CS5100 HOMEWORK 4
Sudharshan Subramaniam Janakiraman
07/27/2020
THEORY

1. .
   1. The Infant trying to learn and speak a language is very difficult. The infant must learn to listen to sounds , recognize the sound and produce the sound. The infant should know how to pronounce the word in a proper way. The key thing in infant learning to speak is its environment. The environment can be composed of other people to whom the infant can listen and pick up words as they speak or perceive a physical object to name it. Ex., the infant can easily learn and say a Ball when it looks at the ball visually rather than someone explaining what a ball looks like.

   The infant learns to speak by listening or viewing or feeling (externally or internally), making ears , eyes, brain and body parts that can feel acting as sensors to the infant's ability to to learn how to speak and mouth acts as the output. The infant also tries to speak through lip reading.

   The performance measure of how the infant speaks depends upon the environment it is present. If the environment does not give necessary input then infant might fare poorly but if the environment is rich enough to provoke the infant to speak then the performance measure will be high.

   2. A same problem can be represented in different ways leading to the possibility of representation of the problem in more than one decision tree and both are correct. This means both decisions trees satisfy the conditions of the problem but are completely different in their representation. This means a tree cannot always give the correct tree but give a tree which can be equivalent to the correct answer satisfying all the original conditions

3. .

$$[x_1, x_2] \xrightarrow{\text{Maps}} [x_1, x_1 * x_2]$$

x_1	x_2	x_1x_2
1	1	1
1	-1	-1
-1	1	-1
-1	-1	1

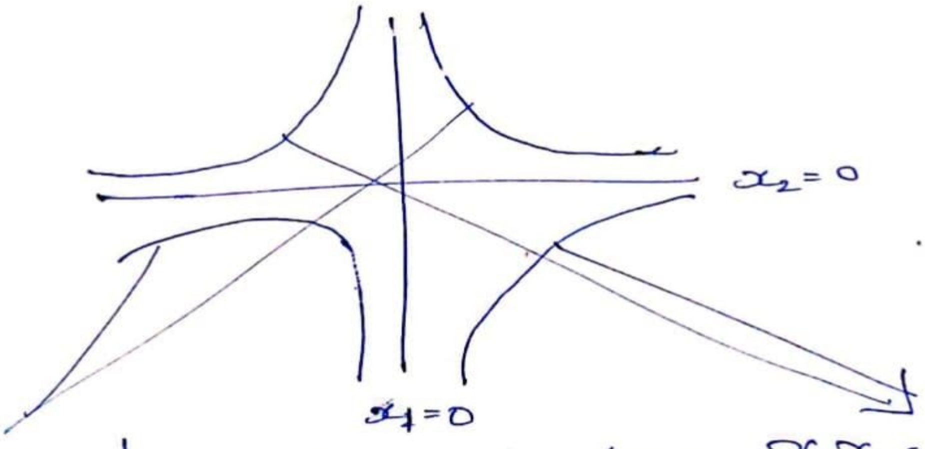
⇒

[x_1	x_2]		[x_1	x_1x_2]
1	1	→	1	1
1	-1	→	1	-1
-1	1	→	-1	-1
-1	-1	→	-1	1

⇒ The above representation Can be mapped by $x_1x_2 = 1$ & $x_1x_2 = -1$

$x_1x_2 = 0$ → Marginal Seperator

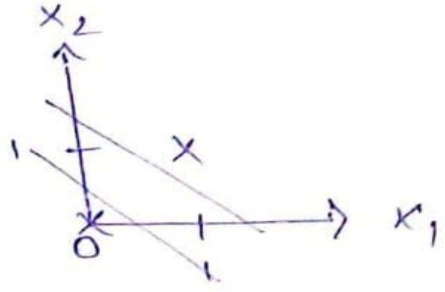
Margin = 1



$x_1x_2 = 1$ → HyperBola ← $x_1x_2 = -1$

4.

