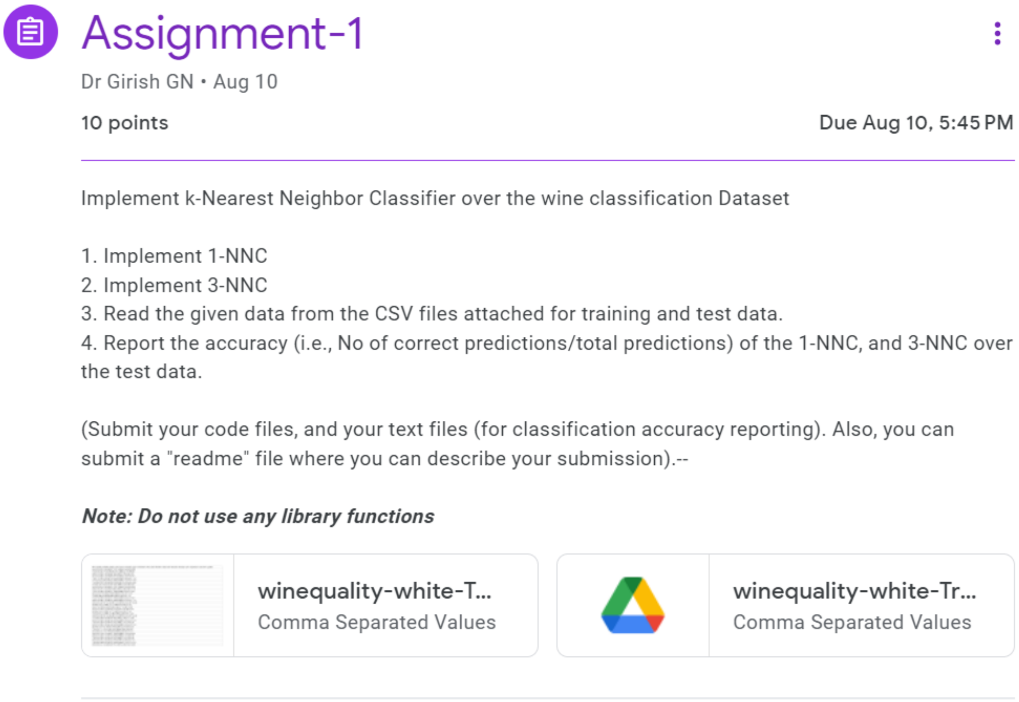
Assignment 1 :-

Implement k-Nearest Neighbor Classifier over the wine classification Dataset

1. Implement 1-NNC
2. Implement 3-NNC
3. Read the given data from the CSV files attached for training and test data.
4. Report the accuracy (i.e., No of correct predictions/total predictions) of the 1-NNC, and 3-NNC over the test data.

(Submit your code files, and your text files (for classification accuracy reporting). Also, you can submit a "readme" file where you can describe your submission).--



Assignment-1

Dr Girish GN • Aug 10

10 points

Due Aug 10, 5:45 PM

Implement k-Nearest Neighbor Classifier over the wine classification Dataset

1. Implement 1-NNC
2. Implement 3-NNC
3. Read the given data from the CSV files attached for training and test data.
4. Report the accuracy (i.e., No of correct predictions/total predictions) of the 1-NNC, and 3-NNC over the test data.

(Submit your code files, and your text files (for classification accuracy reporting). Also, you can submit a "readme" file where you can describe your submission).--

Note: Do not use any library functions

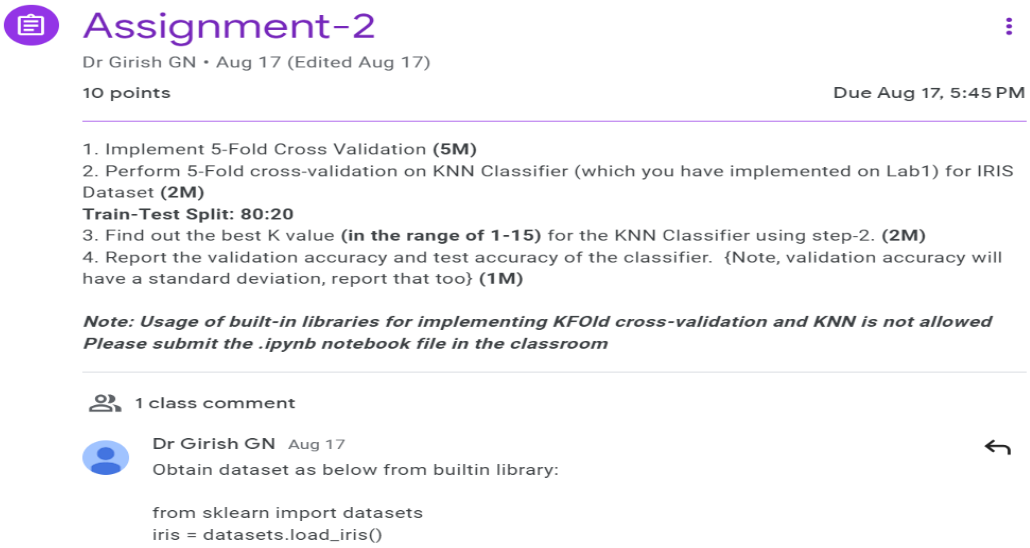
winequality-white-T...
Comma Separated Values

