STA 9890 - Statistical Learning for Data Mining

In-Class Test 1

This is a closed-note, closed-book exam.
You may not use any external resources other than a (non-phone) calculator.

Name: ____

Instructions

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

This exam is divided into three sections:

- True/False (30 points; 10 questions at three points each)
- Short Answer (50 points; 10 questions at five points each)
- Mathematics of Machine Learning (20 points; one long question in 4 parts)

You have one hour to complete this exam from the time the instructor says to begin. The instructor will give time warnings at: 30 minutes, 15 minutes, 5 minutes, and 1 minute.

When the instructor announces the end of the exam, you must stop **immediately**. Continuing to work past the time limit may be considered an academic integrity violation.

Write your name on the line above *now* before the exam begins.

Each question has a dedicated answer space. Place all answers in the relevant spot. Answers that are not clearly marked in the correct location **will not** receive full credit. Partial credit may be given at the instructor's discretion.

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

Additional pages for scratch work are included at the end of the exam packet.

This is a closed-note, closed-book exam.
You may not use any external resources other than a (non-phone) calculator.

True/False: 30 points total at 3 points each

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

TF1. True/False: Linear regression is a supervised learning method because it has a matrix of features $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ and a vector of responses $\boldsymbol{y} \in \mathbb{R}^n$ which we attempt to predict using $\boldsymbol{X}$.

TRUE FALSE

TF2. True/False: Models with higher training error always have higher test error.

TRUE FALSE

TF3. True/False: For the same level of sparsity, best subsets provides smaller training error than the lasso.

TRUE FALSE

TF4. True/False: Because kernel methods are more flexible than pure linear models, they always provide in-sample (training) error improvements.

TRUE FALSE

TF5. True/False: OLS finds the linear model with the lowest test MSE.

TRUE FALSE

TF6. True/False: When cross-validation is used to select the optimal value of λ in lasso regression, the CV estimate of the out-of-sample (test) error of the selected model is unbiased because the cross-validation error is computed on an unseen 'hold-out' set and not on the training data.

TRUE FALSE

TF7. True/False: Ordinary least squares is BLUE when applied to a VAR (vector autoregressive = multivariate time series) model if the underlying data generating process is truly linear, the errors are mean zero and have constant variance (no 'heteroscedasticity').

TRUE FALSE

TF8. True/False: Reducing variance always increases bias.

TRUE FALSE

TF9. True/False: K-Nearest Neighbors can be used for regression and classification.

TRUE FALSE

TF10. True/False: Linear models are preferred in high-dimensional scenarios because they have a low bias.

TRUE FALSE

Short Answer: 50 points total at 5 points each

SA1. Give an example that demonstrates why the ℓ_0-"norm" is not convex.

SA2. List three reasons we may choose to use a *sparse* model

---

SA3. Compare and contrast spline and kernel models. Give at least 2 key similarities ("compare") and 2 key differences ("contrast").

SA4. The *elastic net* is a combination of ridge and lasso regression:

$$\arg\min_{\boldsymbol{\beta}} \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda_1 \|\boldsymbol{\beta}\|_1 + \frac{\lambda_2}{2} \|\boldsymbol{\beta}\|_2^2.$$

In the case where $\boldsymbol{X}$ is the identity matrix, what is $\hat{\boldsymbol{\beta}}$ in terms of $\boldsymbol{y}, \lambda_1, \lambda_2$?

Hint: Note that the solution is the composition of the ridge and lasso shrinkage operators, applied in any order. That is, if the ridge and lasso shrinkage operators are $S_1(\cdot), S_2(\cdot)$*, the solution will be of the form* $S_1(S_2(\cdot)) = S_2(S_1(\cdot))$.

SA5. Name three advantages of *convexity* in formulating machine learning approaches.

