API 222 Problem Set 1

Machine Learning and Big Data Analytics: Spring 2024

Due at 11:59am on February 15 - 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 "Quizzes".

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.

(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)

This dataset focuses on predicting Atherosclerotic Cardiovascular Disease (ASCVD) risk, encompassing clinical, demographic, and lifestyle data. Accurate ASCVD risk prediction is crucial for public health policy, enabling early intervention and informed healthcare strategies. It aids policymakers and health officials in reducing the burden of cardiovascular diseases, a leading global cause of death, and in formulating policies for healthier lifestyles. Utilizing this data in machine learning can lead to improved public health outcomes and more efficient healthcare resource allocation, demonstrating the dataset's significant implications for public health policy and patient care.

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) Compute the mean and standard deviation of the `Risk` score.

For the next few questions, set the seed to 222 and randomly put 20% of your observations 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 `Risk` 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 Risk on only three variables: Age, isDiabetic, and isHypertensive (include an intercept)

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

(b) Create a table that shows the coefficients from both of your models. Use the `stargazer` package to do this if you are working in an RMD. (2 pt)

(c) Provide some intuition for what it means that some of the coefficients changed between the two regressions. (1 pt)

4. When you use your training data to build a KNN model that regresses Risk on all other features in the data, what is your test Mean Squared Error with $K = 2$? (1 pt)

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

6. Between the standard linear regression and the KNN regression, which performed better? (1 pt)

7. From an inference standpoint, which of these models would we rather use? (1 pt)