3/9/2020

Ishita Kapoor

The University of texas at dallas

SVM and Decision Trees

**INTRODUCTION**

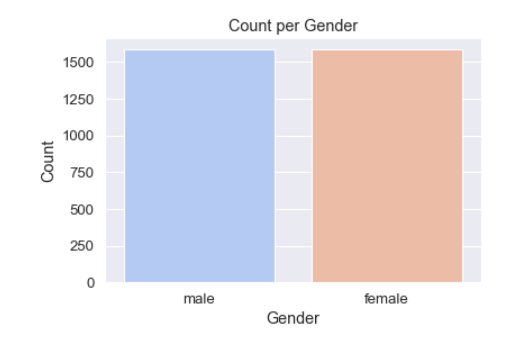
**Dataset Description and Summary Statistics**

Two datasets were used for the implementation of SVM and Decision trees.  
1. Kernel Performance dataset [SGEMM GPU](https://archive.ics.uci.edu/ml/datasets/SGEMM+GPU+kernel+performance)  
2. Voice Detection based on acoustic properties (Attached)

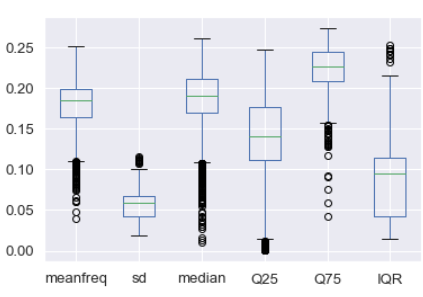
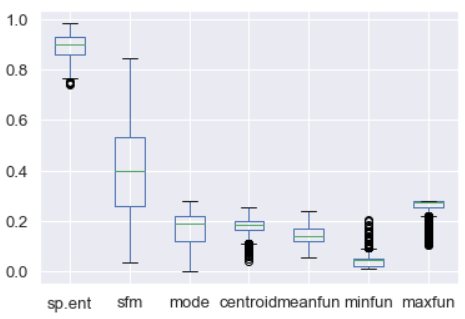
**Below is a brief description of the dataset:**1. The voice detection dataset consists of 3168 observations and 21 features  
2. There is one dependent variable, “Label” which has 2 distinct values (1,0). 0 is assigned to male whereas 1 is for female  
3. There are no null values in the entire dataset  
4. The data is normalized to bring all the variables to one common scale  
5. Data dictionary and dataset has been attached in the deliverable

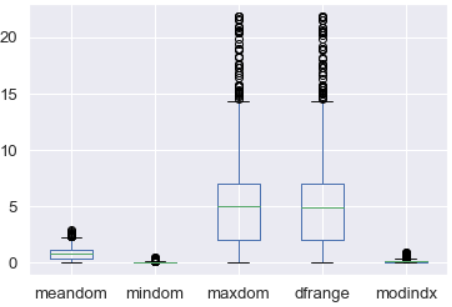
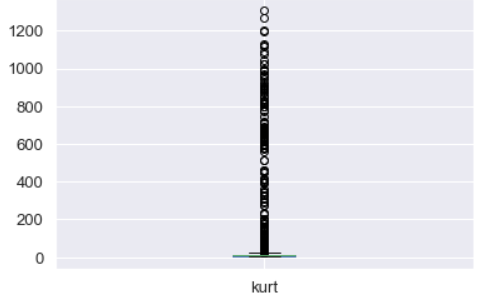
**Data Preparation and Exploratory Data Analysis for Voice Detection data:**

**Frequency Distribution Plot**



**Observation:** From the distribution plot, we observe that both male and female have 1,584 observations in the dataset. The data is equally distributed between both the genders

**Box Plot for Independent variables:**



**Observation:** When we look at the box plot, we can observe many outliers in almost all the variables. However, a possible reason for outliers might be due to the size of the data. In small dataset, values may be outlying - but not because these outlying values are Wrong but rather because the rest of the values clumps together more tightly as they should. So, the outlier is the only datum putting things right. Removing it would unnecessarily introduce bias, what is a rather bad thing.