winequality-white-Tr...
Comma Separated Values

Assignment 2 :-

1. Implement 5-Fold Cross Validation **(5M)**
2. Perform 5-Fold cross-validation on KNN Classifier (which you have implemented on Lab1) for IRIS Dataset **(2M)**
Train-Test Split: 80:20
3. Find out the best K value **(in the range of 1-15)** for the KNN Classifier using step-2. **(2M)**
4. Report the validation accuracy and test accuracy of the classifier. {Note, validation accuracy will have a standard deviation, report that too} **(1M)**

Note: Usage of built-in libraries for implementing KFOld cross-validation and KNN is not allowed
Please submit the .ipynb notebook file in the classroom



Assignment-2

Dr Girish GN • Aug 17 (Edited Aug 17)

10 points

Due Aug 17, 5:45 PM

1. Implement 5-Fold Cross Validation **(5M)**
2. Perform 5-Fold cross-validation on KNN Classifier (which you have implemented on Lab1) for IRIS Dataset **(2M)**
Train-Test Split: 80:20
3. Find out the best K value **(in the range of 1-15)** for the KNN Classifier using step-2. **(2M)**
4. Report the validation accuracy and test accuracy of the classifier. {Note, validation accuracy will have a standard deviation, report that too} **(1M)**

Note: Usage of built-in libraries for implementing KFOld cross-validation and KNN is not allowed
Please submit the .ipynb notebook file in the classroom

1 class comment

Dr Girish GN Aug 17

Obtain dataset as below from builtin library:

from sklearn import datasets
iris = datasets.load_iris()

Assignment 3 :-

1. Use the OCR dataset provided in the following link:
https://sites.google.com/site/viswanathpulabaigari/data-sets.
2. Find the Best K for KNN using 3-fold cross-validation [5 Marks]
3. Find the **Minkowski distance metric (p) for the test using the Best K-NN classifier.** [5 Marks]
Note: Do not use any APIs for KNN, Cross Validation
Datasets : pp_tra.dat is the training set; pp_tes.dat is the test set
It is 192-dimensional data points; in each row, the 193rd entry is the class label.
There are 6670 training examples and 3333 test examples.

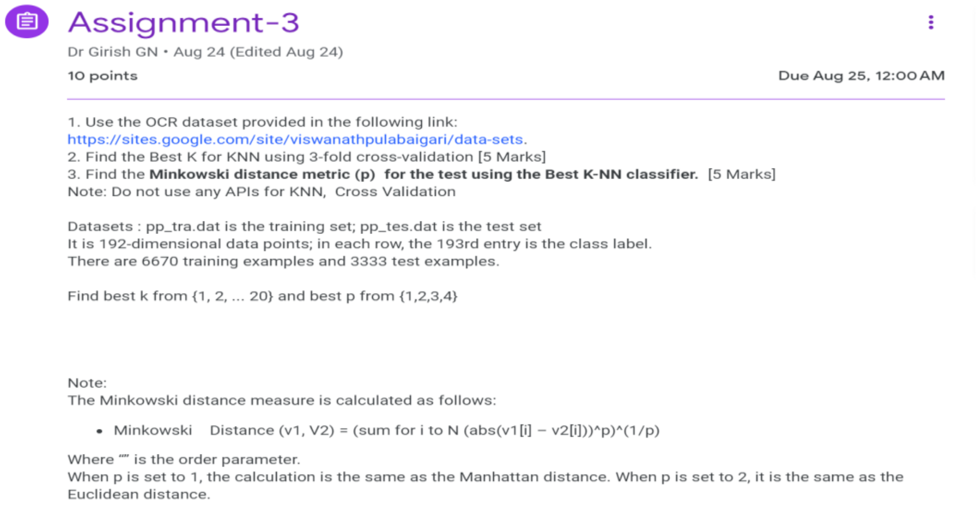
Find best k from {1, 2, ... 20} and best p from {1,2,3,4}
Note:

The Minkowski distance measure is calculated as follows:

Minkowski Distance (v1, V2) = (sum for i to N (abs(v1[i] – v2[i]))^p)^(1/p)

Where “” is the order parameter.

When p is set to 1, the calculation is the same as the Manhattan distance. When p is set to 2, it is the same as the Euclidean distance.



