Final exam

EC 524/424

Due *before* midnight on Thursday, 18 March 2021

Thank you for a great term!

The exam

To wrap up, we have some big-picture, short-answer questions and a fun classification task.

Important: You are to work on this exam alone. Do not discuss it with your classmates until finals week is over. If I suspect you of cheating, you will fail this course, which may cause you to fail out of the program.

You *can* use your notes, books, and the internet.

Coding: Do not ask Stephen or me to troubleshoot your code. I (Ed) will answer clarifying questions. We will not debug your code.

Exam description: I've divided your exam into two main sections.

In the first section ("Section 1"), you will provide **short** answers (1–3 sentences) to the prompts. Section 1 has four subsections. In each subsection you get to choose four of the five questions that you want to answer (meaning you get to skip one question per subsection).

In the second section ("Section 2"), you get to do classification!

To submit: You should submit two items:

1. Your answers to the questions/prompts in the two sections (PDF, HTML, *etc.*).
2. Your predictions from *Section 2*.

Due: Your materials are due *before* midnight (Pacific) on Thursday, 18 March 2021.

Section 1: The bigger picture

Instructions: In each subsection, answer *four* of the five questions. Make it clear which question you are answering.

Keep your answers short (1–3 sentences).

1A

1.01 Compare and contrast "the validation-set approach" and "k-fold cross validation".

1.02 Explain what bootstrapping is.

1.03 Why don't we typically use **accuracy** to split classification-based trees?

1.04 Define L1 and L2 loss. Describe how they will affect your measurement of loss.

1.05 Does ridge regression include variable selection? Explain.

1B

1.06 Explain the role bootstrapping plays in bagging trees. How does it solve some of the problems typical to decision trees?

1.07 Why would we want to use out-of-bag (OOB) error instead of CV?

1.08 Define AUC (area under the curve). Explain why an AUC of 1 desirable.

1.09 How is leave-one-out cross validation related to k-fold cross validation?

1.10 Why do we need to standardize/normalize our predictors for elasticnet, ridge, and lasso?

1C

1.11 **Define and compare/contrast** sensitivity and precision.

1.12 Describe (generally) when/why we should prefer accuracy, sensitivity, and specificity.

1.13 In one sentence: explain the idea behind SVM.

1.14 What is cost-complexity pruning and why do we use it?

1.15 What is ensemble learning?

1D

1.16 Why is cross validation so important in prediction settings (relative to causal-inference settings)?

1.17 In the variance-bias trade off: What do we mean by "variance"? What do we mean by "bias"? In general, how does increasing a model's flexibility affect "variance" and "bias"?

1.18 How do you define a "good" level of accuracy?

1.19 Is it better to overfit or underfit? Explain.

1.20 When cleaning data, why is it a bad idea to drop observations with missing values of variables?

Section 2: Data time

Download the data files. You should have three files:

- The training dataset (`train.csv`).
- The testing dataset (`test.csv`).
- An example of what your submission file should look like (`sample-submission.csv`).

The data come from Yelp restaurants in Portland, Oregon.

Your goal: Use restaurants' observable variables to predict whether they will have a rating of 4.5 `stars` or higher.

Important: You will have to use the numeric `stars` variable (the restaurants' ratings) to create your own outcome variable that is `"Yes"` if the restaurant has a rating of 4.5 stars or higher and `"No"` if the restaurant's average rating is below 4.5. *Note:* You will probably then want to drop `stars` (*i.e.*, don't use it for prediction).

Steps (include your answers to these steps in the write up)

`2.01` Load the data and create a clean, aesthetically pleasing, well-labeled figure that describes the data. Explain *why* your figure does a good job of describing the data.

Note: I understand that this figure will not describe every variable in the dataset. Choose one or two dimensions that you think are important.

`2.02` Choose a method (*e.g.*, logistic elasticnet, random forest, boosted trees, SVM) for this classification task (predicting restaurants with average ratings above 4.5). Why did you choose this method?

Note: You cannot choose plain, OLS regression or plain, logistic regression.

`2.03` Clean the data. This step includes normalization, creating new variables/interactions, *etc*. Again: Briefly explain your decisions.

`2.04` Tune your model's hyperparameters. Describe which hyperparameters you tuned and which values your model "chose".

`2.05` What is your best model's (estimated) out-of-sample accuracy?

`2.06` What is your best model's (estimated) out-of-sample precision?

`2.07` Train your best model on the full training dataset. Predict onto the test dataset. Submit your predictions.