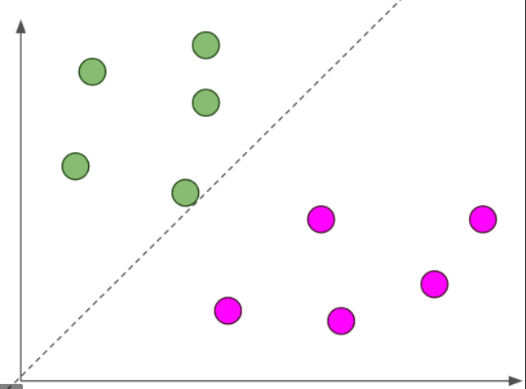
**Reason for choosing Gender Detection dataset:**

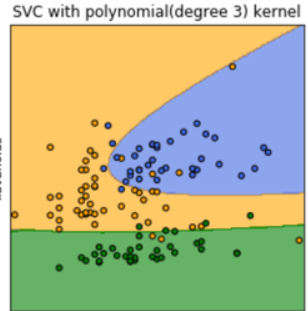
A gender detection may be useful in cases of a mobile healthcare system. For example, there are some pathologies, such as vocal fold cyst, which mainly occur in female patients. If there is an automatic method for gender detection embedded into the system, it is easy for a healthcare professional to assess and prescribe appropriate medication to the patient. The applications of gender detection have increased significantly due to the recent developments in speech recognition, human-computer interaction, and biometric security systems including authentication to access data, and security. Gender detection systems limit the search of an imposter to half of the space in many recognition and security systems, where the goal is the identification of a person. The dataset has 21 features and then SVM and decision trees are used as a classifier.

All the algorithms are run on both gender detection and kernel performance datasets. Various kernels are used to find the algorithm that has the best accuracy for the two datasets.

**Support Vector Machine:**

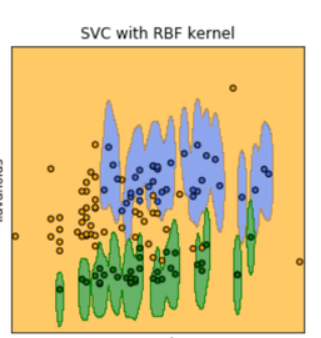
SVM is a supervised machine learning algorithm that can be used for classification. Various kernels such as linear, polynomial, rbf (gaussian) and exponential can be used to classify the data by drawing a hyperplane between the two classed. The hyperplane can be a straight line, curved line or a circle depending upon the data and the kernel used.

**EXPERIMENT 1:  
Experimenting with various kernels**In this experiment we will train the model using various kernel and observe the best performing kernel based on train/test accuracy.

1. **Linear Kernel:** The linear SVM classifier works by drawing a straight line between the classes. All the datapoints falling on one side of the line is labeled as one class and the datapoints falling on other side is labelled as other class
2. **Polynomial Kernel:**

The polynomial kernel looks not only at the given features of input samples to determine their similarity, but also combinations of these

Equation is:



1. **Gaussian Radial Basis Function:**Gaussian RBF (Radial Basis Function) is another popular Kernel method used in SVM models for more. RBF kernel is a function whose value depends on the distance from the origin or from some point.

Equation is:



**Below are the accuracy percentage for the above-mentioned kernels for both test and train dataset**  
We have split the dataset into train and test with train as 70% and test as 30%.

******Gender Detection Dataset: Kernel Performance dataset:**

**Observation:** From the table we observe that RBF has the maximum accuracy for both test and train dataset. This means that the given dataset is not linearly discriminated. The RBF kernel maps the non-linear dataset into a higher dimensional space where we can find a hyperplane that separated the data.

**Decision Trees:**

Decision tree is a supervised learning technique where the data is continuously split according to a certain parameter. It consists of nodes (test for the values of a certain attribute), branch (correspond to an outcome of a test and connect to another node) and leaf nodes (terminal nodes that predict the output). Here we have a classification tree where the output of the data set is 1(female) or 0(male).

**Experiment 1:**  
**Decision Tree without pruning:**An iterative process of splitting the data into partitions and then splitting it up further on each of the branches. We have split the data into train and test with train as 80% and test as 20%.

**Gender Detection Dataset:**

**Observation:** We observe similar accuracy to SVM for Gender Detection data. However, there is a 5% increase in accuracy for kernel performance, when **Kernel Performance dataset:** compared to SVM. The train accuracy Is 100% for both the datasets.

**Advantages of Decision Tree:**  
The computation time for Decision tree is extremely low even for a large dataset whereas the computation time for SVM was very high. This is because decision trees exclude unimportant features.

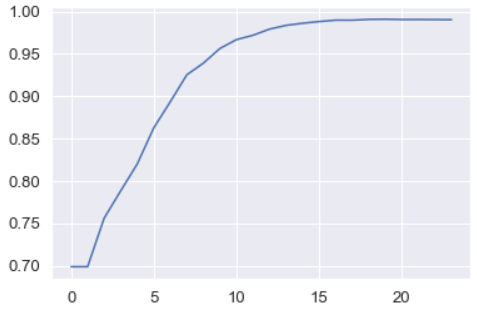
**Disadvantages of Decision tree:**  
Decision tress are often biased toward splits on features having many levels. There is a high chance of overfitting the data. To avoid overfitting, we limit the growth of the tree or prune the tree.

