**ER131: Data, Environment and Society**

In class midterm exam

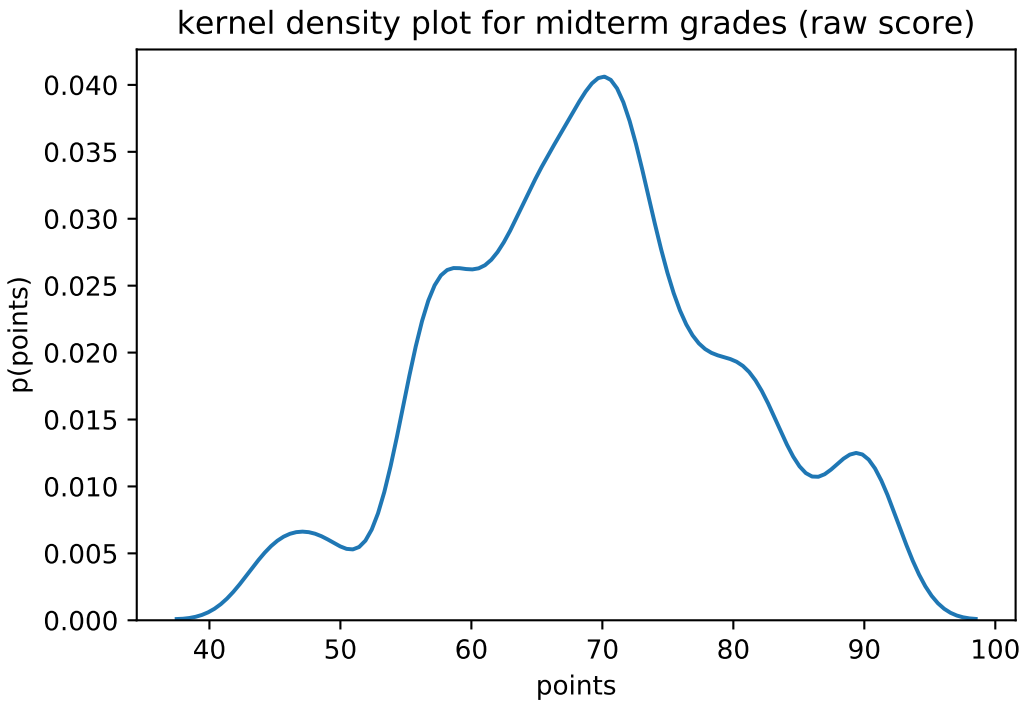
November 19, 2019

SOLUTIONS

**Instructions**:

1. You have 80 minutes to take this test.
2. This booklet should have 7 problems and 8 pages (including this cover sheet). If it doesn't, you must let us know before you begin the exam.
3. There are 95 points total.
4. You may write on both sides of the paper in this booklet. Please be clear about where your work is continued.
5. You must show your work to receive full credit.
6. We will award partial credit.

**Grade distribution (mean 73/95):**



1. (17 points total) Consider the following generic regularized learning problem:

$$\min_{\beta} \sum_{i=1}^{N} \left( y_i - \sum_{j=1}^{p} \beta_j x_{i,j} \right)^2 + \lambda \cdot R(\beta_1, \beta_2, ...\beta_p)$$

(a) (2 points) Using symbols $j, p, \beta_j, x_{0,j}$ and $\hat{y}_0$ write down a prediction for the outcome variable at a "test point" $x_{0,:}$.

$\hat{y}_0 = \sum_{j=1}^{p} \beta_j x_{0,j}$

(b) (2 points) In general, what is the purpose of the regularization term, $R(\cdot)$?

It can be used to tune the bias-variance tradeoff by penalizing the number nonzero coefficients or the size of coefficients.

(c) (3 points) Write down the regularization term used for ridge, best subset selection, and lasso.

Ridge:

$\lambda \sum_{j=1}^{p} \beta_p^2$

Best subset:

$\lambda \sum_{j-1}^{p} I(\beta_p)$

Lasso:

$\lambda \sum_{j-1}^{p} |\beta_p|$

(d) (2 points) Name an advantage to using a ridge loss function. Name a disadvantage.

Advantages: Parameter choices are robust in the presence of collinear features. Fast to solve (closed form).
Disadvantages: does not do "feature selection" (all coefficients will be nonzero). Parameters are biased.

Advantage:

(e) (2 points) Name an advantage to using a lasso loss function. Name a disadvantage.

Advantage: Does feature selection. Faster to solve than subset selection.

Disadvantage: Slower to solve than ridge (no closed form solution). Parameters are biased. Does not handle collinearity well (parameters are unstable to new data samples).

(f) (2 points) Name an advantage to using a best subset loss function. Name a disadvantage.