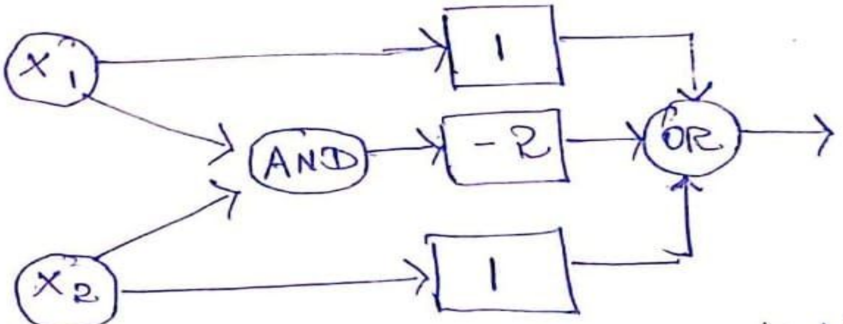
x_1	x_2	$x_1 \oplus x_2$
0	0	0
0	1	1
1	0	1
1	1	0



⇒ Consider x_1 AND x_2 , x_1 OR x_2

x_1	x_2	x_1 & x_2	x_1 (OR) x_2	$x_1 \oplus x_2$
0	0	0	0	0
0	1	0	1	1
1	0	0	1	1
1	1	1	1	0

Comparing $x_1 x_2$, $x_1 + x_2$ & $x_1 \oplus x_2$ we get XOR = OR - AND [when both x_1 & x_2 is 1 else OR]



Threshold = 1 □ → weights

(AND) → Hidden layer

(OR) → output layer

2. .
    1. Predict whether a new loan to be taken by a customer will be repaid within the stipulated time or not

        Linear Regression , Random Forest
    2. To find out which data plays a more important role in evaluating the business model , to find out whether loan will be repaid or not

        Random Forest
    3. To classify whether the customer is eligible for a loan or not

        Logistic regression, Naive Bayes, Support Vector Machines, Decision trees
    4. To forecast credit score based on the usage of credit cards and previous loans

        RNN, SVM, ANN

3. .
    1. .

        a. **Better** because the flexible method allows better curve fitting of the data leading to better use of data and fit it as much as closer to the curve compared to the inflexible method where the performance will be affected due to lack of flexibility.

        b. **Worse** because small observations leads to overfitting of data and the predictors will predict wrong values compared to that of inflexible method

        c. **Better** because flexible methods allows wide range of non linearity

        d. **Worse** because the variance is high , noise will be high and precision of data fitting will be less

    2. .

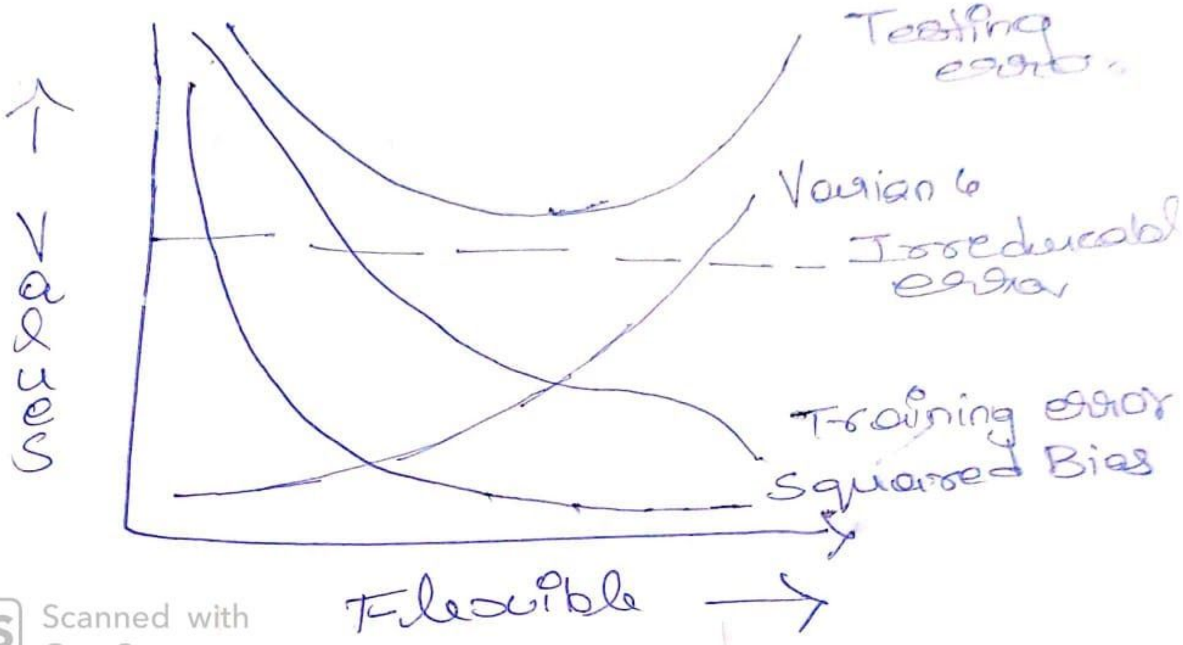
        a. **Regression** because the data needs to be fit along the curve to find out the factor that affects the CEO salary and **Infer** from it. N = 500, p = 3

        b. **Classification** and **Prediction** → Success or Failure n = 20 , p = 13

        c. **Regression** and **prediction** → predict future based on current data n = 52 , p = 3

