API 222 Problem Set 1

Machine Learning and Big Data Analytics: Spring 2024

Due at 11:59am on February 16 - submit on Gradescope

This problem set is worth 30 points in total. To get full credit, submit your code along with a write-up of your answers. This should either be done in R Markdown or Jupyter Notebook, submitted in one knitted PDF.

Brief survey (0 pts)

Please fill out this brief (ungraded) survey to help Professor Saghafian and the teaching assistants get to know you. The link to the survey can also be found on Canvas under "Pages".

Conceptual Questions (15 pts)

1. For each of the following questions, state: (6 pts)

`(1) Whether it is a regression question or a classification question`

`(2) Whether we are interested in inference or prediction`

(a) A health organization is seeking to improve mental health services in a rural area. They want to identify individuals at high risk of developing stress-related disorders. They have demographic data and survey responses about lifestyle and stress levels. For the same group, they also have records indicating whether or not individuals sought mental health services.

(b) In a study exploring gender bias in job recruitment, researchers analyze, using application records and interview feedback, whether female applicants in technology roles are less likely to be called for an interview compared to male applicants.

(c) A team of researchers is investigating the impact of dietary changes on physical fitness levels among middle-aged adults. They first implement a program promoting a balanced diet and then measure the change in the participants' body mass index (BMI) over six months.

2. Flexible models versus inflexible models (5 pts)

(a) Flexible models will generally have lower bias than inflexible models. True or False?

(b) For the same very large number of observations, inflexible models will likely perform better than flexible models when the number of features is small. True or False?

(c) If the underlying data generating process is linear, a flexible model will generally perform worse than an inflexible one. True or False?

(d) Non-parametric models impose stronger assumptions on the underlying data generating process than parametric models. True or False?

(e) KNN and linear regression are both parametric models, as they both have decision rules. True or False?

3. The bias-variance tradeoff (4 pts)

(a) What does bias refer to in the machine learning context?

(b) What does variance refer to in the machine learning context?

(c) Now briefly describe the bias-variance tradeoff.

(d) Briefly explain the issue of overfitting in light of the bias-variance trade-off.

Data Questions (15 pts)

Introduction to NICU Length of Stay Prediction

This dataset presents a simulated scenario of neonatal intensive care unit (NICU) admissions, capturing a variety of clinical factors at the time of admission. Predicting the length of stay (LOS) in the NICU is a critical task that can significantly impact hospital resource planning, cost management, and patient care strategies. For hospital administrators and policymakers, accurately forecasting NICU LOS facilitates better allocation of resources, improves the quality of care, and aids in the development of efficient operational policies. You can read more about the variables in this dataset here. If you are curious about how I created this dataset, you can find the code here.

Also, for any non-integer numbers, please report your numbers to exactly two decimal places for full credit.

1. Preliminary data exploration (5 pts)

(a) How many observations and variables are in the dataset?

(b) Are any of the columns categorical? If so, which ones?

(c) How many rows have at least one missing value?

(d) Compute the mean and standard deviation of the NICU length of stay. *Hint: check the codebook!*

(e) Create a scatter plot of NICU length of stay (y-axis) on gestational age (x-axis). What does the plot tell you about the relationship between these two variables?

For the next few questions, first omit any rows (observations) with NA values and drop the District and Municipality variables. Then put the last 35 observations (after dropping rows with NA values) in a test set and the remaining observations in a training set.

2. When you use your training data to build a linear model that regresses overall 8th grade score on all other features available in the data (plus an intercept), what is your test Mean Squared Error? (2 pt)

3. Now use your training data to build a linear model that regresses overall 8th grade score on only three variables: total per-pupil expenditure, percentage qualifying for reduced-price lunch, and student-teacher ratio (include an intercept)

(a) What is your test MSE? (1 pt)

(b) Which coefficient changes the most between the regression in the previous question and the regression in this question? (1 pt)

(c) What was the coefficient's value under the regression from the previous question? (1pt)

(d) What is the coefficient's value under the current regression? (1 pt)

(e) Provide some intuition for what it means that this coefficient changed significantly between the two regressions. (1 pt)

4. When you use your training data to build a KNN model that regresses overall 8th grade score on all other features in the data, what is your test Mean Squared Error with $K = 2$? (2 pt)

5. When you use your training data to build a KNN model that regresses overall 8th grade score on all other features in the data, what is your test Mean Squared Error with $K = 10$? (1 pt)