**Experiment 2:  
Decision Tree with pruning:**In this experiment, we test the original tree against the pruned versions of it. Leaf nodes are pruned away if the pruned tree is performing better on test data compared to the large tree.  
Below is a plot for accuracy of train/test vs various depths of the tree:

**Entropy and Gini**

**Gini** is less extensive computationally as there is no log term involved while calculating entropy. Therefore, gini is used for large datasets. We have used gini for the kernel performance dataset.

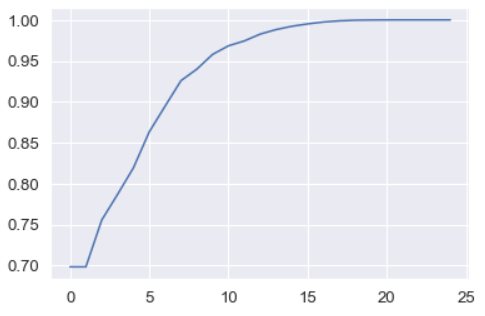
**Entropy** is more extensive computationally; therefore, it is used in smaller datasets. We have used entropy for gender detection dataset in our analysis.

**Kernel Performance Data  
Test data (Out sample)**

**Observation:** We trained the model for multiple depths (1-25). From the plot, we can observe a non-linear relationship between depth and accuracy. Accuracy starts increasing at first and then it becomes stagnant. The highest accuracy is obtained when **depth = 20**. This accuracy is also greater that the accuracy obtained for the original tree

Accuracy

Depth

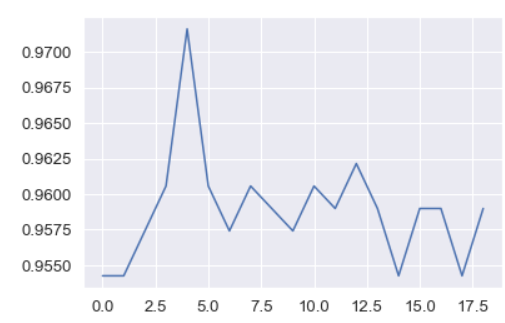
 **Train data (In sample)**

**Observation:** We can observe a similar trend in train data. We obtained 100% accuracy at **depth = 22.**

Accuracy

Depth

**Gender Detection Dataset**

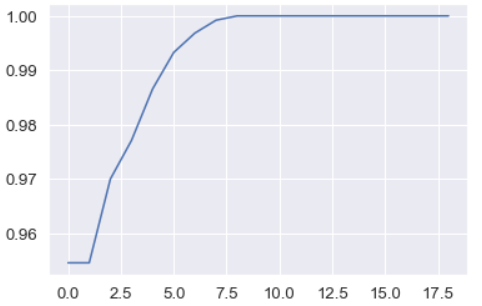
**Test data (Out sample)**

**Observation:** There is no pattern in the accuracy as we depth increases. At depth=5, we see the maximum accuracy. However, even at **depth=5,** the pruned tree is performing better than the original tree.

Accuracy

**Train Data (In sample)**

Depth



**Observation:** We can observe a non-linear relationship for train sample. We obtain 100% accuracy at **depth = 9.**

Accuracy

Depth

**Comparison of accuracy with and without pruning**

**Boosting**

Boosting refers to a family of algorithms which converts week learners to strong learners. In this algorithm, we have used Adaboost. The main idea behind adaboost is to fit a sequence of week learners on repeatedly modified versions of the data. The predictions from all of them are then combined through a weighted majority vote (or sum) to produce the final prediction. For each successive iteration, the sample weights are individually modified, and the learning algorithm is reapplied to the reweighted data. As iterations proceed, examples that are difficult to predict receive ever-increasing influence. Each subsequent weak learner is thereby forced to concentrate on the examples that are missed by the previous ones in the sequence.