3. .

a. .



b. Training Error : Decreases as flexibility increases as more data gets more freedom and fits around the curve

Testing Error : Decreases as the flexibility increase until the data is properly fitted and starts increasing when data becomes over fitted

Irreducible Error : Constant, parallel to flexibility

Squared Bias : Decreases as flexibility increases

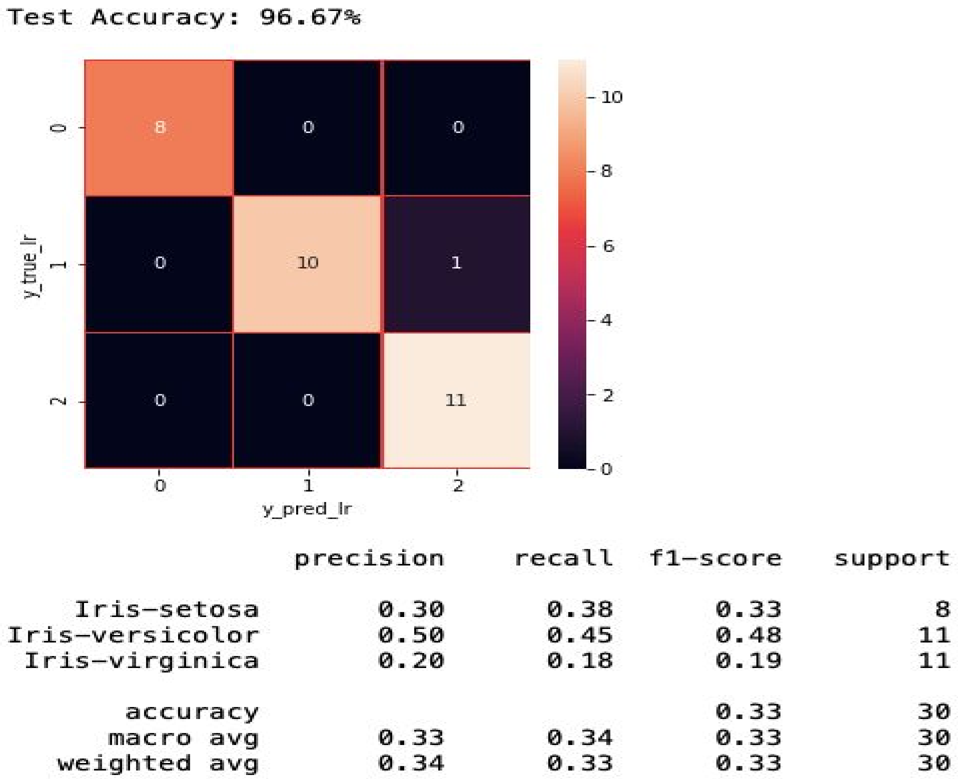
Variance : increases as more flexibility means more noise and less precision

PROGRAMMING ASSIGNMENT

SVM

Since the test train split splits the data randomly we get different outputs for both cases

IRIS DATASET



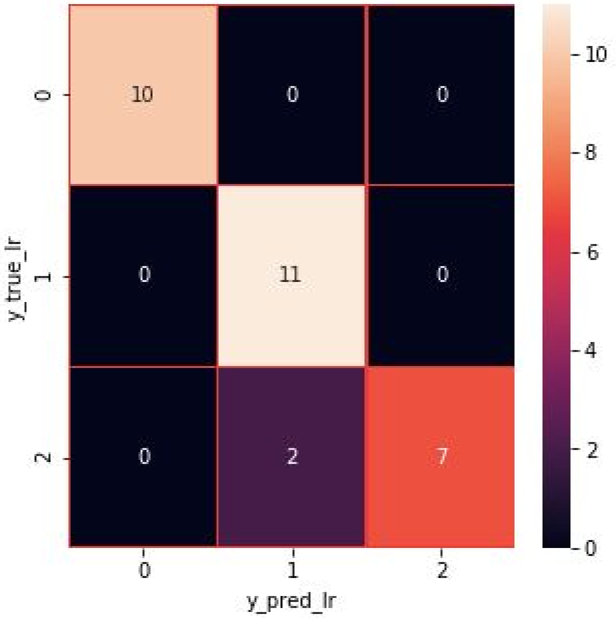
```
Test Accuracy: 96.67%
```

```
                 precision    recall  f1-score   support

    Iris-setosa       0.30      0.38      0.33         8
Iris-versicolor       0.50      0.45      0.48        11
 Iris-virginica       0.20      0.18      0.19        11

       accuracy                           0.33        30
      macro avg       0.33      0.34      0.33        30
   weighted avg       0.34      0.33      0.33        30
```

```
Test Accuracy: 93.33%
```

