Math 390.4 / 650.3 Spring 2020
Midterm Examination One

Professor Adam Kapelner

Thursday, March 26, 2020

Full Name ____

Code of Academic Integrity

Since the college is an academic community, its fundamental purpose is the pursuit of knowledge. Essential to the success of this educational mission is a commitment to the principles of academic integrity. Every member of the college community is responsible for upholding the highest standards of honesty at all times. Students, as members of the community, are also responsible for adhering to the principles and spirit of the following Code of Academic Integrity.

Activities that have the effect or intention of interfering with education, pursuit of knowledge, or fair evaluation of a student's performance are prohibited. Examples of such activities include but are not limited to the following definitions:

Cheating Using or attempting to use unauthorized assistance, material, or study aids in examinations or other academic work or preventing, or attempting to prevent, another from using authorized assistance, material, or study aids. Example: using an unauthorized cheat sheet in a quiz or exam, altering a graded exam and resubmitting it for a better grade, etc.

I acknowledge and agree to uphold this Code of Academic Integrity.

____ ____________________

signature date

Instructions

This exam is 110 minutes (variable time per question) and closed-book. You are allowed **two** pages (front and back) of a "cheat sheet" and scrap paper but no graphing calculator. Please read the questions carefully. No food is allowed, only drinks.

Problem 1 [7min] This question is about modeling in general.

- [10 pt / 10 pts] Record the letters of all the following that are **true**. Your answer will consist of a string (e.g. `aebgd`) where the order of the letters does not matter.

  (a) A model could be false.

  (b) You can prove a model is true via simulation.

  (c) Only very accurate models can be used for prediction.

  (d) Mathematical models are never accurate enough to be useful.

  (e) Mathematical models built by learning from data require at least one feature.

  (f) There can never be more features than observations when building a model by learning from data.

  (g) When building a model for a continuous response by learning from data, there must must be only continous features.

  (h) Values in a nominal feature can be coerced to numeric values.

  (i) Honest validation gives you an idea about how accurate your model is when using it for future prediction.

  (j) Validation can only be performed on mathematical models that were learned from data.

ahi

Problem 2 [10min] This question is about creating a model learned from data and validating that model. Assume a dataset $\mathbb{D} := \langle X, \boldsymbol{y} \rangle$ where X is an $n \times p$ matrix and $\boldsymbol{y}$ is an $n \times 1$ column vector. The dataset is split into a train and test set of n_{train} observations and n_{test} observations. Let $\mathbb{D}_{\text{train}} := \langle X_{\text{train}}, \boldsymbol{y}_{\text{train}} \rangle$ and $\mathbb{D}_{\text{test}} := \langle X_{\text{test}}, \boldsymbol{y}_{\text{test}} \rangle$ just like we did in class and lab by taking a random partition of the indices $1, 2, \ldots, n$. Assume $g = \mathcal{A}(\mathbb{D}_{\text{train}}, \mathcal{H})$ and $g_{\text{final}} = \mathcal{A}(\mathbb{D}, \mathcal{H})$.

- [10 pt / 20 pts] Record the letters of all the following that are **true**. Your answer will consist of a string (e.g. `aebgd`) where the order of the letters does not matter.

  (a) For all $\mathcal{A}$ we have studied, the model g will be the same regardless of the partition of the indices that divides $\mathbb{D}$ into $\mathbb{D}_{\text{train}}$ and $\mathbb{D}_{\text{test}}$.

  (b) For all $\mathcal{A}$ we have studied, the model g will be the same regardless of the order of the data in $\mathbb{D}_{\text{train}}$.

  (c) Honest validation provides an estimate to how g will do in the future.

  (d) Honest validation provides an estimate to how g_{final} will do in the future.

  (e) Assuming stationarity, comparing $g(X_{\text{train}})$ to $\boldsymbol{y}_{\text{train}}$ provides honest validation for the model g.