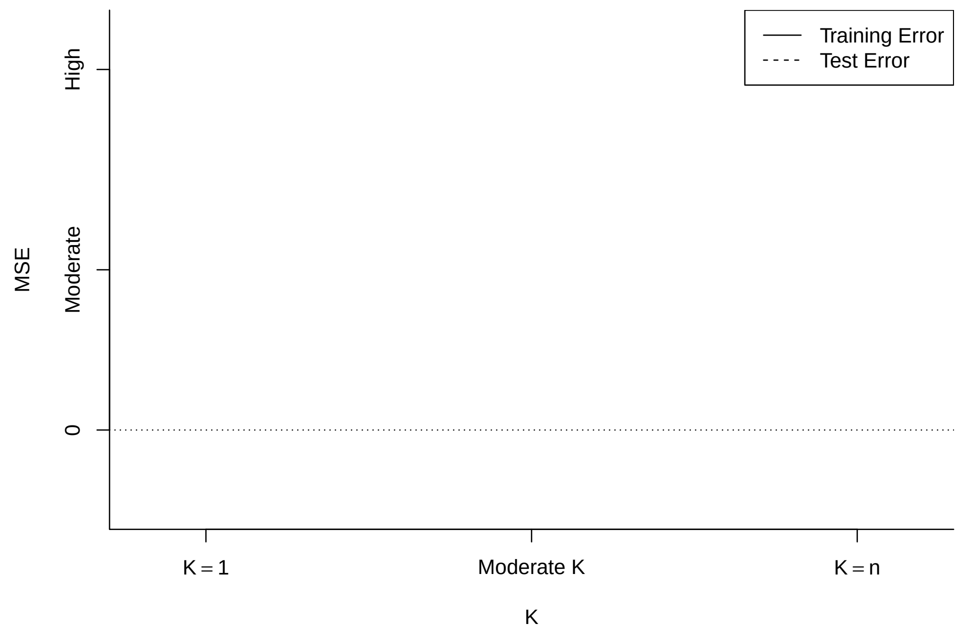
SA6. In your own words, explain why use of a holdout set (or similar techniques) is important for choosing hyperparameters in machine learning.

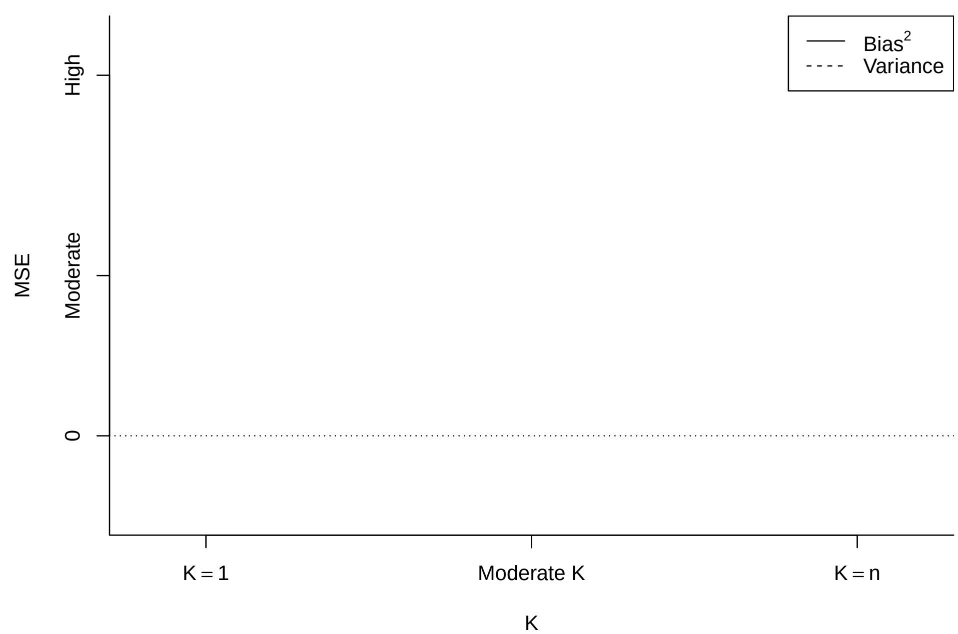
SA7. Give two reasons why is the lasso preferred over best subsets for fitting linear models.