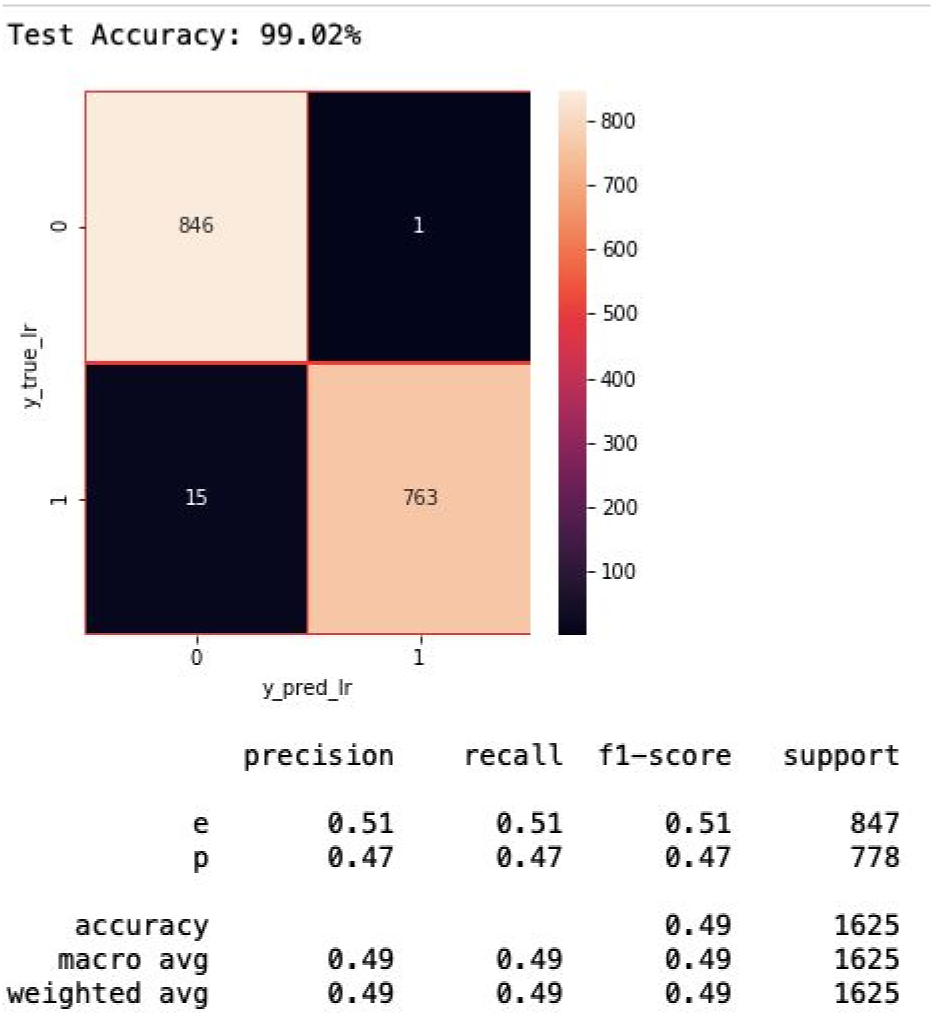


```
                 precision    recall  f1-score   support

    Iris-setosa       0.00      0.00      0.00        10
Iris-versicolor       0.20      0.18      0.19        11
 Iris-virginica       0.40      0.44      0.42         9

       accuracy                           0.20        30
      macro avg       0.20      0.21      0.20        30
   weighted avg       0.19      0.20      0.20        30
```