Assignment-3

Dr Girish GN • Aug 24 (Edited Aug 24)

10 points

Due Aug 25, 12:00 AM

1. Use the OCR dataset provided in the following link:
https://sites.google.com/site/viswanathpulabaigari/data-sets.
2. Find the Best K for KNN using 3-fold cross-validation [5 Marks]
3. Find the **Minkowski distance metric (p) for the test using the Best K-NN classifier.** [5 Marks]
Note: Do not use any APIs for KNN, Cross Validation

Datasets : pp_tra.dat is the training set; pp_tes.dat is the test set
It is 192-dimensional data points; in each row, the 193rd entry is the class label.
There are 6670 training examples and 3333 test examples.

Find best k from {1, 2, ... 20} and best p from {1,2,3,4}

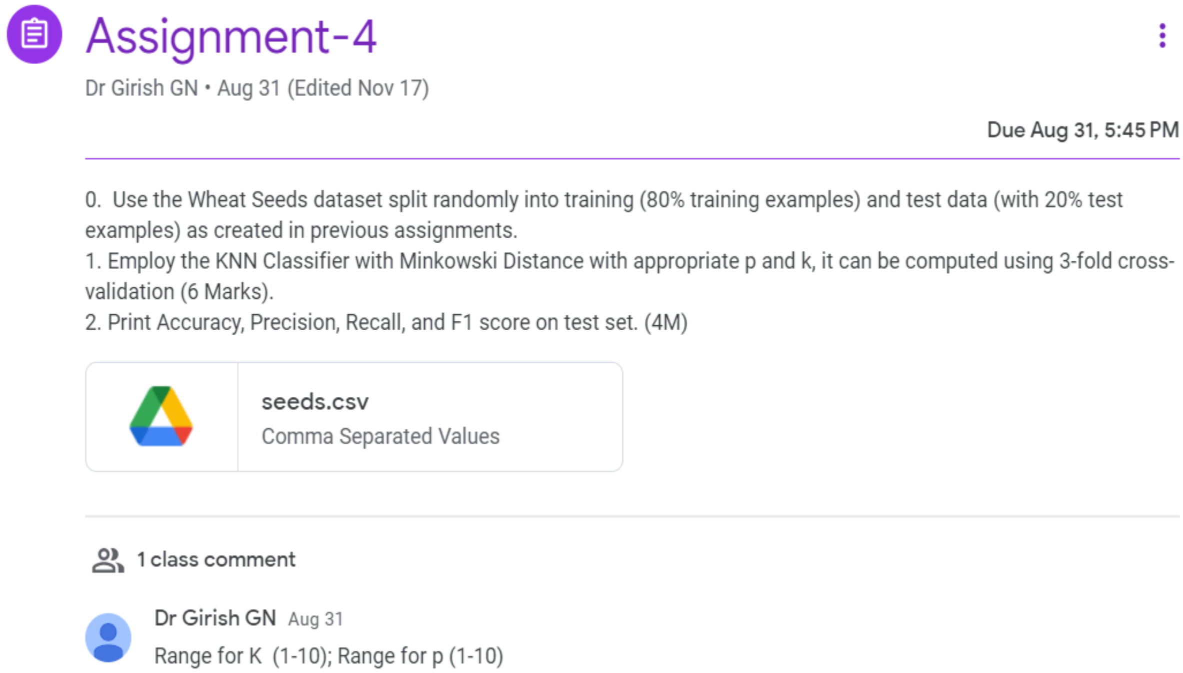
Note:
The Minkowski distance measure is calculated as follows:

- Minkowski Distance (v1, V2) = (sum for i to N (abs(v1[i] – v2[i]))^p)^(1/p)

Where “” is the order parameter.
When p is set to 1, the calculation is the same as the Manhattan distance. When p is set to 2, it is the same as the Euclidean distance.

Assignment 4 :-

0. Use the Wheat Seeds dataset split randomly into training (80% training examples) and test data (with 20% test examples) as created in previous assignments.
1. Employ the KNN Classifier with Minkowski Distance with appropriate p and k, it can be computed using 3-fold cross-validation (6 Marks).
2. Print Accuracy, Precision, Recall, and F1 score on test set. (4M)



Assignment-4

Dr Girish GN • Aug 31 (Edited Nov 17)

Due Aug 31, 5:45 PM

0. Use the Wheat Seeds dataset split randomly into training (80% training examples) and test data (with 20% test examples) as created in previous assignments.
1. Employ the KNN Classifier with Minkowski Distance with appropriate p and k, it can be computed using 3-fold cross-validation (6 Marks).
2. Print Accuracy, Precision, Recall, and F1 score on test set. (4M)

