CS400 Machine Learning

Final Exam (Take-home)

Available – Dec 5th 12:01 AM

Due – Dec 14th 12:00 Noon

Instructions

1. Files to submit – Section1.py, Section2.py, Section3.py, Section4.py, Discussion.pdf.
2. Comment the code files appropriately.
3. Zip all files into a zip file named LastNameFirstNameFinalExam.zip.
4. Submit the zip file in Blackboard.

Section 1

Dataset: Commodity1_price.csv, Commodity2_price.csv, Commodity3_price.csv
Dataset description: It is a list of prices of a perishable, limited consumption good or commodity reported in markets of a country.

a. Date: It's the date commodity was reported in the respective market.
b. Market: Market in which commodity was reported.
c. State: State in which the corresponding market is situated.
d. Variety: Variety of commodity reported.
e. Grade: Grade of commodity reported.
f. Tonnage (Arrival): Tonnage of a crop that arrives at the market
g. Prices: MinimumPrice, ModalPrice, and MaximumPrice columns are the corresponding prices of commodity for the date-state-market-variety-grade combination.

Problem description: We have prices available reported for commodity in different state and markets of the country. Our objective is to forecast the minimum and maximum price of a commodity for a given state, market, variety, and grade. The dataset includes three csv files for three different commodities. Find which model best fits the forecasting of the price for each commodity. Be mindful of overfitting and underfitting.

Submission: Submit the following for this section – one python script showing the final model that fits each commodity. Also, in the document include the highest score for each model that you tested. For each model score include appropriate information, example for KNN – number of neighbors, linear regression – name of final column(s) used for analysis. No mse value needed.

Section 2

Dataset: insurance.csv
Dataset description: It is a list of beneficiaries for an insurance company.

a. age: age of primary beneficiary
b. sex: insurance contractor gender, female, male
c. bmi: Body mass index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight (kg / m ^ 2) using the ratio of height to weight, ideally 18.5 to 24.9
d. children: Number of children covered by health insurance / Number of dependents
e. smoker: Smoking
f. region: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
g. charges: Individual medical costs billed by health insurance

Problem description: We have charges available reported for each insured person. Our objective is to forecast the charges billed to an individual. The task is to find which columns and which model can best predict the charges for an individual. Be mindful of overfitting and underfitting.

Submission: Submit the following for this section – one python script showing the final model along with the final set of columns that fits the prediction. Include the scores for each model that you tested in the document. For each model score include appropriate information, example for KNN – number of neighbors, linear regression – name of final column(s) used for analysis. No mse value needed.

Section 3

Dataset: pda_data_no_dups.csv

Dataset Description: This is a protein data set retrieved from Research Collaboratory for Structural Bioinformatics (RCSB) Protein Data Bank (PDB). The PDB archive is a repository of atomic coordinates and other information describing proteins and other important biological macromolecules. Structural biologists use methods such as X-ray crystallography, NMR spectroscopy, and cryo-electron microscopy to determine the location of each atom relative to each other in the molecule. They then deposit this information, which is then annotated and publicly released into the archive by the wwPDB. The constantly growing PDB reflects the research that is happening in laboratories across the world. This can make it both exciting and challenging to use the database in research and education.

Problem description: Which of these columns best predicts the "classification" of a protein: macromoleculeType, residueCount, resolution, structureMolecularWeight, densityMatthews, densityPercentSol, phValue. Be mindful of overfitting and underfitting.

Submission: Submit the following for this section – one python script showing the final model along with the final set of columns that fits the prediction. Include the scores for each model that you tested in the document. For each model score include appropriate information, example for KNN – number of neighbors, linear regression – name of final column(s) used for analysis. No mse value needed.

Section 4

Dataset: mushrooms.csv

Dataset Description: This dataset includes descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota Family Mushroom drawn from The Audubon Society Field Guide to North American Mushrooms (1981). Each species is identified as edible (e) or poisonous (p).

Problem description: Which of the columns best predicts the "classification" of a mushroom? Be mindful of overfitting and underfitting.

Submission: Submit the following for this section – one python script showing the final model along with the final set of columns that fits the prediction. Include the scores for each model that you tested in the document. For each model score include appropriate information, example for KNN – number of neighbors, linear regression – name of final column(s) used for analysis. No mse value needed.