SA8. Rank the following models in terms of (statistical) complexity with 1 being the lowest complexity and 5 being the highest: .

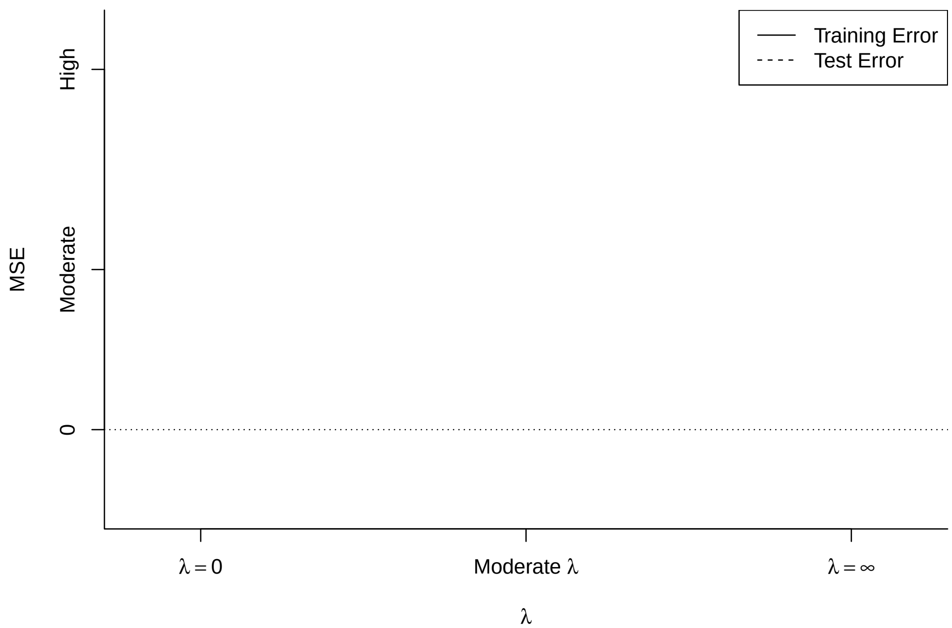
- OLS ____________________
- 1-Nearest Neighbor Regression ____________________
- Cubic Spline Regression ____________________
- Piecewise Cubic Polynomial Regression ____________________
- Ridge Regression ____________________

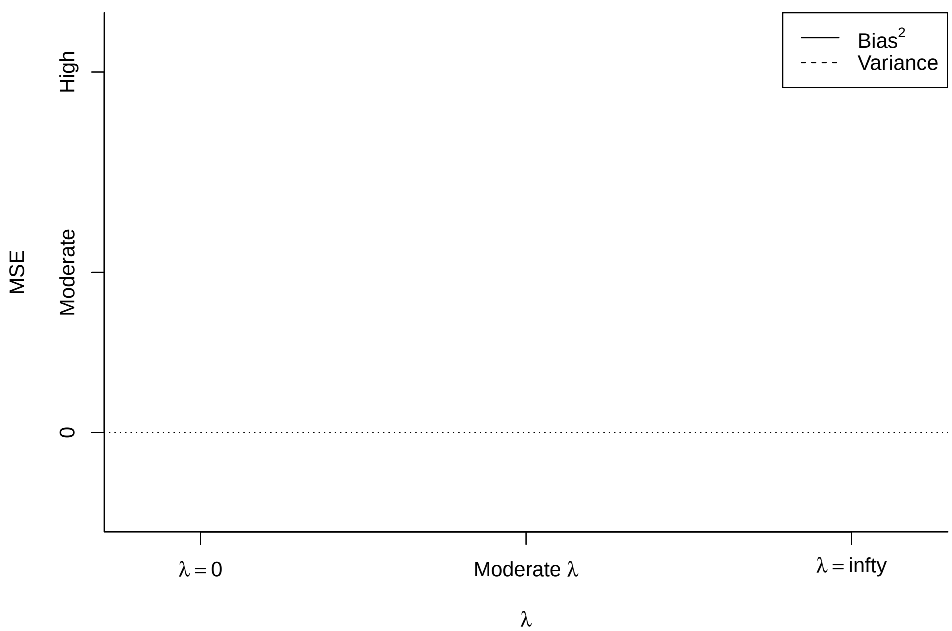
SA9. On the first plot, draw a curve of typical training and test errors for a K-nearest neighbor regressor. On the second plot, darw a curve of typical bias and variance for K-NN regression.