  (f) If stationarity cannot be assumed, comparing $g(X_{\text{train}})$ to $\boldsymbol{y}_{\text{train}}$ provides honest validation for the model g

  (g) Assuming stationarity, comparing $g(X_{\text{test}})$ to $\boldsymbol{y}_{\text{test}}$ provides honest validation for the model g.

  (h) If stationarity cannot be assumed, comparing $g(X_{\text{test}})$ to $\boldsymbol{y}_{\text{test}}$ provides honest validation for the model g.

  (i) If $\mathcal{Y} \subseteq \mathbb{R}$, oos standard error of the residuals is given by the formula

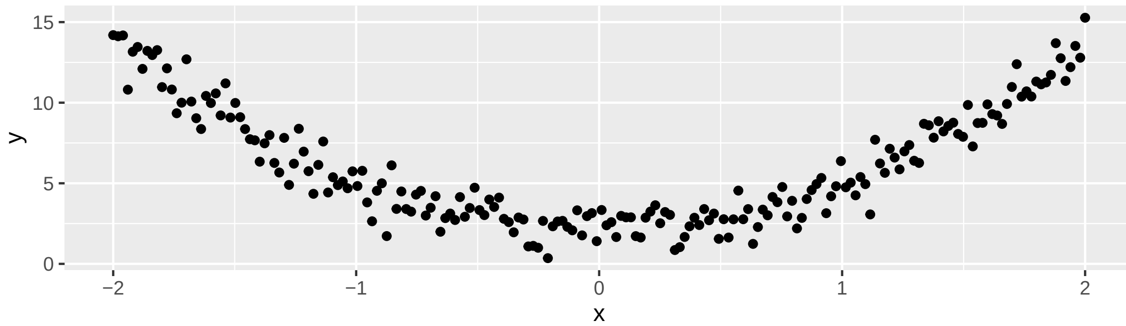
  $$\frac{1}{\sqrt{n_{\text{test}}}} \left|\left|\boldsymbol{y}_{\text{test}} - g(X_{\text{test}})\right|\right|.$$

  (j) If $\mathcal{Y} = \{0, 1\}$, then the oos misclassification rate is given by

  $$\frac{1}{n_{\text{test}}} \left|\boldsymbol{y}_{\text{test}} - g(X_{\text{test}})\right|.$$

$\boxed{\text{bcgi}}$

Problem 3 [7min] A dataset of $n = 200$ and $p = 1$ is collected. Here is a plot of the raw $\mathbb{D}$:



Let X be the random variable (r.v.) that realized x and let Y be the r.v. that realized y.

- [10 pt / 30 pts] Record the letters of all the following that are **true**. Your answer will consist of a string (e.g. `aebgd`) where the order of the letters does not matter.

  (a) X and Y are likely independent.

  (b) X and Y are likely dependent.

  (c) X and Y are likely associated.

  (d) X and Y are likely not associated.

  (e) X and Y likely have covariance zero.

  (f) X and Y likely have covariance nonzero.

  (g) X and Y likely have correlation zero.

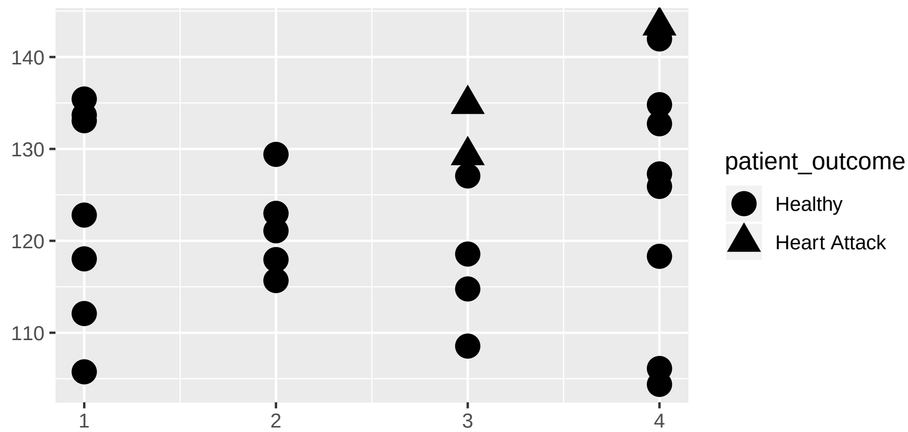
  (h) X and Y likely have correlation nonzero.

