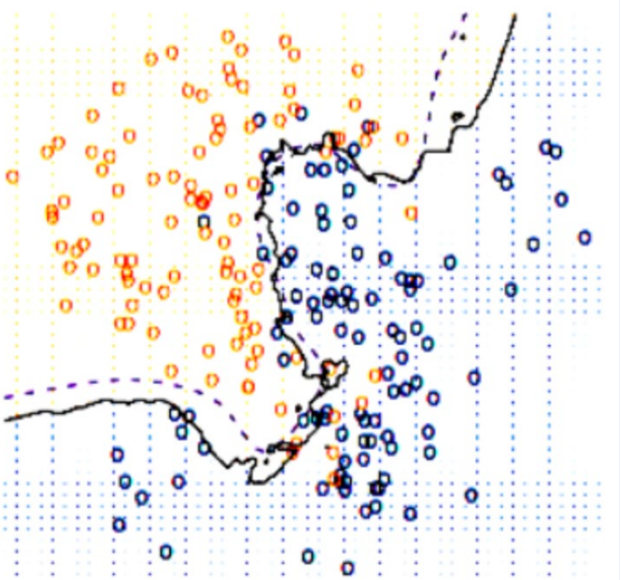
COMP 642 Assignment --- MODULE 4

Assignment

1. KNN
   a. We discussed data normalization (also known as data standardization). Which of the following statements are true for doing data standardization on training and test sets.
      i. Calculate the mean and standard deviation of the whole dataset (training and test set). And apply the mean and standard deviation to both training set and test set.
      ii. Calculate the mean and standard deviation of the training dataset. And apply the mean and standard deviation of training set to both training set and test set.
      iii. Calculate the mean and standard deviation of the training dataset. And do not need to do the data standardization for test set.
      iv. Calculate the mean and standard deviation of the training dataset and test set respectively. And apply the the mean and standard deviation to training and test set respectively.

   b. Which value of k in the k-nearest neighbors algorithm generates the solid decision boundary depicted here? There are only 2 classes. (Ignore the dashed line, which is the Bayes decision boundary.) k = 1 k = 10 k = 2 k = 100.

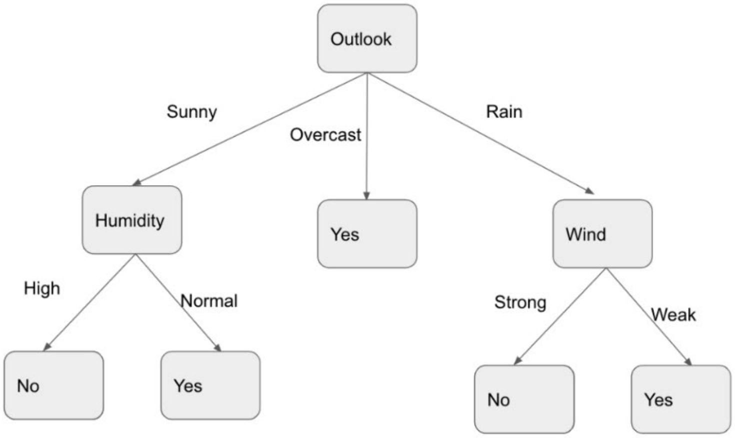


      i. K = 1
      ii. K = 2
      iii. K = 10
      iv. K = 100

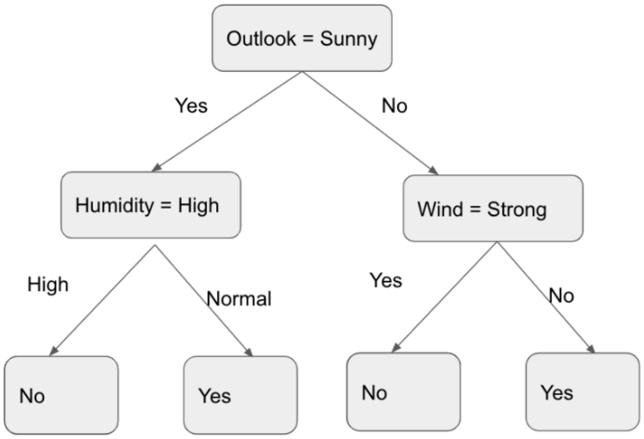
c. Go to knn/dataset folder, execute get_datasets.sh to get the CIFAR-10 object recognition dataset for the coding exercise. (On Mac, just type in ./get datasets.sh at the shell prompt.)
A new folder will be created and it will contain 50,000 labeled images for training and 10,000 labeled images for testing. There is a function to read this data in data_utils.py. Each image is a 32 × 32 array of RGB triples. It is preprocessed by subtracting the mean image from all images.

d. Fill out the corresponding code cells and answer the questions in hw_4_knn.ipynb.

2. Decision Tree
    a. One typical decision tree algorithm ID3 is used for binary classification. There is another decision tree algorithm named CART. The difference between CART and ID3 is CART does binary splits at every node, whereas ID3 exhausts one attribute once it is used.
    Example of ID3:



Example of CART:



The difference between these two algorithms is obvious by looking at the above tree structure. They use similar ways to calculate how to split a node. At each node of CART, the algorithm will iterate all the features and feature values and evaluate each split according to a cost function to select the feature and feature value with the largest cost reduction.

$$Reduction(Feature_i, Value_j) = cost(D) - [\frac{|D_{left}|}{|D|} * cost\left(D_{left}\right) + \frac{|D_{right}|}{|D|} * cost\left(D_{right}\right)]\,)$$

Consider a data set comprising 400 data points from class C1 and 400 data points from class C2. Suppose that a decision stump model A splits these into two leaves at the root node; one containing (300,100) and the other containing (100,300) where (n,m) denotes n points are from class C1 and m points are from class C2. Similarly a second decision stump model B splits the examples as (200,400) and (200,0). Calculate the reduction in cost using misclassification rate for models A and B. Which is the preferred split (model A or model B) according to the cost calculations?

Misclassification rate:

$$cost(D) = \frac{1}{|D|} \sum_{(x,y)\in D} I\ (y\ \neq \hat{y})$$

$$\hat{y} = majority\ label\ in\ D$$

b. If using the entropy to measure the cost, what is the answer to the question 2.a ?

Hint:
Entropy:

Let p = fraction of position examples in D

$$cost(D) = \ -p \log_2 p - (1-p)\log_2 (1-p)$$

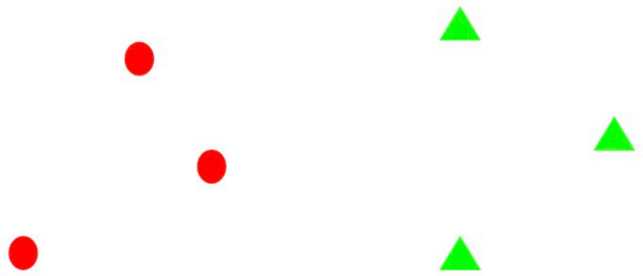
c. If using the Gini index to measure the cost, what is the answer to the question 2.a

Hint:
Gini Index:

$$cost(D) = \ 2p\ (1-p)$$

d. Go to decisiontree folder and execute the hw_4_dt.ipynb, fill the code cells and answer the questions in notebook.

3. Support Vector Machine
    a. Draw the possible decision boundary learned by svm model, and point out all the support vectors.



    b. Does changing the support vectors change the decision boundary? Does changing the non support vectors change the decision boundary?

    c. Go to svm folder and execute the hw_4_svm.ipynb, fill the code cells and answer the questions in the notebook.

Submit a .doc or .pdf with your written answers. Submit your Jupyter notebook. Submit a PDF of your Jupyter notebook.