Business Analytics

Session 8b. Midterm Review

Renyu (Philip) Zhang

New York University Shanghai

Spring 2019

Final Exam

- Time: 6:00pm-9:00pm on Monday, May 20.
- Coverage: All the materials covered in this course.
- Policy: Open book, open note, calculator allowed; cellphone, laptop, iPad, and other electronic devices not allowed.
- Format: A few T/F questions; a few (small scale) quantitative analysis questions, and a few logical reasoning questions.
- Difficulty: Comparable with homework problem sets (the questions without involving data).
- Sample exam: To be distributed as we approach the end of this semester.

Linear Regression

- Please interpret the following linear regression model:

$$\textit{kid\_score} \approx 25.73 + 5.95\textit{mom\_hs} + 0.56\textit{mom\_iq}$$

  - *kid_score* =IQ score of a kid
  - *mom_hs* =Whether mom has attended high school (1 if yes, 0 if no)
  - *mom_iq* =IQ score of the mom

- Alternatively, interpret the next linear regression model:

$$\textit{kid\_score} \approx -11.48 + 51.27\textit{mom\_hs} + 0.97\textit{mom\_iq} - 0.48\textit{mom\_hs} \cdot \textit{mom\_iq}$$

Linear Regression

- We have the following training set:

X	1	2	3	4
y	0	2	1	3

The linear regression model trained on this data sample is

$$y \approx -0.5 + 0.8X$$

The testing set is:

X	0	5
y	-1	4

- What is the in-sample R^2?
- What is the out-of-sample R^2?

Logistic Regression

- Consider the following logistic regression model:

$$\mathbb{P}[Y = 1|p] \approx \frac{\exp(20 - 0.1p)}{1 + \exp(20 - 0.1p)}$$

  - Y =Whether the customer purchases the product (1 if yes, 0 if no)
  - p =price of the product
- How would you interpret the coefficient of price -0.1?
- Suppose that the model predicts $Y = 1$ if the predicted probability $\mathbb{P}[Y = 1] > 0.5$. Will the model predict that the customer will purchase the product if the price is 210?
- The confusion matrix of the classification model with the classification threshold $t = 0.5$ is given by the following:

	Predicted $Y = 1$	Predicted $Y = 0$
Actual $Y = 1$	20	10
Actual $Y = 0$	15	25

What is the false positive rate? If we want to decrease the false positive error rate to 0.25, how should we adjust the value of t? If we make this adjustment of t, what will happen to the false negative error rate?

Logistic Regression

- For the logistic regression model built in the slide 5, the out-of-sample AUC is 0.78. How can you interpret AUC in this specific context?

- Use your own words to explain why, if the model complexity increases, the model variance will increase whereas the model bias will decrease.

- For Lasso and Ridge linear regression, what will be the fitted model parameters $(\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p)$ if the regularization parameter α goes to infinity?

Trees

- Back in 2014, Leonardo Dicaprio was disappointed and would like to know what it takes to (finally) win an Academy Award. He complied a data set of people who have and have not won, with few features. He would use a decision tree to classify award winners from non-winners. See the following table for his data set.

Person	age≥ 50	has children	gender	is L. Dicaprio	class
Anthony Hopkins	Y	Y	male	N	winner
Honey Boo Boo	N	N	female	N	loser
Leonardo Dicaprio	N	N	male	Y	loser
Meryl Streep	Y	Y	female	N	winner
Morgan Freeman	Y	N	male	N	winner
Jennifer Lawrence	N	N	female	N	winner
Sandra Bullock	N	Y	female	N	winner

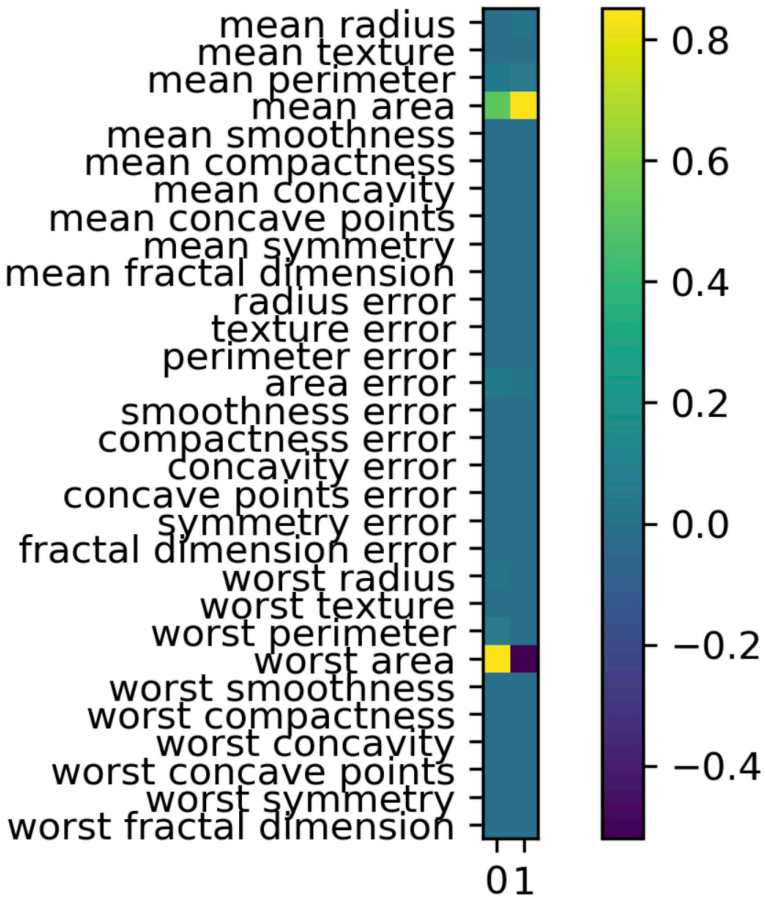
- Define the outcome variable Y and the features X for the data set.
- Draw a decision tree that has 0 error on the training set.
- Build a decision tree iteratively to minimize the Gini index in each step and with 0 error.