  (i) If all values of $x > 0$ were dropped from $\mathbb{D}$, then it would appear that X and Y likely have covariance positive.

  (j) If all values of $x > 0$ were dropped from $\mathbb{D}$, then it would appear that X and Y likely have covariance negative.

bcfhj

Problem 4 [13min] The raw $\mathbb{D}$ with $n = 27$ is plotted below where x_1 is on the horizontal axis and x_2 is on the vertical axis. The binary response y measures patient outcome and is depicted by different shapes (see the illustration's legend).



- [10 pt / 40 pts] Record the letters of all the following that are **true**. Your answer will consist of a string (e.g. `aebgd`) where the order of the letters does not matter.

(a) x_1 is most likely an ordinal variable.

(b) x_1 is most likely a nominal variable.

(c) If $\mathcal{A} =$ OLS, then it would not be able to return a g.

(d) If $\mathcal{A} =$ perceptron learning algorithm without a limit on iterations, then it would not be able to return a g.

(e) If $\mathcal{A} =$ SVM with the Vapnik function, then you would need to specify the value of λ to be able to return a g.

(f) If $\mathcal{A} =$ SVM with the Vapnik function, then g would have zero average hinge error.

(g) If $\mathcal{A} =$ SVM with the Vapnik function and $\lambda = 0$, then g would divide $\mathbb{D}$ only by using x_1.

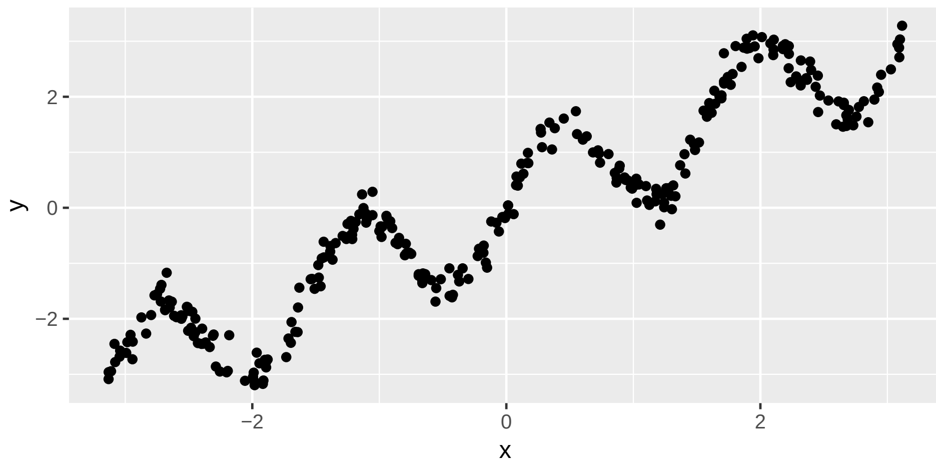
(h) Regardless of the $\mathcal{A}$ used, R^2 would be a preferred metric to assess model accuracy.

(i) It is possible to design an $\mathcal{A}$ that could return a model g that gives a perfect fit.

(j) It $\mathcal{A} =$ KNN where $K = n$, then $g(\boldsymbol{x})$ will be the same for all observations $\boldsymbol{x} \in \mathbb{D}$.

$\boxed{\text{adeij}}$

Problem 5 [15min] A raw $\mathbb{D}$ is plotted below:



- [10 pt / 50 pts] Record the letters of all the following that are **true**. Your answer will consist of a string (e.g. `aebgd`) where the order of the letters does not matter.