SA10. On the first plot, draw a curve of typical training and test errors for ridge-regression. On the second plot, darw a curve of typical bias and variance for ridge regression.





Mathematics of Machine Learning: 20 points total

In this section, you will analyze so-called *generalized* ridge and lasso penalties. These methods apply the ridge or lasso penalties to something other than the vector of estimated coefficients ($\hat{\boldsymbol{\beta}}$):

$$\arg\min_{\boldsymbol{\beta}} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \frac{\lambda}{2}\|\boldsymbol{D}\boldsymbol{\beta}\|_2^2 \qquad \text{(Generalized Ridge)}$$

$$\arg\min_{\boldsymbol{\beta}} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{D}\boldsymbol{\beta}\|_1 \qquad \text{(Generalized Lasso)}$$

Most commonly, $\boldsymbol{D}$ is taken to be some sort of first-or-second order difference matrix so that

$$\|\boldsymbol{D}_1\boldsymbol{\beta}\|_2^2 = \sum_{i=1}^{p-1}(\beta_{i+1} - \beta_i)^2 \text{ and } \|\boldsymbol{D}_1\boldsymbol{\beta}\|_1 = \sum_{i=1}^{p-1}|\beta_{i+1} - \beta_i| \qquad \text{(First order difference)}$$

and

$$\|\boldsymbol{D}_2\boldsymbol{\beta}\|_2^2 = \sum_{i=1}^{p-2}(\beta_{i+2} - 2\beta_{i+1} + \beta_i)^2 \text{ and } \|\boldsymbol{D}_2\boldsymbol{\beta}\|_1 = \sum_{i=1}^{p-2}|\beta_{i+2} - 2\beta_{i+1} + \beta_i| \qquad \text{(2nd order difference)}$$

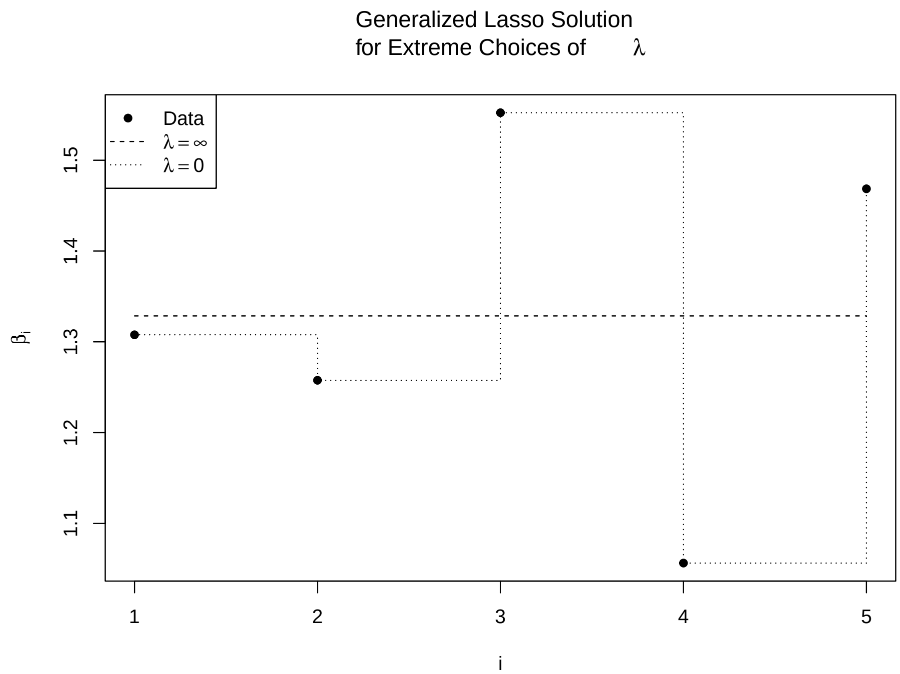
but other choices are also popular.