Clustering and PCA

- Consider the following data set:

Data Index	1	2	3	4	5	6	7	8	9	10
X	1	-2	3	-1	4	7	10	6	-5	-10

Use $k-$means algorithm to cluster the data set into 3 groups with initial centers $X = -10$, $X = 0$, and $X = 10$. What would be the resulting



- Recall the PCA method with the Breast cancer data set. The PCA result is reported in the left figure. Which features have the highest variabilities based on the PCA result?
- What will happen to the PCA result if we standardize all the features by subtracting the mean and dividing by the standard deviation? Why do we standardize the features before performing PCA?

Feature Engineering and Selection

- Give an example to illustrate why the robust scaler is not influenced by outliers.
- To impute missing values in a data set, what is the advantage of $k-$nearest neighbors approach over directly imputing the sample mean/median?
- Why overall accuracy is a misleading model performance measure for classification if the data is highly imbalanced?
- What is the gold standard for feature engineering and model selection?

Potential Outcome Models and Randomized Experiments

- If you directly compare the average health score of the patients who are hospitalized and those who are non-hospitalized, will you overestimate or underestimate the causal effect of hospital? Why?

- Suppose you want to examine the effectiveness of a new drug. How would you design an experiment to achieve this goal? Explain clearly why your analysis is valid.

- Explain why in a well-randomized experiment, you do not need to include the covariates in the regression to estimate the treatment effect.

Observational Studies and Instrumental Variables

- In observational studies, what is the key assumption for identifying the causal effect of a treatment? Give a example to explain what this assumption means.
- If you directly regress log-of-earning on education length (in schooling years), will you overestimate or underestimate the causal effect of education?
- How do you check the strong-first stage assumption of an IV?
- We randomly send out encouragement to adopt a new feature to a group of users for an App. Compared with the control group, the users who are sent the encouragement have about 20% higher chance to adopt this feature; and they, on average, spend 2.5 more minutes on the App per day. What is the estimated causal effect of this new feature on a user's App time per day?

Final Group Project

Goal of the Project

- To have the opportunity to go through entire procedure of data-driven decision making/analytics.
  - Problem definition
  - Model formulation.
  - Analysis with analytics techniques
  - Interpretation of results
  - Implementation/recommendation to the original problem

- To practice the analytics methodologies we have learned in this course.
  - Prediction
  - Causal Inference
  - Optimization

Goal of the Project

- To have the opportunity to go through entire procedure of data-driven decision making/analytics.
  - Problem definition
  - Model formulation.
  - Analysis with analytics techniques
  - Interpretation of results
  - Implementation/recommendation to the original problem

- To practice the analytics methodologies we have learned in this course.
  - Prediction
  - Causal Inference
  - Optimization

- To have fun by solving an interesting problem!

Choice of Final Project

1. Heterogeneous Treatment Effect with Decision Trees
2. Multi-Armed Bandit Experiments
3. Elevator Routing and Scheduling
4. Sports Analytics (aka Moneyball Theory)
5. A project of your own choice (need approval from me)

Examples of Final Project in Previous Classes

- Predicting 2016 US presidential election result
- Predicting NCAA Madness March result
- Predicting prices of AirBnb listings
- Optimizing class schedule
- Optimizing bike redistribution in bike-sharing systems

Timeline of Final Project

- April 14, Sunday, 10:00pm: Choice of final project due
- April 28, Sunday, 10:00pm: Final project progress report due
- Last Session (May 13 or 14): Final project presentation
  - Each member of a team should come to the stage to present
  - Time controlled
  - Each student should ask at least one question
- May 20, Monday, 10:00 PM: Final project due
  - Final project report (do not write too long!)
  - Data set(s)
  - Code
  - Slides
  - Mutual group member evaluation

Feel free to discuss with me or send your project report to me for feedback any time before the due date.

Grading of Final Project

- Final project counts for 40% of the final grade.

- Grading scheme of the final project:
  - Clearly defined problem and correctly formulated analytics model (10%)
  - Rigorous analysis and meaningful results (18%)
  - Results clearly interpreted
    - Insights to the original problem (4%)
    - Limitations of the results/method clearly discussed (4%)
  - Compelling project presentation (2%)
  - Well-written project report (2%)
  - Extra credit: Innovative ideas, thorough analysis, significant results, practical impact, etc. (up to 10%, case by case evaluation)

Group Member Mutual Evaluation

- No free-riders.
- For each member of your group, evaluate his/her contribution in (actively involved, involved, (almost) not involved):
  - Deciding the topic and getting the data set(s)
  - Formulating the model
  - Analysis
  - Writing the report
  - Preparing for and participating in the presentation/slides
- Give a percentage contribution for each member in your group
  - e.g., 30%(Yourself)+20%(Member 1)+25%(Member 2)+25%(Member 3)
- Do not talk to each other when filling the mutual-evaluation form.
- Individual grade for final project:
  - Grade of the group project + peer-evaluations from your group members

Homework

- Finish Homework 8 (NO need to submit it).
- A bonus assignment (5% extra credits).
  - Due at 10:00pm on May 20.
- Read the required reading for Session 9.
- Review the questions discussed today.
- Start think about your final project. Confirm with your group information with me.