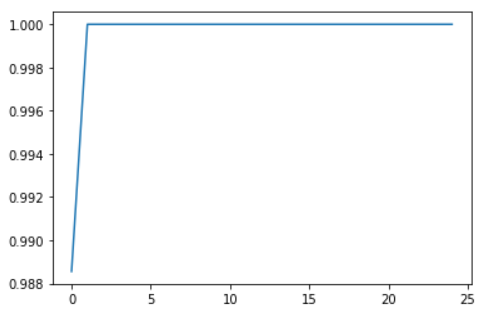
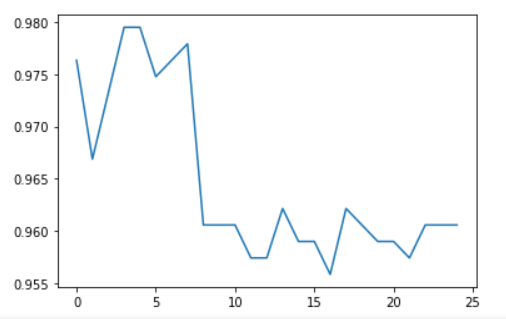
**Experiment 1:  
Boosting without pruning**

**Gender Detection Dataset**

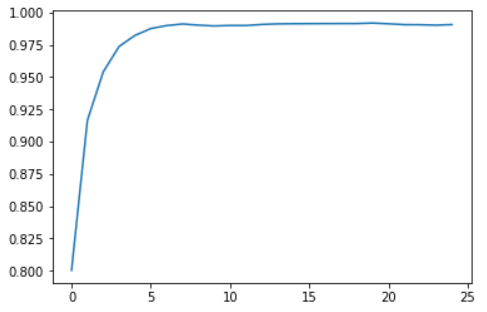
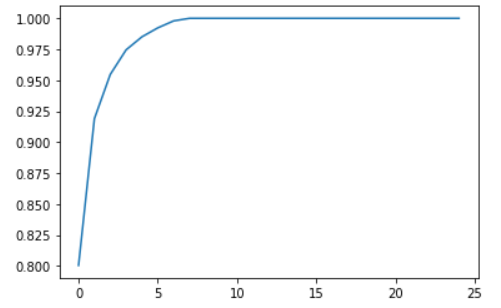
**Observation:** After applying adaboost, we observe a high accuracy for both **Kernel Performance Dataset** the datasets. The accuracy observed for boosting is similar to that without boosting.

**Experiment 2:  
Boosting with pruning and criteria:** In this experiment, we ran both the data for entropy and gini criteria. By default, the number of estimators considered in adaboost is 50. Below is plot showing accuracy vs dept (1,25) for both the datasets.  
 **Gender Detection Data**

**Test Data Accuracy**  **Train Data Accuracy**



**Observation:** We ran boosting algorithm for various depths and for criteria as entropy. We can observe that the accuracy is 100% for train at depth =2. For test, we can observe an accuracy of 97.3% at depth = 3.

**Kernel Performance Dataset  
 Test Data Accuracy Train Data Accuracy**

**Observation:** We ran boosting algorithm for various depths and for criteria as entropy. We can observe that the accuracy is 100% for train at depth =9. For test, we can observe an accuracy of 99.12% at depth = 8.

**Cross Validation:**

K-Fold CV is where a given data set is split into a *K* number of sections/folds where each fold is used as a testing set at some point. Let’s take the scenario of 5-Fold cross validation(K=5). Here, the data set is split into 5 folds. In the first iteration, the first fold is used to test the model and the rest are used to train the model. In the second iteration, 2nd fold is used as the testing set while the rest serve as the training set. This process is repeated until each fold of the 5 folds have been used as the testing set.

**Experiment 1:**    
In this experiment, we did cross validation for SVM for both the provided datasets.

**SVM:  
Gender Detection Dataset**



**Run Dataset**

 **Observation:  
Gender Detection:** We observe that the test accuracy is very low when compared to the train accuracy. This means that the model has overfit. Since we can see that the cross validation has a low accuracy when compared to the previous models, this means that the sampling is creating a bias. **Run dataset:** We observe a low test and train accuracy when compared to SVM without cross validation. This means that there is error in sampling.

**Experiment 2:**

**Decision Tree and Boosting**

**Gender Detection Dataset**



**Run Dataset**



**Observation:**

**SVM:**

We observe that the test and train accuracy for cross validation SVM is very low when compared to SVM without cross validation. The cross-validation sampling is creating a bias.

**Decision Tree:**    
**Gender Detection:** We observe that the test accuracy is very low when compared to the train accuracy. This means that the model has overfit. Since we can see that the cross validation has a low accuracy when compared to the previous models, this means that the sampling is creating a bias.

**Run Dataset:** Even in this dataset, we observe a low accuracy which is again due to sampling error.

**Conclusion:**

In the above experiments, we ran kernel performance and Gender detection dataset for various classification algorithms. For the above two datasets, we get the maximum accuracy in **boosting with pruning**. In boosting, we allow many weak learners to learn from their mistake sequentially with the aim that they can correct their high bias problem while maintaining the low-variance property.