Advantage: Finds the "true model" (coefficients are unbiased). Does feature selection. Handles correlated features well.

Disadvantage: Slow to solve.

(g) (4 points) Name another loss function that combines the best attributes of ridge and lasso? Name two of those attributes. Name a disadvantage to the method.

The elastic net. It can do feature selection *and* handle correlated features well while still being relatively fast to solve. Disadvantage: You need to tune another hyperparameter.

2. (12 points total) Gradient search. We're going to work with the following data set: $Y = \{0, 1, 2\}$. Consider the constant model:

$$y_i = \beta + \epsilon_i$$

We're going to estimate that model using:

$$\hat{y}_i = \hat{\beta}$$

(a) (2 points) Write down a sum of squares loss function. The model should appear in the loss function.

$\sum_{i=1}^{n}(\hat{\beta} - y_i)^2$

(b) (2 points) Write down an absolute value loss function. The model should appear in the loss function.

$\sum_{i=1}^{n}|\hat{\beta} - y_i|$

(c) (4 points) Assuming the learning rate is 1, do two gradient search steps for the sum of squares loss function and the data set above. Begin with $\hat{\beta}_0 = 0$. (The subscript indexes iteration number.) Solve for $\hat{\beta}_1$ and $\hat{\beta}_2$, the estimate on $\beta$ at the end of the first and second gradient search steps.

$$\begin{aligned}
\frac{dL}{d\beta} &= 2\sum_{i=1}^{n}(\beta - y_i) \\
\hat{\beta}_{k+1} &= \hat{\beta}_k - \alpha\frac{dL}{d\beta} \\
&= \hat{\beta}_k - 2\alpha\sum_{i=1}^{n}(\beta - y_i) \\
\hat{\beta}_0 &= 0 \\
\hat{\beta}_1 &= 0 - 2 * [(0-0) + (0-1) + (0-2)] = 6 \\
\hat{\beta}_2 &= 6 - 2 * [(6-0) + (6-1) + (6-2)] = -24
\end{aligned}$$

(d) (2 points) If you were implementing gradient search for the sum of squares loss function, would you choose a larger or smaller learning rate than the one we used here? Justify your answer.

Smaller. $\alpha = 1$ is causing the algorithm to over-shoot the solution.

(e) (2 points) If you were implementing gradient search for the absolute value loss function, what would need to happen to the learning rate as you approach the optimal $\beta$?

It would need to gradually decrease as it approached the solution. This is because the slope is not zero at the optimum.

3. (12 points) Suppose I give you a data set with a single quantitative feature $X$ and a single quantitative output variable $Y$, and 100 observations. The *linear* regession model for this data is

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Throughout this problem we'll assume $\sigma_\epsilon > 0$.

(a) (2 points) What is the purpose of the $\epsilon_i$ term?

It captures variation in the data that cannot be described by a linear relationship, or more generally variation in $y$ that is uncorrelated with $x$.

(b) (2 points) Write down a *cubic* regression model (include lower order polynomial terms as well). Label your coefficients with $\beta$'s.

$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \epsilon_i$

(c) (2 points) Suppose the *true* relationship between $Y$ and $X$ is linear. If you fit both the linear and cubic model to the data, circle which condition will hold for the *training* error:
$\text{MSE}_{\text{linear}} \geq \text{MSE}_{\text{cubic}}$, $\text{MSE}_{\text{linear}} > \text{MSE}_{\text{cubic}}$, or $\text{MSE}_{\text{linear}} < \text{MSE}_{\text{cubic}}$. Justify your answer.

For training error, $\text{MSE}_{\text{linear}} > \text{MSE}_{\text{cubic}}$. Since $\sigma_\epsilon > 0$, there is variation in the data that a linear model can't describe; a more flexible cubic model will capture that variation.

(d) (2 points) For the same linear relationship in the previous problem, circle which condition will hold for the *cross validated* error:
$\text{MSE}_{\text{linear}} \geq \text{MSE}_{\text{cubic}}$, $\text{MSE}_{\text{linear}} > \text{MSE}_{\text{cubic}}$, or $\text{MSE}_{\text{linear}} < \text{MSE}_{\text{cubic}}$. Justify your answer.

For cross validated error, $\text{MSE}_{\text{linear}} < \text{MSE}_{\text{cubic}}$. The cubic model will over-fit the data, and this will be revealed in the cross-validated error, since it was not used to train the model.

(e) (2 points) Suppose the *true* relationship between $Y$ and $X$ is *non*-linear. Circle which condition will hold for the *training* error:
$\text{MSE}_{\text{linear}} \geq \text{MSE}_{\text{cubic}}$, $\text{MSE}_{\text{linear}} > \text{MSE}_{\text{cubic}}$, or $\text{MSE}_{\text{linear}} < \text{MSE}_{\text{cubic}}$. Justify your answer.

