Shiraz University

Statistical Pattern Recognition (Fall 2020)

Homework#3: Bayesian Classification, Naïve Bayes Classification

Due date: 5th January 2021

In order to do this homework, you have to go through Bayesian and Naïve Bayes classification theories and concepts.

Part A: Bayesian Classification (Quadratic Multiclass Classification)

For this part of homework, you have to generate two datasets for three classes each with 500 samples from three Gaussian distribution for each dataset described below:

Dataset#1:

$$\text{Class1:} \quad \mu = \begin{pmatrix} 3 \\ 6 \end{pmatrix} \quad \sum = \begin{pmatrix} 1.5 & 0 \\ 0 & 1.5 \end{pmatrix}$$

$$\text{Class2:} \quad \mu = \begin{pmatrix} 5 \\ 4 \end{pmatrix} \quad \sum = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$$

$$\text{Class3:} \quad \mu = \begin{pmatrix} 6 \\ 6 \end{pmatrix} \quad \sum = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

Dataset#2:

$$\text{Class1:} \quad \mu = \begin{pmatrix} 3 \\ 6 \end{pmatrix} \quad \sum = \begin{pmatrix} 1.5 & 0.1 \\ 0.1 & 0.5 \end{pmatrix}$$

$$\text{Class2:} \quad \mu = \begin{pmatrix} 5 \\ 4 \end{pmatrix} \quad \sum = \begin{pmatrix} 1 & -0.20 \\ -0.20 & 2 \end{pmatrix}$$

$$\text{Class3:} \quad \mu = \begin{pmatrix} 6 \\ 6 \end{pmatrix} \quad \sum = \begin{pmatrix} 2 & -0.25 \\ -0.25 & 1.5 \end{pmatrix}$$

- Consider the first 80% of the data in **each class** for train and the rest 20% for test.
- Use a Bayesian classifier to classify both the train and test datasets and calculate both accuracies (for each dataset separately).
- Plot the decision boundary and classification results while representing the misclassified samples with a different color or shape (for each dataset separately).
- Plot estimated PDFs. (3D for each dataset separately)
- Contour estimated PDFs along with the decision boundary. (2D for each dataset separately)
- What is the main difference between two datasets? Explain your answer using your results and plots. compare this part with partb of homework#2?

Part B: Naïve Bayes Classification

Dataset: Sentiment Labelled Sentences

https://archive.ics.uci.edu/ml/datasets/Sentiment+Labelled+Sentences

In this part of the homework you have to implement a Naive Bayes classifier to detect *positive* and *negative* sentiment from reviews. The "Sentiment Labelled Sentences" dataset is provided and you have to refer to the "readme" file so as to find out more about this dataset. this dataset contains sentences labelled with positive or negative sentiment, extracted from reviews of products, movies, and restaurants. The sentences come from three different websites: imdb.com, amazon.com, and yelp.com. For each website, there exist 500 positive and 500 negative sentences.

- Consider the first 80% of the data in **each website review** for train and the rest 20% of **each website review** for test. (for each website review separately)
- Classify the data using Naïve Bayes classifier and report the total and per-class accuracies (train and test for each website review separately).
- Use each word as a feature and construct a dictionary. (for each website review separately. So, you will have 3 dictionaries)
- if you multiply many small probabilities you may run into problems with numeric precision, what is the problem? To handle it, we recommend that you compute the logarithms of the probabilities instead of the probabilities. Explain why this approach can help?
- If you need, you should use **Laplace Smoothing** as a part of your job.
- What is the base assumption of Naïve Bayes classifier? Why is it important?

Notes:

- **Pay extra attention to the due date. It will not extend.**
- **Be advised that submissions after the deadline would not grade.**
- **Your implementation should be functional.**
- **Prepare your full report in PDF format and include the figures and results.**
- **Do not use sklearn or any similar library for Bayesian, Naïve Bayes classification and write your own code.**
- **The allowed programming languages are any language and feel free.**
- **Feel free for using sklearn in python for split train and test dataset.**
- **Feel free for using numpy, pandas, or any regular library in python.**
- **Submit your assignment using a zipped file with the name of "StdNum_FirstName_LastName.zip" to csehws.shirazu.ac.ir or for someone has problem with it to compuscien@gmail.com with SPR-Fall 2020-HW#3 subject.**