**CS697A – Topic in Computer Science – Machine Learning**

Assignment 3

**Submitted by: Prachi Panse**

**NetID: ve3568**

**Q1:** Decision Trees, classification: Use **sklearn.tree** library’s DecisionTreeClassifier algorithm. For the DecisionTreeClassifier determine the value of the **tree depth** parameter (experiment with depth=2, 3, 5, 10) that results in the best test error for each of the 4-training data sets you created. How does the best depth value change as the number of instances and classes change?

***Answer:***

Program file: <Assignment3Q1.ipynb>

* Output For the given training and random testing data:

X100\_C69

tree\_depth :2 testing\_error : 0.02000 accuracy\_score : 0.98000

tree\_depth :3 testing\_error : 0.02000 accuracy\_score : 0.98000

tree\_depth :5 testing\_error : 0.02000 accuracy\_score : 0.98000

tree\_depth :10 testing\_error : 0.02000 accuracy\_score : 0.98000

X100\_CAll

tree\_depth :2 testing\_error : 0.68500 accuracy\_score : 0.31500

tree\_depth :3 testing\_error : 0.53400 accuracy\_score : 0.46600

tree\_depth :5 testing\_error : 0.35100 accuracy\_score : 0.64900

tree\_depth :10 testing\_error : 0.24100 accuracy\_score : 0.75900

X500\_69

tree\_depth :2 testing\_error : 0.00600 accuracy\_score : 0.99400

tree\_depth :3 testing\_error : 0.00600 accuracy\_score : 0.99400

tree\_depth :5 testing\_error : 0.00600 accuracy\_score : 0.99400

tree\_depth :10 testing\_error : 0.00600 accuracy\_score : 0.99400

X500\_CAll

tree\_depth :2 testing\_error : 0.70080 accuracy\_score : 0.29920

tree\_depth :3 testing\_error : 0.61620 accuracy\_score : 0.38380

tree\_depth :5 testing\_error : 0.31100 accuracy\_score : 0.68900

tree\_depth :10 testing\_error : 0.15100 accuracy\_score : 0.84900

* Analysis:

As we increase the tree depth, training error always reduces or stays constant, but it never gets increased with more depth of the tree. Whereas, for testing errors, initially with it deceases with increase in depth till a certain point and then increases due to overfitting. Considering depth value=2/3/5/10, minimum error is found at depth 10 for all datasets. With variance in the number of classes in datasets, the best tree depth value remained the same in this case of random testing instances. Below is the reference graph to analyze data behavior,