  (a) $\mathcal{X}$ is likely $[-2, 2]$.

  (b) $\mathcal{Y}$ is likely $\subseteq \mathbb{R}$.

  (c) If $\mathcal{H} = \{a \,:\, a \in \mathbb{R}\}$ then a rational $\mathcal{A}$ would return g where $a \approx 0$.

  (d) If $\mathcal{H} = \{a + b\mathbb{1}_{x>0} \,:\, a, b \in \mathbb{R}\}$ then it is likely $f \in \mathcal{H}$.

  (e) If $\mathcal{H} = \{a + b\mathbb{1}_{x>0} \,:\, a, b \in \mathbb{R}\}$ then a rational $\mathcal{A}$ would return g where $b > a$.

  (f) If $\mathcal{H} = \{a + b\mathbb{1}_{x>0} \,:\, a, b \in \mathbb{R}\}$ and $\mathcal{A} = \text{OLS}$, then it would be impossible to compute R^2 for g.

  (g) If $\mathcal{H}$ = the space of all continuous functions, then $t \in \mathcal{H}$.

  (h) If $\mathcal{H} = \{a + bx \,:\, a, b \in \mathbb{R}\}$ then a rational $\mathcal{A}$ would return g where $a < 0$.

  (i) With $\mathcal{H}$ specified optimally, the error due to estimation will likely be low.

  (j) With $\mathcal{H}$ specified optimally, the error due to misspecification will likely be low.

$\boxed{\text{bcehij}}$

Problem 6 [17min] Let $\boldsymbol{X} = [\mathbf{1}_n \mid \boldsymbol{x}_1 \mid \ldots \mid \boldsymbol{x}_p] \in \mathbb{R}^{n \times (p+1)}$ and $\text{rank}[\boldsymbol{X}] = p + 1$ and $\boldsymbol{y} \in \mathbb{R}^n$. Assume that the vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$ are mean-centered (i.e. the sample average of all their entries is zero) and that $\boldsymbol{y}$ is also mean-centered. Your modeling task is to model the response using the n observations. Your $\mathcal{A} = \text{OLS}$. Let $\boldsymbol{b}$ be the vector of OLS estimates for the $p+1$ features, let $\boldsymbol{\beta}$ be the slope coefficients in the optimal linear model, $\boldsymbol{H}$ be the orthogonal projection matrix onto the colsp $[\boldsymbol{X}]$, $\hat{\boldsymbol{y}}$ is the vector of predictions for the n observations and $\boldsymbol{e}$ are the residuals.

- [10 pt / 60 pts] Record the letters of all the following that are **true**. Your answer will consist of a string (e.g. `aebgd`) where the order of the letters does not matter.

  (a) $\boldsymbol{b} = \underset{\boldsymbol{w} \in \mathbb{R}^{p+1}}{\arg\min} \left\{ (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}) \right\}$.

  (b) $\boldsymbol{X}^T\boldsymbol{X}$ is a full rank $n \times n$ matrix.

  (c) $\boldsymbol{H}$ is a full rank $n \times n$ matrix.

  (d) $\boldsymbol{H}\boldsymbol{e} = \mathbf{0}_n$.

  (e) $\boldsymbol{H}\boldsymbol{X}\boldsymbol{b} - \hat{\boldsymbol{y}} = \mathbf{0}_n$.

  (f) $\boldsymbol{H}\boldsymbol{X}\boldsymbol{\beta} - \hat{\boldsymbol{y}} = \mathbf{0}_n$.

  (g) $\mathbf{1}_n^\top \boldsymbol{H} \mathbf{1}_n = p + 1$.

  (h) $\boldsymbol{x}_2^\top \boldsymbol{H} \boldsymbol{x}_2 = (n-1)s_{x_2}^2$.

  (i) $\text{rank}[\boldsymbol{I} - \boldsymbol{H}] = n - p - 1$.

