Midterm Exam
STA9890
Statistical Learning for Data Mining

Assignment Parameters:
Date Assigned: 2024-03-24
Date Due: 2024-04-02 @ 5:45pm

This is a closed-note, closed-book exam.
You may not use any external resources.

Instructions

This exam will be graded out of **100 points**.

You will have 1.5 hours (90 minutes) to complete this exam.

You will have 100 minutes from the time you download this exam on Blackboard to upload your solutions: that is, you have 90 minutes to complete the exam and 10 minutes to scan and upload any written answers. *You are responsible for uploading your exam in time, so plan accordingly.*

Late exams will not be accepted.

This exam is divided into four equally weighted parts:

- Multiple Choice
- Short Answer
- Issue Spotter
- Mathematics of Machine Learning

Though all sections are worth the same total number of points, they are not all equally difficult. Individual questions within a section also vary in difficulty. You need to use your time wisely. Skip questions that are not easy to answer quickly and return to them later.

For the multiple choice and short answer questions, please answer in the space provided on the exam sheet. If you need additional space for the issue spotter or mathematical questions, use the additional pages at the end of the PDF.

Mark your answers clearly: if I cannot easily identify your intended answer, you may not get credit for it.

I *will not* answer questions during the exam period. If a question is ambiguous, do your best to answer it. If you need to make additional assumptions to answer a question, please state them in full.

This is a closed-note, closed-book exam. You may not use any external resources.

Multiple Choice (25 points; 2.5 points each)

For each question, **CIRCLE** your answer(s).

Q1. True/False: Supervised learning problems have a continuous response variable (y).

TRUE FALSE

Q2. True/False: OLS finds the linear model with the lowest training MSE.

TRUE FALSE

Q3. True/False: Low-bias models should always be preferred to maximize out-of-sample accuracy.

TRUE FALSE

Q4. True/False: Models with higher training error always have higher test error

TRUE FALSE

Q5. True/False: Selecting the model with the lowest cross-validation error will always minimize in-sample (training) error

TRUE FALSE

Q6. True/False: The lasso does a better job selecting 'true' variables in low correlation settings

TRUE FALSE

Q7. Select all that apply: Which of these describes kernel methods?

- They allow us to fit spline methods efficiently
- They allow us to do 'infinite-dimensional feature expansion'
- Because they are more flexible, they always provide out of sample (test accuracy) improvements

Q8. Select all that apply: The "ℓ_0 norm" is

- The number of non-zero elements in a vector
- An example of an ℓ_p-norm
- The sum of the absolute values of a vector

Q9. Select all that apply: The following are convex penalties

- Lasso
- SCAD
- Best Subsets

Q10. Select all that apply: The following are sparsity-inducing penalties

- Local Polynomial Penalty
- Ridge
- MC-PLUS

Short Answer (25 points; 5 points each)

Q1. Sketch the bias-variance trade-off curve for K-Nearest Neighbors. Place K on the horizontal (x) axis and draw two lines (one for bias and one for variance). Identify the point with the best test error. Label the lines clearly.

Q2. Give three reasons we may want to use a *sparse* linear model:

Q3. When might we prefer to use an ℓ_1-loss instead of an ℓ_2-loss for regression?

Q4. Under what conditions is OLS unbiased?

Q5. Rank the following models in terms of complexity: OLS; 1-Nearest Neighbor Regression; Cubic Spline Regression; Ridge Regression; Cubic Polynomial Regression.

Issue Spotter (25 points)

This is an *issue spotter* question. Below, I describe (in words) a hypothetical application of the ML techniques we have discussed in this class. Your task is to find **5 mistakes** in the ML pipeline and to:

i) describe the problem (2 points each); and

ii) say how that step could have been performed more effectively / accurately. (3 points)

Scenario:

> Baruch College is looking to better understand the post-graduation outcomes of its students. The Office of Institutional Research (OIR) sent a survey to 1,000 randomly selected students to have graduated during the past 20 years and received 750 responses. Students were contacted at their last reported work email. OIR collected the following information:
>
> - Years of Study at Baruch
> - Year of Graduation
> - Initial post-graduation salary
> - Baruch GPA
>
> OIR noted that more recent graduating classes had a higher response rate, but because the sample was selected uniformly from all graduates, it is representative.
>
> The CUNY Board has set a goal of ensuring over 80% of Baruch graduates have salaries putting them in the middle class, which they define as a household income of $50,000 or more annually (roughly 2/3 of the NYC median household income of $75,000). Using this information, OIR creates a response variable
>
> $$y_i = \begin{cases} 1 & \text{Initial post-graduation salary } > 50,000 \\ 0 & \text{Initial post-graduation salary } < 50,000 \end{cases}$$
>
> and builds the following OLS model to predict y:
>
> $$\hat{y} = -0.1 * \#\text{ Years of Study} + 0.3 * \text{Baruch GPA} + 0.25 * \text{Year of Graduation}$$
>
> The CUNY Board examines this model and recommends the following policy changes:
>
> - Terminating all Bachelor's (4 year) programs and focusing only on Associate's (2 year) programs to decrease the average number of years of study at Baruch.
> - Instructing the faculty to relax grading standards to raise average GPAs by 0.5.
>
> OIR notes that, if these proposals had been enacted, their model predicts 35% more middle class outcomes; the CUNY Board responds enthusiastically and implements these changes without a pilot program. OIR also notes that the fraction of middle-class salaries has increased rapidly over the past 4 years. In recognition of this achievement, the CUNY Board authorizes substantial performance-recognition compensation (bonuses) for Baruch leadership.

Mistake #1:

Fix #1:

Mistake #2:

Fix #2:

Mistake #3:

Fix #3:

Mistake #4:

Fix #4:

Mistake #5:

Fix #5:

Mathematics of Machine Learning (25 points)

Consider the *generalized ridge regression* estimator given

$$\hat{\boldsymbol{\beta}}_{\text{GRR}} = \text{argmin}_{\boldsymbol{\beta}} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{D}\boldsymbol{\beta}\|_2^2$$

where $\boldsymbol{D}$ is a known matrix.

Q1.) Derive a closed form expression for $\hat{\boldsymbol{\beta}}_{\text{GRR}}$ (5 points).

Q2.) Write out a gradient descent algorithm that can be used to solve for $\hat{\boldsymbol{\beta}}_{\text{GRR}}$ (10 points).

Q3.) Now, suppose $\boldsymbol{D}$ is chosen so that

$$\|\boldsymbol{D}\boldsymbol{\beta}\|_2^2 = \sum_{j=2}^{p}(\beta_j - \beta_{j-1})^2$$

What happens as λ gets larger? When might you want to use this type of penalty? (5 points)

Q4.) How does this method differ from the *fused lasso*, given by:

$$\hat{\boldsymbol{\beta}}_{\text{FL}} = \text{argmin}_{\boldsymbol{\beta}} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{j=2}^{p} |\beta_j - \beta_{j-1}|$$

? (5 points)

There are many valid answers to this problem, but the simplest might be to take $\boldsymbol{X}$ to be the identity matrix and to draw solutions for $\hat{\boldsymbol{\beta}}_{GRR}$ and $\hat{\boldsymbol{\beta}}_{FL}$.

(Extra page for longer answers)

(Extra page for longer answers)

(Blank page for scratch work - not graded)

(Blank page for scratch work - not graded)

(Blank page for scratch work - not graded)