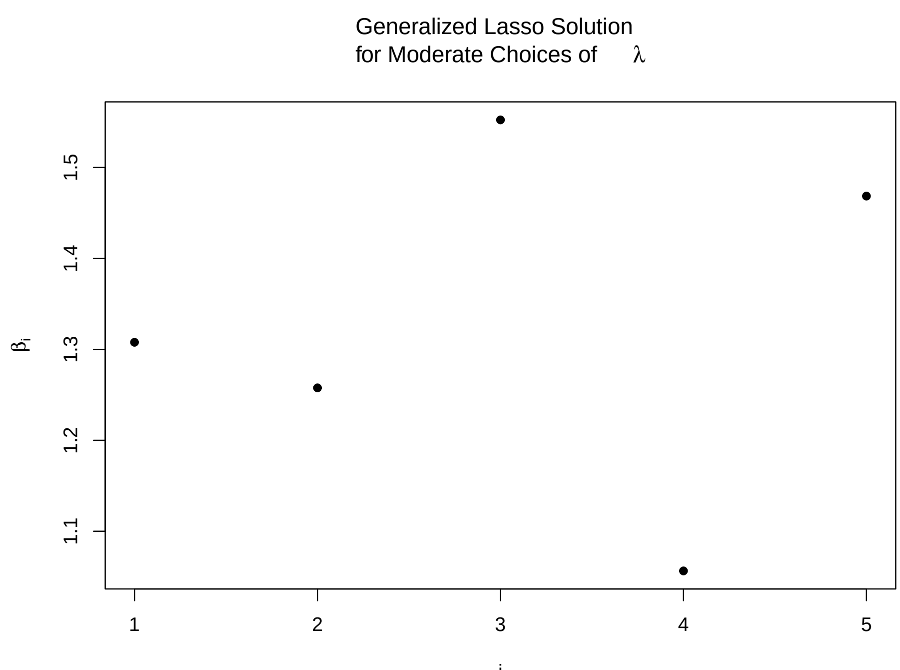
MML1. By analyzing the stationarity (zero gradient) condition, derive the closed-form solution for generalized ridge regression (arbitrary $\boldsymbol{D}$). Box your answer so that it is clearly identifiable. You may assume all relevant matrices are full-rank and/or invertible. (8 points)

Hint: Note that $\partial\|\boldsymbol{D}\boldsymbol{\beta}\|_2^2/\partial\boldsymbol{\beta} = 2\boldsymbol{D}^\top\boldsymbol{D}\boldsymbol{\beta}$.

MML2. In order to build intuition, let us consider the case where $\boldsymbol{X} = \boldsymbol{I}$ and consider what happens when we change λ in the generalized lasso problem. On the following plot, I have drawn the vector of observation $\boldsymbol{y}$ (dots) as well as the $\boldsymbol{D}_1$-generalized lasso solution for $\lambda = 0$ (small dashes) and $\lambda = \infty$ (large dashes).

On the blank set of axes, draw the estimated $\hat{\boldsymbol{\beta}}$ vectors from the $\boldsymbol{D}_1$-generalized lasso at a both 'small-ish' and 'large-ish' value of λ. Label each solution carefully. (4 points)





MML3. What is the relationship between $\boldsymbol{D}_2$-generalized ridge regression and spline methods? (4 points)

MML4. Describe a situation in which generalized ridge or generalized lasso regression would be appropriate. What value of $\boldsymbol{D}$ would you choose for your application? (4 points)

(Blank page for scratch work - not graded)

(Blank page for scratch work - not graded)