined as

$$I(X_i, Y) = \sum_{x \in \{0,1\}} \sum_{y \in \{1,2\}} \log \left(\frac{P(X_i = x, Y = y)}{P(X_i = x)P(Y = y)} \right) P(X_i = x, Y = y)$$

for word $i = 1, 2, \ldots, d$. We then rank words according to this measure of relevance, and select the top $K \in \{1, 2, \ldots, d\}$ features to fit the naive Bayes classifier.

1. Show analytically that $I(X_i, Y) = 0$ if X_i and Y are independent. Explain why the mutual information provides a sensible measure of word relevance for the classification task. [2 marks]

2. Show analytically that

$$\begin{aligned} I(X_i, Y) = &\log \left(\frac{1 - \theta_{i,1}}{1 - (1 - \pi)\theta_{i,1} - \pi\theta_{i,2}} \right) (1 - \pi)(1 - \theta_{i,1}) \\ &+ \log \left(\frac{1 - \theta_{i,2}}{1 - (1 - \pi)\theta_{i,1} - \pi\theta_{i,2}} \right) \pi(1 - \theta_{i,2}) \\ &+ \log \left(\frac{\theta_{i,1}}{(1 - \pi)\theta_{i,1} + \pi\theta_{i,2}} \right) (1 - \pi)\theta_{i,1} \\ &+ \log \left(\frac{\theta_{i,2}}{(1 - \pi)\theta_{i,1} + \pi\theta_{i,2}} \right) \pi\theta_{i,2} \end{aligned}$$

   where $P(Y = 2) = \pi$, $P(X_i = 1|Y = 1) = \theta_{i,1}$ and $P(X_i = 1|Y = 2) = \theta_{i,2}$. [3 marks]

3. Using the parameters estimated in Question 2.10 and the expression in Question 3.2, compute the mutual information for each word in the bag of words. [6 marks]

4. What are the words with the top five highest mutual information? By running a Google search on these words (if they are unfamiliar), explain why they are most relevant for this classification task. [4 marks]

5. Using the augmented training dataset in Question 2.10, compute the misclassification rate when selecting only the top K features to fit the naive Bayes classifier for all $K \in \{1, 2, \ldots, d\}$. Plot the misclassification rate against K. Report the value of K that minimizes the misclassification rate and briefly discuss your findings [7 marks]