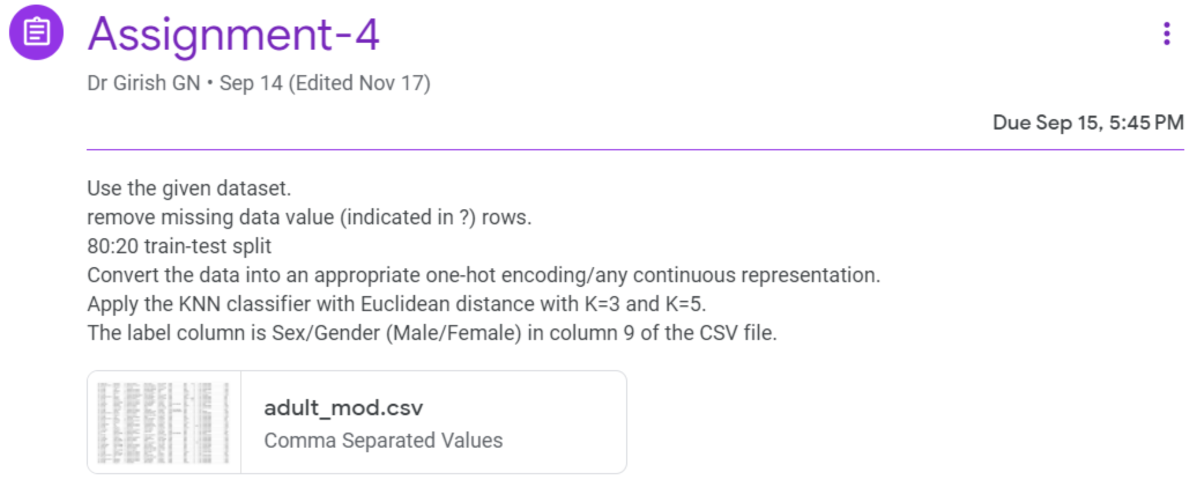
seeds.csv
Comma Separated Values

1 class comment

Dr Girish GN Aug 31
Range for K (1-10); Range for p (1-10)

Assignment 4 :- (ungraded)

Use the given dataset.
remove missing data value (indicated in ?) rows.
80:20 train-test split
Convert the data into an appropriate one-hot encoding/any continuous representation.
Apply the KNN classifier with Euclidean distance with K=3 and K=5.
The label column is Sex/Gender (Male/Female) in column 9 of the CSV file.



Assignment-4

Dr Girish GN • Sep 14 (Edited Nov 17)

Due Sep 15, 5:45 PM

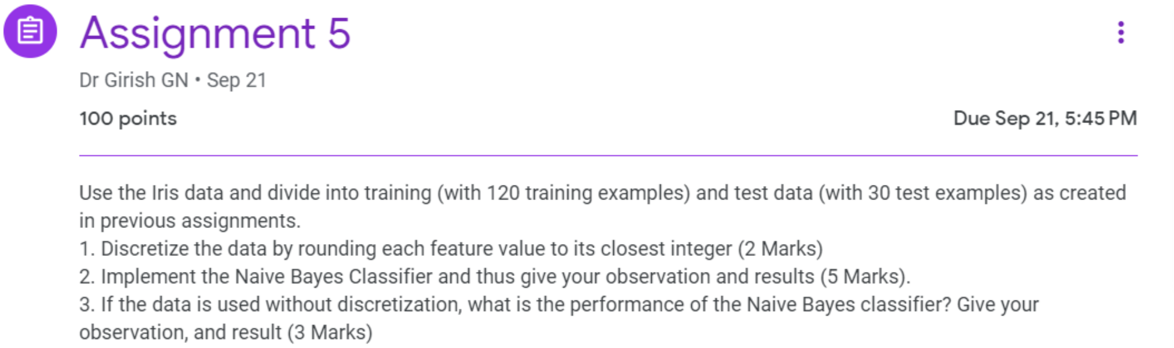
Use the given dataset.
remove missing data value (indicated in ?) rows.
80:20 train-test split
Convert the data into an appropriate one-hot encoding/any continuous representation.
Apply the KNN classifier with Euclidean distance with K=3 and K=5.
The label column is Sex/Gender (Male/Female) in column 9 of the CSV file.

adult_mod.csv
Comma Separated Values

Assignment 5 :-

Use the Iris data and divide into training (with 120 training examples) and test data (with 30 test examples) as created in previous assignments.

1. Discretize the data by rounding each feature value to its closest integer (2 Marks)
2. Implement the Naive Bayes Classifier and thus give your observation and results (5 Marks).
3. If the data is used without discretization, what is the performance of the Naive Bayes classifier? Give your observation, and result (3 Marks)