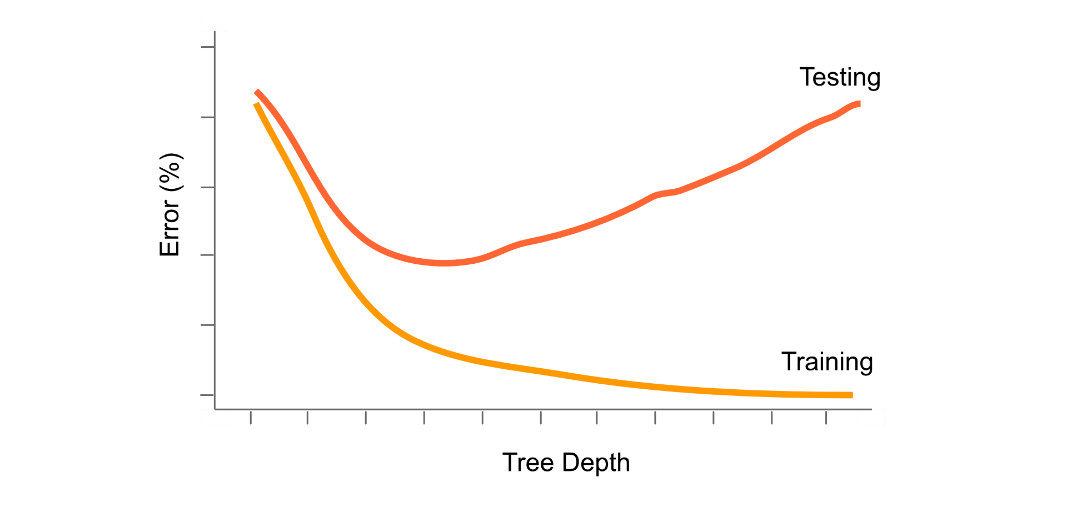
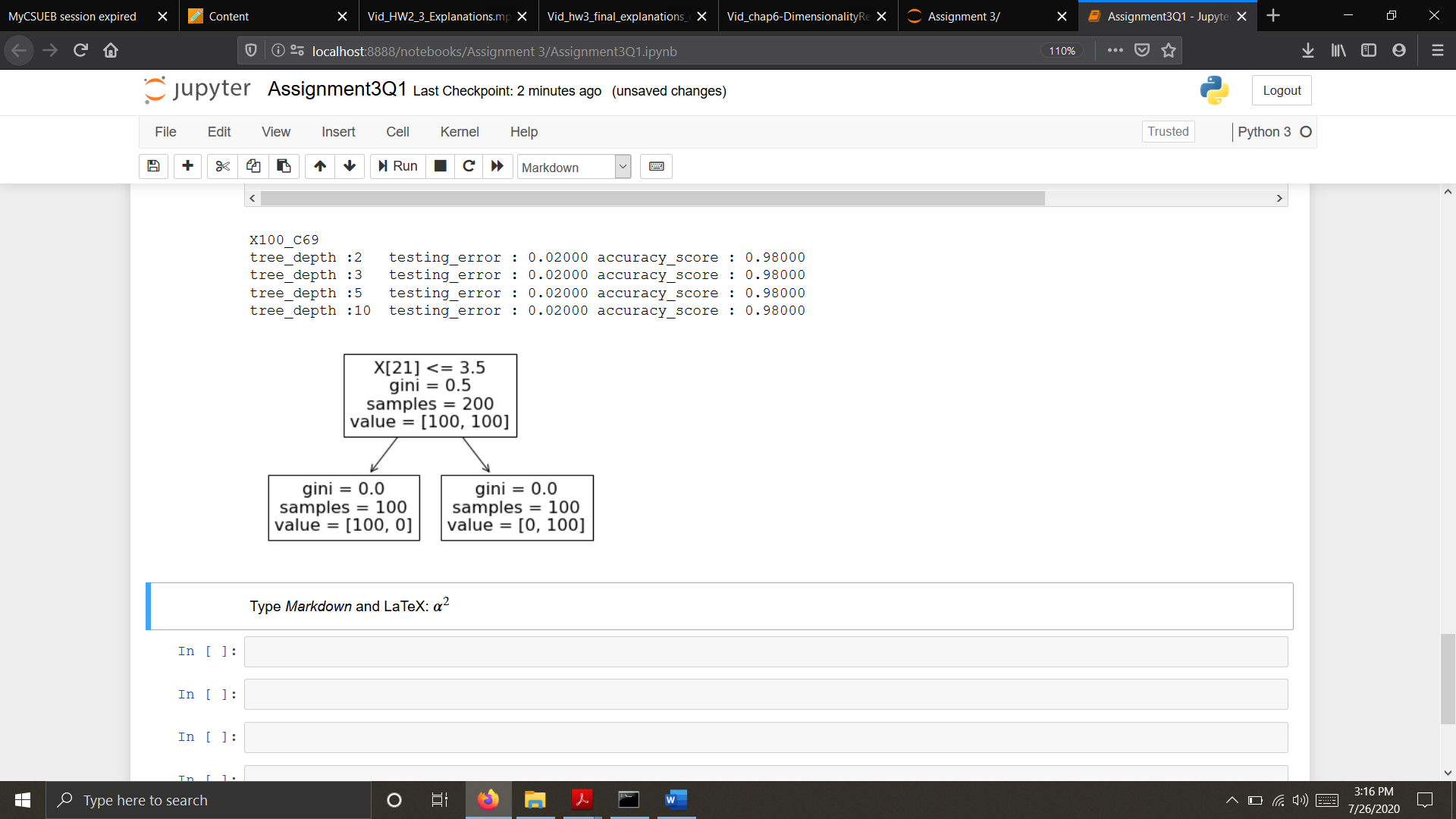


Image reference: <https://allmodelsarewrong.github.io/tree-basics.html>

* Tree-plotting for X100\_C69:



Likewise, it can be done for all the datasets.

**Q2:** Nonparametric Classification: Use **sklearn.tree** library’s KneighborsClassifier algorithm. For the KneighborsClassifier determine the value of the best k parameter (experiment with k=1, 3, 5, 9) that results in the best test error for each of the 4 training data sets you created. How does the best k value change as the number of instances and classes change?

***Answer:***

Program file: <Assignment3Q2.ipynb>

* Output For the given training and random testing data:

X100\_C69

k\_value :1 testing\_error : 0.00000 accuracy\_score : 1.00000

k\_value :3 testing\_error : 0.00000 accuracy\_score : 1.00000

k\_value :5 testing\_error : 0.00000 accuracy\_score : 1.00000

k\_value :9 testing\_error : 0.00000 accuracy\_score : 1.00000

X100\_CAll

k\_value :1 testing\_error : 0.05100 accuracy\_score : 0.94900

k\_value :3 testing\_error : 0.04300 accuracy\_score : 0.95700

k\_value :5 testing\_error : 0.04600 accuracy\_score : 0.95400

k\_value :9 testing\_error : 0.04300 accuracy\_score : 0.95700

X500\_69

k\_value :1 testing\_error : 0.00000 accuracy\_score : 1.00000

k\_value :3 testing\_error : 0.00000 accuracy\_score : 1.00000

k\_value :5 testing\_error : 0.00000 accuracy\_score : 1.00000

k\_value :9 testing\_error : 0.00000 accuracy\_score : 1.00000

X500\_CAll

k\_value :1 testing\_error : 0.02000 accuracy\_score : 0.98000

k\_value :3 testing\_error : 0.02420 accuracy\_score : 0.97580

k\_value :5 testing\_error : 0.02360 accuracy\_score : 0.97640

k\_value :9 testing\_error : 0.02320 accuracy\_score : 0.97680

* Analysis:

For datasets with only 2 classes, class 6 and class, accuracy rate was always 1, that means it has 0 errors. When it comes to all classes, the best k value changed as the number of instances did. For X100\_CAll, minimum testing error was at k value = 3 and 9, but for X500\_Call, best k value was 9 because it gave the minimum error.

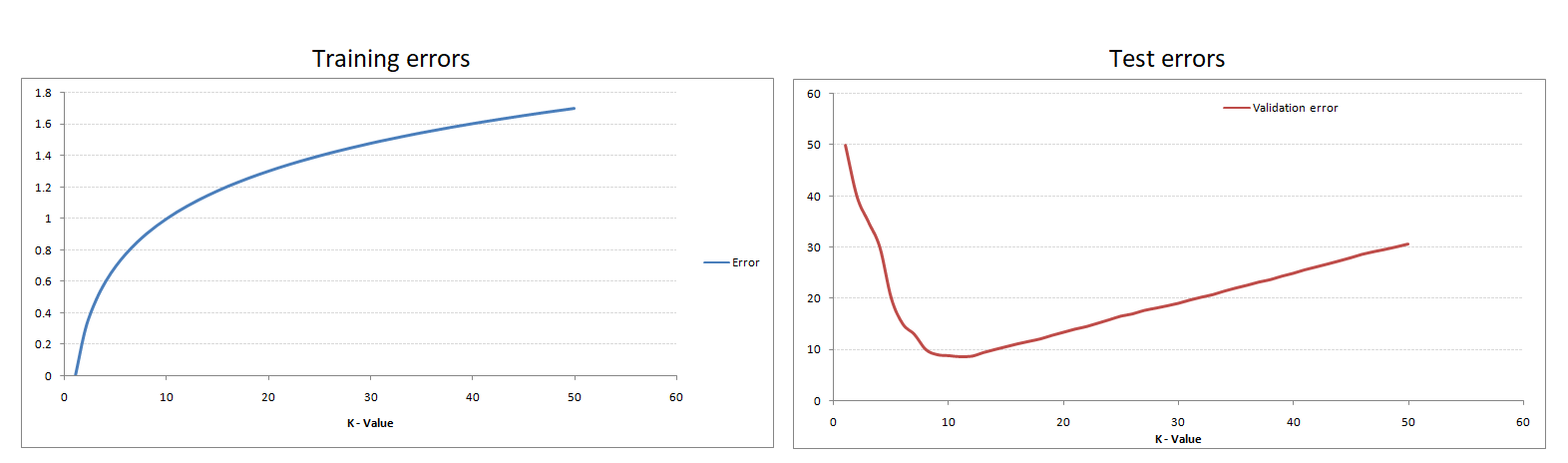


Image reference: [https://miro.medium.com/max/3144/0\*GXpGf2DXU4ogtpiI.png](https://miro.medium.com/max/3144/0*GXpGf2DXU4ogtpiI.png)

**Q3:** Decision Trees, regression for digit completion: Using only the data in X500\_69 for training, use the first 48 features as inputs and predict the next 16 features, i.e. create 16 decision tree regression models, using the sklearn library. Report the test error (use only the instances from classes 6 and 9) for each of the 16 regression models. Which pixels are easier to predict?

(Clarification, each of your models will have the same set of features, namely features 1…48.)

***Answer:***

Program file: <Assignment3Q3.ipynb>

* Output For the given training and testing data:

X500\_69

mserror for feature 49 : 0.0

mserror for feature 50 : 2.5290858725761773

mserror for feature 51 : 21.545706371191137

mserror for feature 52 : 14.903047091412743

mserror for feature 53 : 20.398891966759003

mserror for feature 54 : 19.822714681440445

mserror for feature 55 : 13.595567867036012

mserror for feature 56 : 1.074792243767313

mserror for feature 57 : 0.0

mserror for feature 58 : 0.09141274238227147

mserror for feature 59 : 3.7174515235457064

mserror for feature 60 : 7.080332409972299

mserror for feature 61 : 8.60387811634349

mserror for feature 62 : 20.803324099722992

mserror for feature 63 : 22.670360110803323

mserror for feature 64 : 0.7257617728531855

* Analysis:

The features which have minimum error will be considered as the best features and they will be easiest to predict. In the given dataset, feature 49 and feature 57 have the minimum error as 0. Therefore, predicting those pixels are easier.