MUSHROOM DATASET



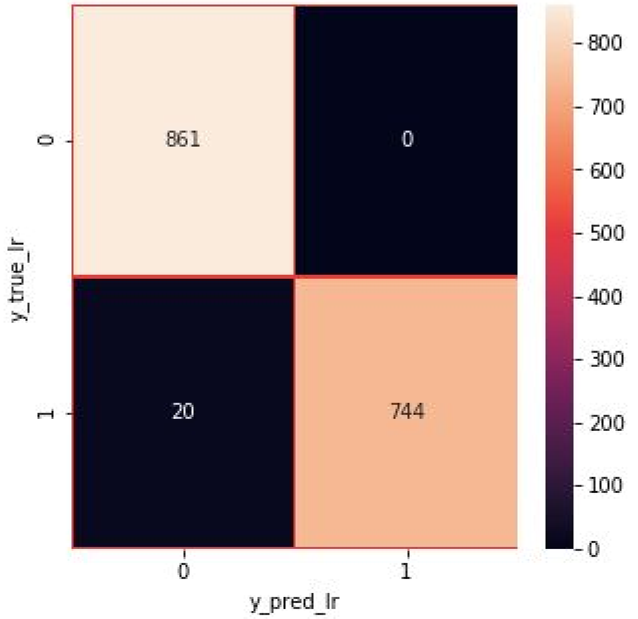
```
Test Accuracy: 99.02%
```

```
              precision    recall  f1-score   support

           e       0.51      0.51      0.51       847
           p       0.47      0.47      0.47       778

    accuracy                           0.49      1625
   macro avg       0.49      0.49      0.49      1625
weighted avg       0.49      0.49      0.49      1625
```

```
Test Accuracy: 98.77%
```

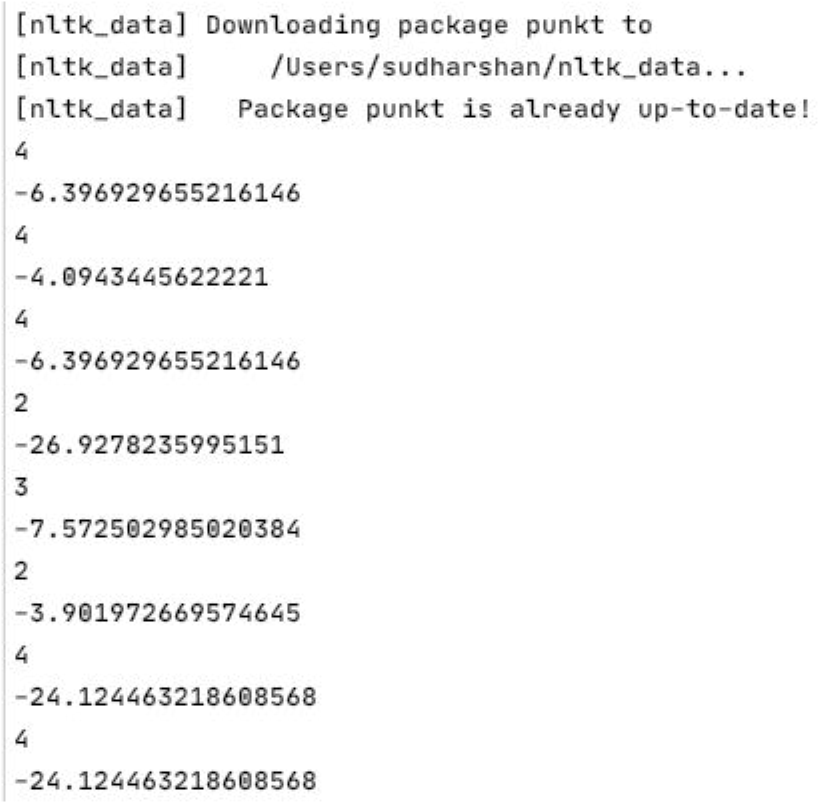


```
              precision    recall  f1-score   support

           e       0.55      0.53      0.54       861
           p       0.49      0.50      0.49       764

    accuracy                           0.52      1625
   macro avg       0.52      0.52      0.52      1625
weighted avg       0.52      0.52      0.52      1625
```

NAIVE BAYES AND MARKOV CHAINS



```
[nltk_data] Downloading package punkt to
[nltk_data]     /Users/sudharshan/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
4
-6.396929655216146
4
-4.0943445622221
4
-6.396929655216146
2
-26.9278235995151
3
-7.572502985020384
2
-3.901972669574645
4
-24.124463218608568
4
-24.124463218608568
```