Assignment 5

Dr Girish GN • Sep 21

100 points

Due Sep 21, 5:45 PM

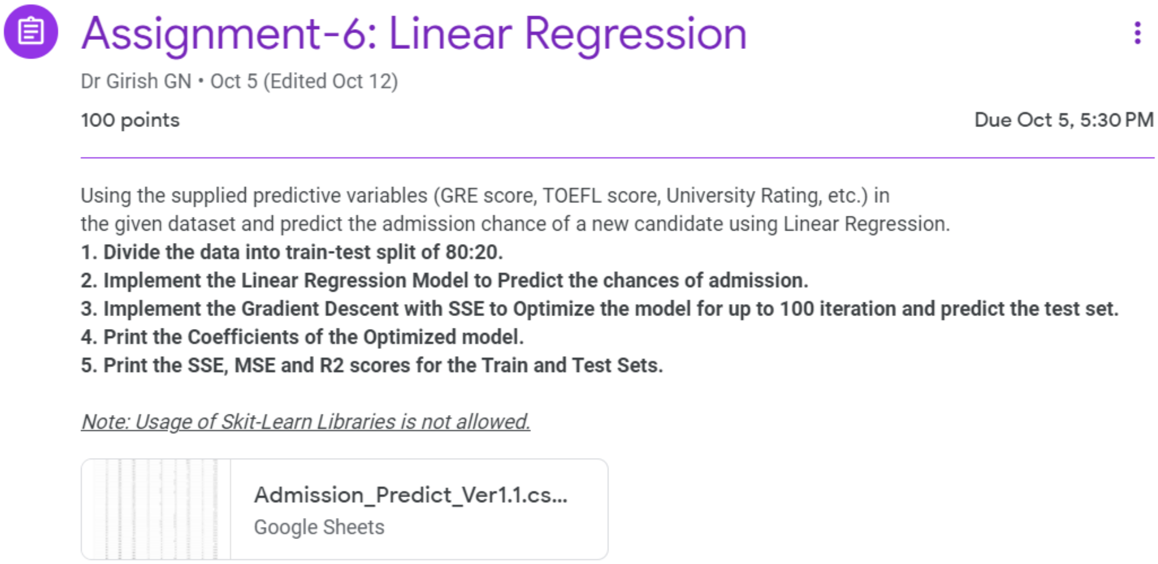
Use the Iris data and divide into training (with 120 training examples) and test data (with 30 test examples) as created in previous assignments.
1. Discretize the data by rounding each feature value to its closest integer (2 Marks)
2. Implement the Naive Bayes Classifier and thus give your observation and results (5 Marks).
3. If the data is used without discretization, what is the performance of the Naive Bayes classifier? Give your observation, and result (3 Marks)

Assignment 6 :-

Using the supplied predictive variables (GRE score, TOEFL score, University Rating, etc.) in the given dataset and predict the admission chance of a new candidate using Linear Regression.

1. Divide the data into train-test split of 80:20.
2. Implement the Linear Regression Model to Predict the chances of admission.
3. Implement the Gradient Descent with SSE to Optimize the model for up to 100 iteration and predict the test set.
4. Print the Coefficients of the Optimized model.
5. Print the SSE, MSE and R2 scores for the Train and Test Sets.

Note: Usage of Skit-Learn Libraries is not allowed.



Assignment-6: Linear Regression

Dr Girish GN • Oct 5 (Edited Oct 12)

100 points

Due Oct 5, 5:30 PM

Using the supplied predictive variables (GRE score, TOEFL score, University Rating, etc.) in the given dataset and predict the admission chance of a new candidate using Linear Regression.
1. Divide the data into train-test split of 80:20.
2. Implement the Linear Regression Model to Predict the chances of admission.
3. Implement the Gradient Descent with SSE to Optimize the model for up to 100 iteration and predict the test set.
4. Print the Coefficients of the Optimized model.
5. Print the SSE, MSE and R2 scores for the Train and Test Sets.

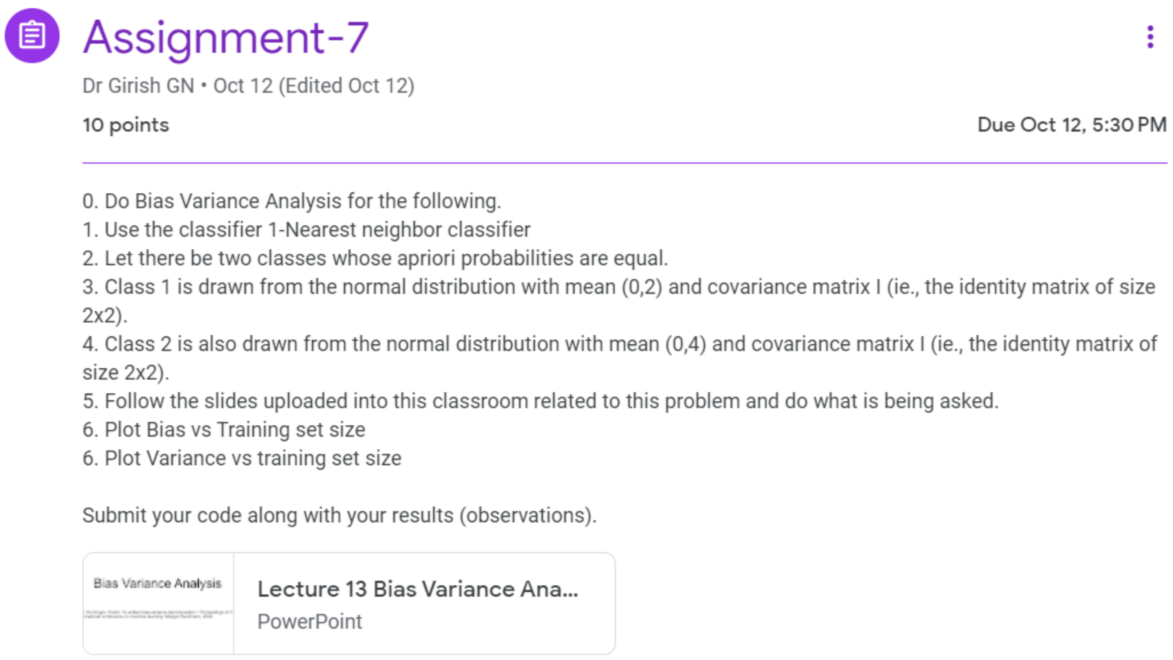
Note: Usage of Skit-Learn Libraries is not allowed.

