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.

Classification. Prediction. The goal is to identify individuals at high risk of developing stress-related disorders, which is a classification problem. The organization is interested in predicting the risk of developing stress-related disorders, which is a prediction problem. The problem says nothing about understanding the relationship between the predictors and the response, so we are not interested in inference.

(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.

Classification. Inference. The goal is to analyze the effect of being female on the likelihood of being called for an interview, which is a classification problem.

(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.

Regression. Inference. The goal is to measure the change in the participants' BMI, which is a regression problem. The researchers are interested in understanding the impact of dietary changes on physical fitness levels, which is an inference problem.

2. Flexible models versus inflexible models (5 pts) EXPLAIN ANSWERS

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

True. Flexible models generally have a lower bias but higher variance compared to inflexible models.

(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?

False. With a large sample size and a small number of features, flexible models can pick up meaningful yet complex patterns in the data and are less prone to overfitting issues. Therefore, they will likely outperform inflexible models.

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

True. Since the underlying process is linear, a linear model (which sits under the umbrella of inflexible models) is appropriate. A flexible model is more likely than the inflexible model to incorporate the noise in the training data when making predictions.

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

False. Non-parametric model generally require no functional form assumption and thus impose weaker assumptions on the underlying data generating process than paramteric models.

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

False. KNN is a nonparametric model and linear regression is a parametric model.

3. The bias-variance tradeoff (4 pts)

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

Bias refers to the error produced by representing a real world problem by a statistical learning method.

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

Variance denotes the degree of change in the prediction function when estimated on a different training set.

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

In an ideal world, we would find a model that has low variance and low bias, because that would yield a good and consistent model. In practice, you usually have to allow bias to increase in order to decrease variance and vice versa.

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

Overfitting occurs when we use a flexible model fits exactly against the training data, resulting in low bias. However, the model performs poorly on out-of-sample data due to high variance.

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?

```
# Load the data
data <- read.csv("data/heartRisk.csv")

dim(data)
```

```
## [1] 1000    9
```

There are 1000 observations and 9 variables.

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

```
str(data)
```

```
## 'data.frame':    1000 obs. of  9 variables:
##  $ isMale        : int  1 0 0 1 0 0 1 1 0 1 ...
##  $ isSmoker      : int  0 0 1 1 1 1 1 1 1 0 ...
##  $ isDiabetic    : int  1 1 1 1 0 0 0 1 0 1 ...
##  $ isHypertensive: int  1 1 1 0 1 1 0 0 1 1 ...
##  $ Age           : int  49 69 50 42 66 52 40 75 42 65 ...
##  $ Systolic      : int  101 167 181 145 134 154 104 136 169 196 ...
##  $ Cholesterol   : int  181 155 147 166 199 174 187 189 179 187 ...
##  $ HDL           : int  32 59 59 46 63 22 52 59 99 46 ...
##  $ Risk          : num  11.1 30.1 37.6 13.2 15.1 17.3 2.1 46 1.7 48.5 ...
```

`isMale`, `isSmoker`, `isDiabetic`, and `isHypertensive` are categorical.

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

```
m <- mean(data$Risk, na.rm = TRUE)

sd <- sd(data$Risk, na.rm = TRUE)

print(paste("Mean:", round(m,2), "Standard Deviation:", round(sd,2)))
```

```
## [1] "Mean: 19.67 Standard Deviation: 17.04"
```

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.

```
set.seed(222)
test <- sample(1:1000, 200)
train <- setdiff(1:1000, test)

train_data <- data[train,]
test_data <- data[test,]
```

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)

```
library(stargazer)
model.1 <- lm(Risk ~ ., data = train_data)

test_pred <- predict(model.1, newdata = test_data)
test_mse <- mean((test_pred - test_data$Risk)^2)
test_mse
```

```
## [1] 81.00468
```

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)

```
model.2 <- lm(Risk ~ Age + isDiabetic + isHypertensive, data = train_data)

test_pred <- predict(model.2, newdata = test_data)
test_mse <- mean((test_pred - test_data$Risk)^2)
test_mse
```

```
## [1] 146.7809
```

(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)

```
stargazer(list(model.1,model.2), type = "latex",
          column.labels = c("Model 1", "Model 2"))
```

Table 1:

	Dependent variable:	
	Risk	
	Model 1	Model 2
	(1)	(2)
isMale	4.894***	
	(0.535)	
isSmoker	9.165***	
	(0.536)	
isDiabetic	10.569***	10.449***
	(0.536)	(0.826)
isHypertensive	4.717***	5.281***
	(0.535)	(0.825)
Age	0.920***	0.878***
	(0.023)	(0.036)
Systolic	0.209***	
	(0.009)	
Cholesterol	0.070***	
	(0.013)	
HDL	−0.094***	
	(0.011)	
Constant	−85.725***	−40.725***
	(2.974)	(2.196)
Observations	800	800
R^2	0.799	0.518
Adjusted R^2	0.797	0.517
Residual Std. Error	7.539 (df = 791)	11.635 (df = 796)
F Statistic	393.184*** (df = 8; 791)	285.600*** (df = 3; 796)
Note:		*p<0.1; **p<0.05; ***p<0.01

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

The coefficient for isHypertensive increased in the second model. This means that when we control for the other variables in the model, the effect of being hypertensive on ASCVD risk decreases. This is likely due to multicollinearity between the variables in the first model.

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)

```
library(FNN)
model.3 <- knn.reg(train = train_data[, -9],
                   test = test_data[, -9],
                   y = train_data$Risk, k = 2)

test_mse <- mean((model.3$pred - test_data$Risk)^2)
test_mse
```

```
## [1] 155.3028
```

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)

```
model.4 <- knn.reg(train = train_data[, -9],
                   test = test_data[, -9],
                   y = train_data$Risk, k = 10)

test_mse <- mean((model.4$pred - test_data$Risk)^2)
test_mse
```

```
## [1] 126.5375
```

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

The linear regression performed better than the KNN regression. The test MSE for the KNN regression with $K = 10$ was 126.5375, while the test MSE for the standard linear regression was 81.00468.

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

We would rather use the standard linear regression model. The KNN model is a black box model, and it is difficult to interpret the relationship between the features and the target variable. The standard linear regression model allows us to interpret the coefficients of the features and understand the relationship between the features and the target variable.