For training error, $\text{MSE}_{\text{linear}} > \text{MSE}_{\text{cubic}}$. If the relationship is nonlinear then a cubic model will capture this variation whereas a linear model will not.

(f) (2 points) Continuing to assume the *true* relationship between $Y$ and $X$ is *non*-linear, circle which condition will hold for the *cross validated* error:
$\text{MSE}_{\text{linear}} \geq \text{MSE}_{\text{cubic}}$, $\text{MSE}_{\text{linear}} > \text{MSE}_{\text{cubic}}$, or $\text{MSE}_{\text{linear}} < \text{MSE}_{\text{cubic}}$. Justify your answer.

For cross validated error, $\text{MSE}_{\text{linear}} > \text{MSE}_{\text{cubic}}$. If the true relationship is nonlinear, both the cross validation data and training data will exhibit this relationship and the nonlinear model will perform better than the linear one.

4. (14 points) Environmental Justice questions.

2-3 sentence question rubric:

- -1 point: basically correct but not completely clear
- -2 points: severely lacking in clarity or not correctly reasoned, but some details correct
- -3 points: lacking any appropriate connections to the question

(a) (2 points) What federal legislation forms the foundation for all federal regulation on EJ?

Title VI of the Civil Rights Act of 1964.

(b) (3 points) In Pastor's *Reclaiming Nature* article he mentions that disparate impact may be the result of "market outcomes." In 2-3 sentences expalin what he means by this.

Pastor is referring the to possibility that hazards could be placed into poor communities because it is less expensive to locate facilities there or even because communities view it as a way to encourage economic development. This in turn could lead to disparate impact by race because communities of color are more likely to be poor than other communities.

(c) (3 points) The Trump administration has proposed a rule to alter how disparate impact gets treated in housing discrimination cases. In 2-3 sentences, explain how statistical learning models matter to this rule.

A prior disparate impact standard put more onus on defendants to explain any practices that appear discriminatory. Trump's proposed rule would require *plaintiffs* to show that a different policy would have averted the outcome. This puts plaintiffs in the position of having to understand the details of the defendant's data and algorithms well enough to propose alternative approaches.

(d) (3 points) The U.S. Department of Housing and Urban Development has proposed a change to the rule for implementation of the Fair Housing Act's Disparate Impact Standard. How could the this change affect the likelihood that market outcomes (as defined by Pastor) are the sole determinant for where people live? Give your answer in 2-3 sentences.

Based on the prior answer, loan approval algorithms could discriminate but be deemed legal if plaintiffs are unable to show that an alternative approach is possible. This might force people of color into homes of lower value.

(e) (3 points) In 2-3 sentences, explain how your answer to the above question matters for environmental justice?

If people of color are unable to obtain high-limit loans, they would be more likely to locate in communities with low property values. If hazardous industries also locate in places where property values are low, as discussed above, then people of color would be more likely to live in the vicinity of hazardous industries.

5. (13 points) Classification methods

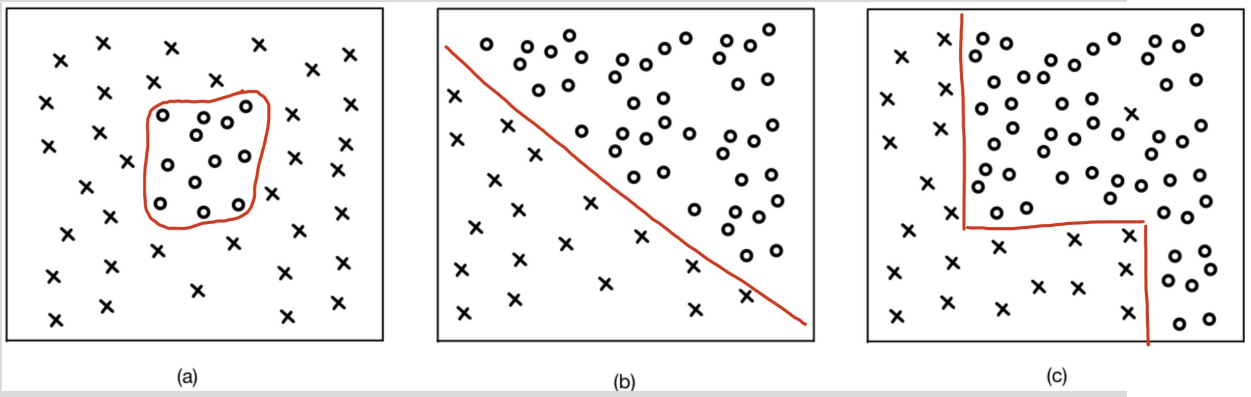
(a) (2 points) Explain why a decision tree built with recursive binary splitting may not be "optimal," in the sense that another tree might deliver a lower mean squared error.