Admission_Predict_Ver1.1.cs...
Google Sheets

Assignment 7 :-

0. Do Bias Variance Analysis for the following.
1. Use the classifier 1-Nearest neighbor classifier
2. Let there be two classes whose apriori probabilities are equal.
3. Class 1 is drawn from the normal distribution with mean (0,2) and covariance matrix I (ie., the identity matrix of size 2x2).
4. Class 2 is also drawn from the normal distribution with mean (0,4) and covariance matrix I (ie., the identity matrix of size 2x2).
5. Follow the slides uploaded into this classroom related to this problem and do what is being asked.
6. Plot Bias vs Training set size
6. Plot Variance vs training set size

Submit your code along with your results (observations).



Assignment-7

Dr Girish GN • Oct 12 (Edited Oct 12)

10 points

Due Oct 12, 5:30 PM

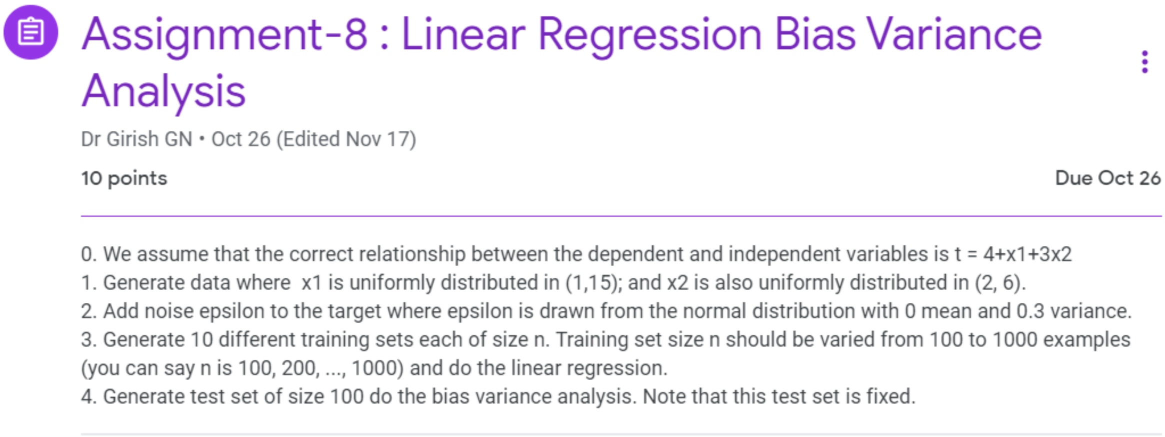
0. Do Bias Variance Analysis for the following.
1. Use the classifier 1-Nearest neighbor classifier
2. Let there be two classes whose apriori probabilities are equal.
3. Class 1 is drawn from the normal distribution with mean (0,2) and covariance matrix I (ie., the identity matrix of size 2x2).
4. Class 2 is also drawn from the normal distribution with mean (0,4) and covariance matrix I (ie., the identity matrix of size 2x2).
5. Follow the slides uploaded into this classroom related to this problem and do what is being asked.
6. Plot Bias vs Training set size
6. Plot Variance vs training set size

Submit your code along with your results (observations).

Assignment 8 :-

0. We assume that the correct relationship between the dependent and independent variables is t = 4+x1+3x2
1. Generate data where x1 is uniformly distributed in (1,15); and x2 is also uniformly distributed in (2, 6).
2. Add noise epsilon to the target where epsilon is drawn from the normal distribution with 0 mean and 0.3 variance.
3. Generate 10 different training sets each of size n. Training set size n should be varied from 100 to 1000 examples (you can say n is 100, 200, ..., 1000) and do the linear regression.

