INST 737

MILESTONE 2.

DUE MARCH 31st at MIDNIGHT.

Deliverables: You will need to upload the following documents to the shared Google drive:

(a) Report (40 pts)

(b) Code (20 pts)

(c) Slides for a 10-minute presentation of the main Milestone 2 findings and a recorded video of your presentation published on Youtube. (10 pts). Remember to upload link to video in this file: http://bit.ly/2vrZapd

1. Report

Explain additional data collection, data preparation and data cleaning efforts that you have done after Milestone 1. These additional efforts might include, for example, the collection of more tweets for a specific time range or the provision of labels for a given dataset of images.

Next, provide answers for the following four questions. If you believe that some of the requested methods are not applicable to your research question, please justify why.

Question 1. Linear Regressions.

Divide your dataset into training and testing sets as we have seen in class and report:

a [5pts]. Linear Regression Parameters (show plots where applicable)

For each independent variable in your model, compute a linear regression with respect to the dependent feature and report:

- What is the intercept?

- What is the coefficient?

- Is it a predictive feature?

- Which are the most predictive features according to the training data?

- What are the residuals? Is a linear regression applicable to your problem?

- Use the trained model to predict on your testing data. Show results together with confidence and prediction bands. Report prediction accuracy using (1) the correlation between the predicted and real values and (2) the mean square error between the two.

b [4pts] . Multivariate regressions

Show whether considering combinations of independent features improves the prediction results. Evaluate different combinations of features as applicable and report those that improve the results shown in question (a).

- Which are the coefficients for each feature?

- Which are the most predictive features according to the training data?

Divide your dataset into training and testing sets and report prediction accuracy using (1) the correlation between the predicted and real values and (2) the mean square error between the two.

c [3pts] . Regularization

Repeat experiments in (a) and (b) adding regularization. Do you observe any improvements in the prediction results?

d [3pts]. Repeat a-c multiple times with different randomly selected training and testing sets and report differences or similarities across runs.

Question 2. Logistic Regression and NB.

a [5pts]. With the knowledge gathered from question 1(b), compute a logistic regression model with respect to different sets of independent features on your training dataset and report:

- What is the intercept?

- What are the coefficients for each of the features? Are they statistically significant?

- What are the log-odds and odd ratios of the outcome for a unit increase in each independent variable?

- Which are the most predictive features according to the training data?

- Use the trained model to predict on your testing dataset. Explain your results.

b [4pts]. Next, we are going to report classification results using Naïve Bayes. Remember to categorize all your variables before running the classifier.

- Divide your dataset into training and testing set and train the classifier. Report the confusion matrix.

- Repeat the process above with the Laplace estimator. Do the results improve?

Question 3. Decision Trees and Random Forests.

In this question we will evaluate the efficacy of decision trees and random forests. If your problem is multi-class (not binary), you will need to use all the classes to test the accuracy. If your independent variable is continuous, divide it into ranges between min and max values and consider each range a class.

a [2pts]. Split your dataset into training and testing sets assuming that samples are not randomly ordered (Hint: generate random numbers). Show that the distribution after the split is similar to the original.

b [3pts] . Train a decision tree and interpret the main resulting if-then rules (together with their corresponding plots). Test the trained tree with your testing dataset and compare the confusion matrices obtained during training and testing: compute percentages of correctly and incorrectly classified samples and compare results in training and testing.

c [3pts]. Apply boosting with different numbers of trees and analyze the impact on the prediction results: what is the impact of the number of trees in the accuracy of the classifier?

d [3pts]. Do the same analysis with bagging and random forests: train with bagging and then train with random forests using at least four different numbers of trees. Compare prediction results over the testing sets. Which are the most important features in each random forest?

Question 4. Comparative Analysis.

[5 pts] Write a summary of all classifiers, their predictive quality and which one would you use to answer your research question(s).

2. Code (20 pts)

The code needs to have comments that explain what each routine does. Please indent and comment your code and be well organized.

3. Presentation (10pts)

You are expected to record a video with your project presentation and to upload it to Youtube. The duration of the presentation should be at most **10 minutes**. Please

make sure you share the youtube link with all the class by adding it to the ‘projects’ file in the shared Google drive.

Please watch all the project presentations from the other teams since Milestone 3 will require you to provide a critical summary of each project.