Recursive binary splitting does not "look ahead" to splits that might happen after the next one. In this way recursive binary splitting could miss a sequence of splits that is collectively optimal, even if the first in the sequence of splits is worse than other options.

(b) (9 points) Consider the following three data sets. The horizontal and vertical locations of the symbols represent two separate features. Pick one of the following options to predict "x" versus "o" outcomes for each data set, using each option only once: (i) a single classification tree (i.e. not a model built with an ensemble method), (ii) support vector machine with a linear kernel (iii) support vector machine with a radial kernel and the constraint on margin violations set to zero ($\sum \epsilon_i = 0$). *Draw the decision boundary you think the you'd get in each case*. Justify your answers.

Rubric:

- 1 point for each correct decision boundary
- 1 point for each correct model assignment
- 1 point for each correct justification




Model for (a) and justification:

SVM with radial kernel. The radial kernel allows closed regions. If the model allowed constraint violations, then SVM with radial kernel could also have fit (c).

Model for (b) and justification:

SVM with linear kernel. The boundary is linear.

Model for (c) and justification:

These data are not separable, so the SVM with radial kernel can't fit since it allows no boundary violations. Since (b) is clearly linear, we are left with regression tree for this data set. The boundary is drawn as a composition of sides of rectangles.

(c) (2 points) Suppose you want to build a model that identifies wildfires from satellite data. Give a reason why you'd use a "human in the loop" to facilitate building the data set you'd train your model with.

A data set of "true" wildfires is needed for training. Machines don't know what wildfires are until we tell them.

6. (12 points) Reading question: Hino et al. "Machine learning for environmental monitoring" *Nature Sustainability* (2018).

   (a) (2 points) What do the author's set out to predict?

   The likelihood that a facility will fail a water quality inspection.

   (b) (2 points) Are they doing classification or regression?

   Regression. They are predicting a "risk score," which is a quantitative variable.

   (c) (2 points) Which statistical learning algorithm do they settle on, and what others do they try?

   Regression trees. They tried LASSO, elastic net, random forests.

   (d) (2 points) What resource allocation problem do the authors set out to solve?

   Where should water quality inspectors spend their time?

   (e) (2 points) Explain why following the recommendations made by this paper's model might not yield real-world results.

   (i) Facilities might behave "strategically": it might be easier to reduce their risk score than it is to reduce their likelihood of inspection failure, (ii) their model does not account for potential changes over time.

   (f) (2 points) How could a field trial be used to test how well the algorithm in this paper works?

   If a random set of inspectors use this model, one could compare the number of inspection failures they find to the number of inspection failures found by inspectors that don't use the model (a control group). If the number of failures identified by the inspectors using the model is significantly more than the number of failures identified by the control group, then the algorithm works.

7. (15 points total) Prediction policy problems. Read the following and explain for each whether you would recommend that the agency (i) do a causal analysis for impact assessment, or (ii) build a prediction model for resource allocation? Explain each answer.

   Justify your answer

Rubric:

- -1 point for small lack of clarity, but basically sound reasoning.
- -2 points for some correct insights
- -3 points for gross lack of clarity, but attempted answer

(a) (5 points) An aid agency wants to know if extending the grid to villages in Tanzania will improve per capita income in those villages. They have historical data on what villages received grid extensions in the past and village-level per capita income throughout the country.

This would require a causal analysis for impact assessment. The agency wants to know if a single factor – electrification – will improve economic development. Note that it would be necessary to control for many factors across the villages – or to run a randomized control trial on which villages receive the electrification "treatment." The answer to the problem was somewhat obscured by language indicating a desire to know something in the future. But for a single factor like this, a prediction model will need to built with causal inference methods.

(b) (5 points) PG&E wants to know if their infrastructure will cause a wildfire tomorrow. Data in their possession: (i) historical weather and vegetation data throughout their service territory (ii) location as well as location and timing of grid-caused ignition events. They also have data on forecasted daily weather conditions.

. This would require a predictive model. PG&E wants to take all the information it has at its disposal to predict what will happen tomorrow. As stated, their objective is not to identify the cause of wildfires.

Note that the question is phrased sufficiently ambiguously that we accepted either answer (impact analysis or prediction), so long as the justification was sound.

(c) (5 points) The California Energy Commission wants to know if increasing subsidies toward the purchase of solar PV will increase PV adoption in the future. They have historical data on subsidy amounts and PV adoption.

This would require a causal analysis for impact assessment. The Commission wants to know if a single factor – subsidies – cause more PV adoption. This would help them decide if they should apply more subsidies in the future.