4. Generate test set of size 100 do the bias variance analysis. Note that this test set is fixed.



Assignment-8 : Linear Regression Bias Variance Analysis

Dr Girish GN • Oct 26 (Edited Nov 17)

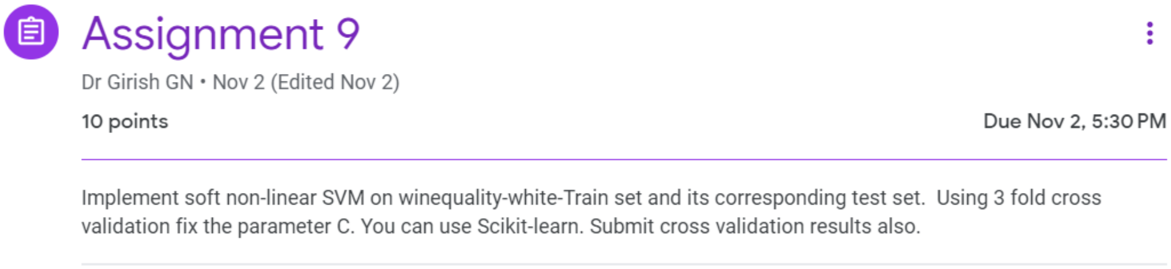
10 points

Due Oct 26

0. We assume that the correct relationship between the dependent and independent variables is t = 4+x1+3x2
1. Generate data where x1 is uniformly distributed in (1,15); and x2 is also uniformly distributed in (2, 6).
2. Add noise epsilon to the target where epsilon is drawn from the normal distribution with 0 mean and 0.3 variance.
3. Generate 10 different training sets each of size n. Training set size n should be varied from 100 to 1000 examples (you can say n is 100, 200, ..., 1000) and do the linear regression.
4. Generate test set of size 100 do the bias variance analysis. Note that this test set is fixed.

Assignment 9 :-

Implement soft non-linear SVM on winequality-white-Train set and its corresponding test set. Using 3 fold cross validation fix the parameter C. You can use Scikit-learn. Submit cross validation results also.



Assignment 9

Dr Girish GN • Nov 2 (Edited Nov 2)

10 points

Due Nov 2, 5:30 PM

Implement soft non-linear SVM on winequality-white-Train set and its corresponding test set. Using 3 fold cross validation fix the parameter C. You can use Scikit-learn. Submit cross validation results also.

Assignment 10 :-

Predict “tomorrow whether it will rain or not” using the target variable “RainTomorrow” in Australian weather dataset using Random Forest Classifier. [10 Marks]

a) If the target variable or feature value is NaN then drop the corresponding data points. Also, drop the Date and WindDirection 9 am features from the dataset [1 mark]

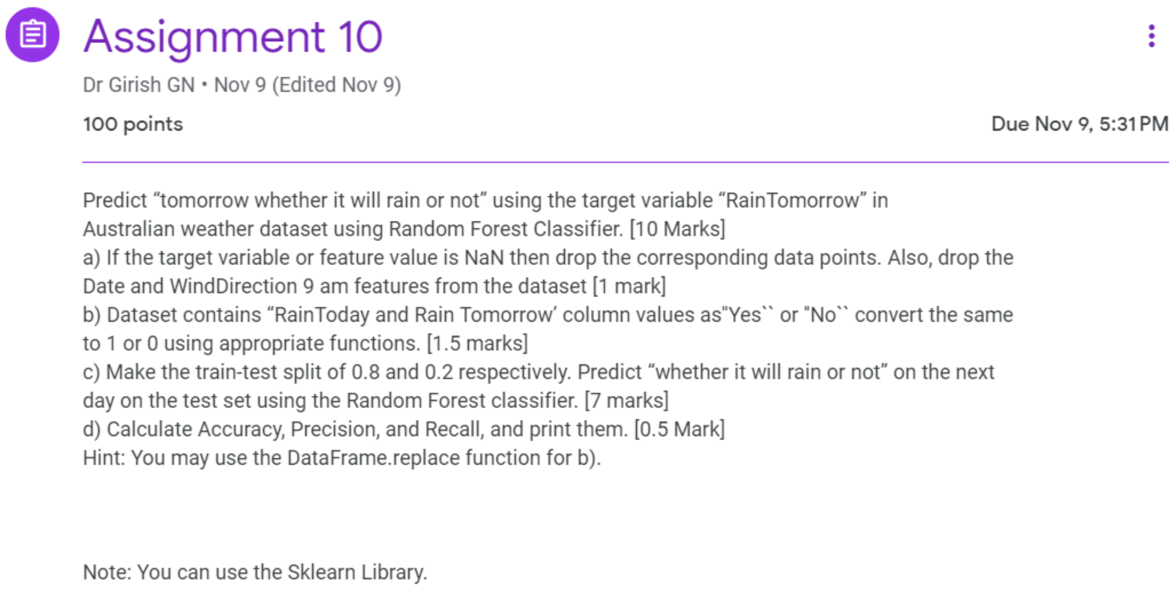
b) Dataset contains “RainToday and Rain Tomorrow’ column values as"Yes`` or "No`` convert the same to 1 or 0 using appropriate functions. [1.5 marks]

c) Make the train-test split of 0.8 and 0.2 respectively. Predict “whether it will rain or not” on the next day on the test set using the Random Forest classifier. [7 marks]

d) Calculate Accuracy, Precision, and Recall, and print them. [0.5 Mark]

Hint: You may use the DataFrame.replace function for b).