  (j) If $p = 0$ then $\hat{\boldsymbol{y}} = \mathbf{0}_n$.

$\boxed{\text{adehij}}$

Problem 7 [13min] We use the same setup as in Problem 6. Let $\boldsymbol{X} = [\mathbf{1}_n \mid \boldsymbol{x}_1 \mid \ldots \mid \boldsymbol{x}_p] \in \mathbb{R}^{n \times (p+1)}$ and $\text{rank}\,[\boldsymbol{X}] = p + 1$ and $\boldsymbol{y} \in \mathbb{R}^n$. Assume that the vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$ are mean-centered (i.e. the sample average of all their entries is zero) and that $\boldsymbol{y}$ is also mean-centered. Your modeling task is to model the response using the n observations. Your $\mathcal{A} =$ OLS. Let $\boldsymbol{b}$ be the vector of OLS estimates for the $p+1$ features, let $\boldsymbol{\beta}$ be the slope coefficients in the optimal linear model, $\boldsymbol{H}$ be the orthogonal projection matrix onto the $\text{colsp}\,[\boldsymbol{X}]$, $\hat{\boldsymbol{y}}$ is the vector of predictions for the n observations and $\boldsymbol{e}$ are the residuals. But now we progressively add columns consisting of entries of random noise to the matrix $\boldsymbol{X}$ which remains full rank each time a column is appended.

- [10 pt / 70 pts] Record the letters of all the following that are **true**. Your answer will consist of a string (e.g. `aebgd`) where the order of the letters does not matter.

  With each additional column appended, ...

  (a) SST decreases.
  (b) $||\hat{\boldsymbol{y}}||^2$ increases.
  (c) $\dim\,[\boldsymbol{H}]$ remains the same.
  (d) $\text{rank}\,[\boldsymbol{H}]$ remains the same.
  (e) the RMSE increases.
  (f) the generalization error of the model increases.
  (g) each residual's absolute value will decrease.
  (h) each slope coefficient (the entries in $\boldsymbol{b}$) will decrease.
  (i) the angle between $\hat{\boldsymbol{y}}$ and $\boldsymbol{y}$ gets closer to zero.
  (j) with $p > n$, $\mathcal{A}$ fails to run.

$\boxed{\text{bcfij}}$

Problem 8 [12min] We use the same setup as in Problem 6. Let $\boldsymbol{X} = [\mathbf{1}_n \mid \boldsymbol{x}_1 \mid \ldots \mid \boldsymbol{x}_p] \in \mathbb{R}^{n\times(p+1)}$ and $\text{rank}\,[\boldsymbol{X}] = p+1$ and $\boldsymbol{y} \in \mathbb{R}^n$. Assume that the vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$ are mean-centered (i.e. the sample average of all their entries is zero) and that $\boldsymbol{y}$ is also mean-centered. Your modeling task is to model the response using the n observations. Your $\mathcal{A} = \text{OLS}$. Let $\boldsymbol{b}$ be the vector of OLS estimates for the $p+1$ features, let $\boldsymbol{\beta}$ be the slope coefficients in the optimal linear model, $\boldsymbol{H}$ be the orthogonal projection matrix onto the $\text{colsp}\,[\boldsymbol{X}]$, $\hat{\boldsymbol{y}}$ is the vector of predictions for the n observations and $\boldsymbol{e}$ are the residuals. But now we use Q-R decomposition to find matrices such that $\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{R}$ where $\boldsymbol{q}_j$ denotes the jth column of $\boldsymbol{Q}$.

- [10 pt / 80 pts] Record the letters of all the following that are **true**. Your answer will consist of a string (e.g. `aebgd`) where the order of the letters does not matter.

  (a) The dimension of $\boldsymbol{Q}$ is the same as $\boldsymbol{X}$.

  (b) $\text{colsp}\,[\boldsymbol{X}] = \text{colsp}\,[\boldsymbol{Q}]$.

