Homework 2

CSE3102: Applied Probability for Computer Science

Due date: 2019/11/28

Note: Use any programming language (C/C++, JAVA, Matlab, etc.). You are not allowed to use a library except data loader and graph visualization. Absolutely no copying.

– Iris flower identification using Bayesian Classifier –

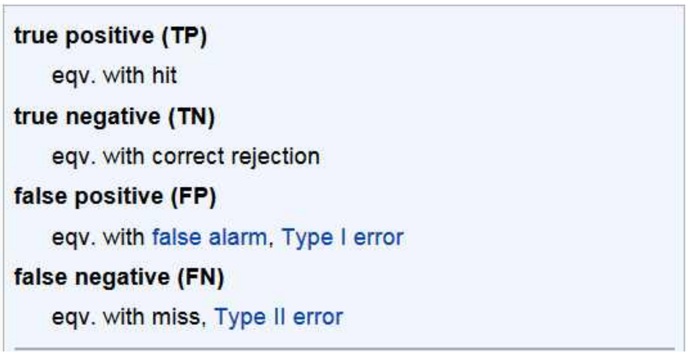
Part A. Preliminaries

1. (10pts) Cross Validation

- Summarize the k-fold cross-validation method for evaluating a classifier (less than one page)
- Important: Use your own explanation. Absolutely DO NOT copy any paragraph from any source (paper, internet, blog, etc.) This will result in 0 point for your entire homework.

2. (0pts) Testing your classifier

- Understand the concept of the precision-recall. I recommend you search further literature in addition to the following definition.

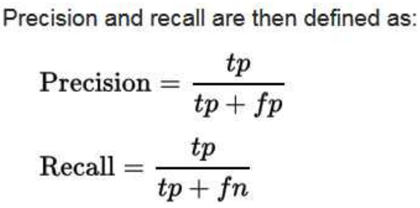


true positive (TP)
eqv. with hit

true negative (TN)
eqv. with correct rejection

false positive (FP)
eqv. with false alarm, Type I error

false negative (FN)
eqv. with miss, Type II error



Precision and recall are then defined as:

$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

3. (0pts) Acquiring dataset

(1) Data file: "iris.data" on the *I-class*.

- The data file has 150 instances (50 in each of three classes)
- Each instance consists of 4 numeric attributes and a label: sepal length in cm, sepal width in cm, petal length in cm, petal width in cm, class categories(Iris Setosa, Iris Versicolour, Iris Virginica)

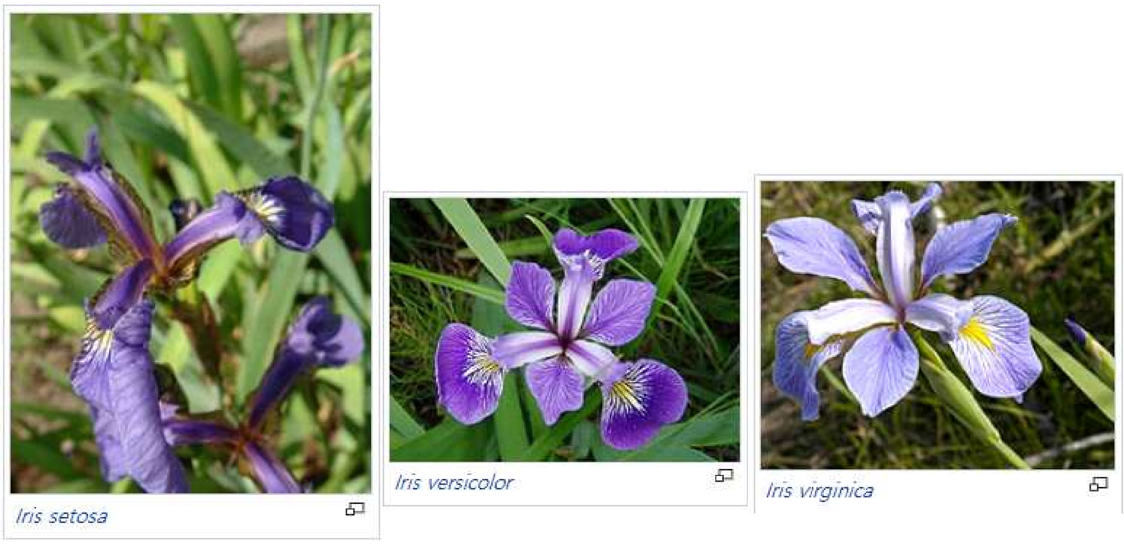


Fig.1: Example images for three classes of iris

(2) More information about the dataset is available in “iris.names”

Part B. Iris classification using text data

1. (20pts) For four different features, show four separate histograms where each shows three classes in different colors. Find your optimal decision boundaries to minimize the costs. Report your analysis about the result.

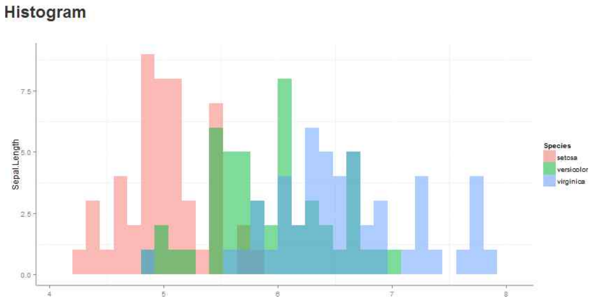


Fig.1: Example 1D plot for three classes of iris

2. (25pts) Visualize the data in 2D space to show the distribution of the attributes. You should show all possible combinations of pairs of attributes, six cases in total. Find your optimal decision boundary in each case to minimize the costs. Report your analysis about the result.

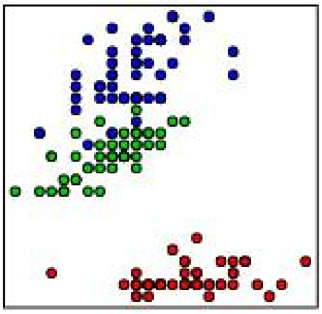


Fig.2: Example plot of (sepal width, petal width)

3. (20pts) Write a code to classify three classes of iris based on the Bayesian decision rule; by finding a pmf(probability mass function) of likelihood and defining a prior. Evaluate your classifier by computing precision and recall. Use 5−fold cross validation for evaluation. You may set or test your own priors.

4. (25pts) Write a code to classify three classes of iris based on the Bayesian decision rule; by modeling with a Gaussian distribution. Evaluate your classifier by computing precision and recall. Use 5−fold cross validation for evaluation. You may set or test your own priors.

5. (Extra credit. 30pts max) Extend your work by using a multi−variate Gaussian model. Discuss about your results.

What to submit: A report with (1) the answers from the problems and (2) all source codes(not source codes, only full solution). (3) Please check your own code perfectly and attach procedure of execution. If your code does not run, no point.

If you have any question,

send an email to TA (Jongmin Lee, 24jj@naver.com)