Note: You can use the Sklearn Library.



Assignment 10

Dr Girish GN • Nov 9 (Edited Nov 9)

100 points

Due Nov 9, 5:31 PM

Predict “tomorrow whether it will rain or not” using the target variable “RainTomorrow” in Australian weather dataset using Random Forest Classifier. [10 Marks]

a) If the target variable or feature value is NaN then drop the corresponding data points. Also, drop the Date and WindDirection 9 am features from the dataset [1 mark]

b) Dataset contains “RainToday and Rain Tomorrow’ column values as"Yes`` or "No`` convert the same to 1 or 0 using appropriate functions. [1.5 marks]

c) Make the train-test split of 0.8 and 0.2 respectively. Predict “whether it will rain or not” on the next day on the test set using the Random Forest classifier. [7 marks]

d) Calculate Accuracy, Precision, and Recall, and print them. [0.5 Mark]

Hint: You may use the DataFrame.replace function for b).

Note: You can use the Sklearn Library.

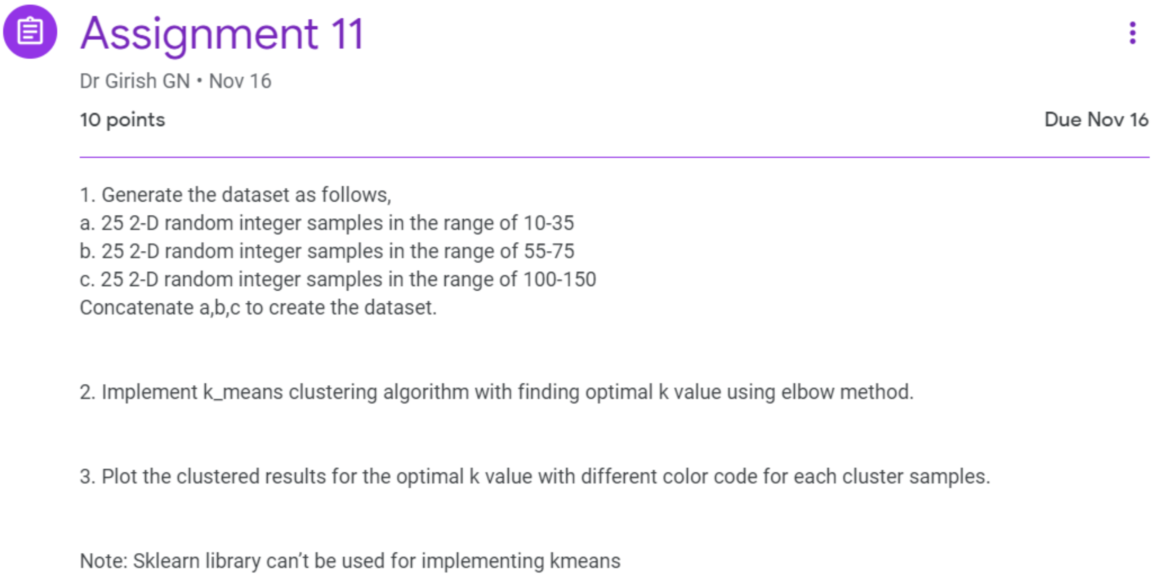
Assignment 11 :-

1. Generate the dataset as follows,
a. 25 2-D random integer samples in the range of 10-35
b. 25 2-D random integer samples in the range of 55-75
c. 25 2-D random integer samples in the range of 100-150
Concatenate a,b,c to create the dataset.

2. Implement k_means clustering algorithm with finding optimal k value using elbow method.

3. Plot the clustered results for the optimal k value with different color code for each cluster samples.

Note: Sklearn library can’t be used for implementing kmeans



Assignment 11

Dr Girish GN • Nov 16

10 points

Due Nov 16

1. Generate the dataset as follows,
a. 25 2-D random integer samples in the range of 10-35
b. 25 2-D random integer samples in the range of 55-75
c. 25 2-D random integer samples in the range of 100-150
Concatenate a,b,c to create the dataset.

2. Implement k_means clustering algorithm with finding optimal k value using elbow method.

3. Plot the clustered results for the optimal k value with different color code for each cluster samples.

Note: Sklearn library can’t be used for implementing kmeans