  (c) $\boldsymbol{R}$ is a full rank $n \times n$ matrix.

  (d) $\boldsymbol{R}$ cannot be inverted.

  (e) $\boldsymbol{y} \in \text{colsp}\,[\boldsymbol{Q}]$.

  (f) $\hat{\boldsymbol{y}} \in \text{colsp}\,[\boldsymbol{Q}]$.

  (g) The first column of $\boldsymbol{Q}$ is a scalar multiple of $\mathbf{1}_n$.

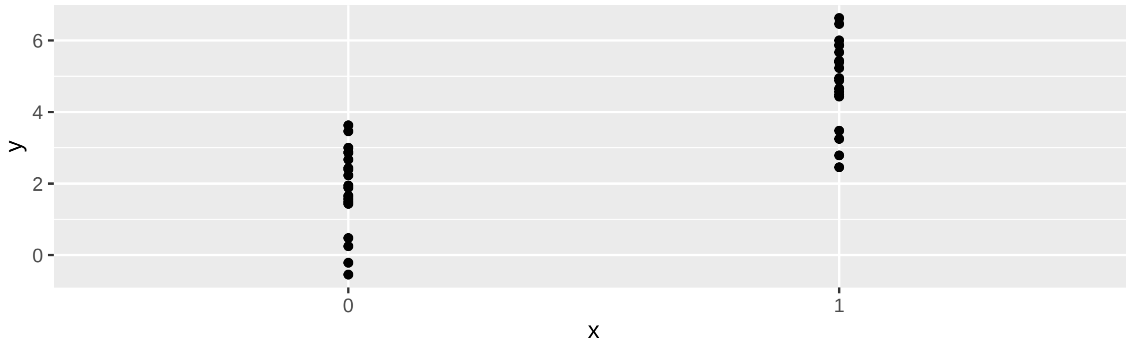
  (h) The last column of $\boldsymbol{Q}$ is a scalar multiple of $\boldsymbol{x}_p$.

  (i) $\boldsymbol{H} = \boldsymbol{Q}\boldsymbol{Q}^\top$.

  (j) $\hat{\boldsymbol{y}} = \text{Proj}_{\boldsymbol{q}_1}[\boldsymbol{y}] + \text{Proj}_{\boldsymbol{q}_2}[\boldsymbol{y}] + \ldots + \text{Proj}_{\boldsymbol{q}_{p+1}}[\boldsymbol{y}]$

abfgij

Problem 9 [14min] A dataset of $n = 40$ and $p = 1$ is collected. Here is a plot of the raw $\mathbb{D}$:



And here is some `R` code that uses the raw $\mathbb{D} = \langle \boldsymbol{x}, \boldsymbol{y} \rangle$ with its output:

```
> round(summary(y[x == 0]), 2)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  -0.54    1.47    1.91    1.86    2.72    3.63
> round(summary(y[x == 1]), 2)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   2.46    4.47    4.91    4.86    5.72    6.63
> mod = lm(y ~ x)
```

- [10 pt / 90 pts] Record the letters of all the following that are **true**. Your answer will consist of a string (e.g. `aebgd`) where the order of the letters does not matter.

  (a) x is termed a "dummy variable".

  (b) Modeling the response y here is termed a "classification problem".

  (c) The model g in the object `mod` is a linear model.

  (d) The R^2 is likely to be low for the model g in the object `mod`.

  (e) The R^2 is likely to be high for the model g in the object `mod`.

  (f) When `coef(mod)` is called, it returns $b_1 = 4.86$.

  (g) When `coef(mod)` is called, it returns $b_1 = 3.00$.

  (h) If `mod = lm(y ~ 0 + x)` and `coef(mod)` is called, it returns $b_1 = 4.86$.

  (i) If `mod = lm(y ~ 0 + x)` and `coef(mod)` is called, it returns $b_1 = 3.00$.

  (j) The model g in the object `mod` can only predict for $x \in \{0, 1\}$.

acdgh