Data 100, Midterm 2

Summer 2020

Instructions:

- This exam consists of 8 questions, worth a total of 65 points.
- This exam must be completed and submitted in the **105 minute** time period ending at **8:45 PM PDT**, unless you have accommodations supported by a DSP letter or are taking an alternate.
- This exam is open-book and open-Internet, but **collaboration is strictly forbidden**.
- **Please show your work for computation questions as we may award partial credit. For multiple-choice questions, if you are writing on blank paper, clearly write the letter of the choice you are selecting.**
- Please write your initials on the top of every page.

Statement of Academic Integrity

Please **COPY, SIGN, and DATE** the following statement on your exam page. **You must do this even if you are writing the exam on blank sheets of paper.**

As a member of the UC Berkeley community, I act with honesty, integrity, and respect for others. I will neither give nor receive assistance while taking this exam. I understand that I must write this exam on paper.

Modeling and Loss Functions [11 Pts]

1. Suppose we observe the growth (g) of some novel bacteria over time (t). We collect n pairs of data, $\{(t_1, g_1), (t_2, g_2), ..., (t_n, g_n)\}$ and decide to model growth as $\hat{g} = \theta \log(t)$. We decide to use squared loss as our loss function. Assume that we are not using regularization.

(a) [2 Pts] Suppose we are provided with the following data.

g_i	$\log(t_i)$
3	2
6	3
9	4

What is the **average** loss for our model on this dataset, as a function of θ? (The only variable your answer should contain is θ. No need to simplify.)

Solution:

$$R(\theta) = \frac{1}{3}[(3 - 2\theta)^2 + (6 - 3\theta)^2 + (9 - 4\theta)^2]$$

(b) [2 Pts] Write an expression for the average loss of our model, $R(\theta)$, for any set of data with n observations. (Your answer should be in terms of θ, g_i, t_i, and n, and should not involve the specific set of values provided in part a.)

Solution:

$$R(\theta) = \frac{1}{n}\sum_{i=1}^{n}(g_i - \theta\log(t_i))^2$$

(c) [4 Pts] What is the value of $\hat{\theta}$ that minimizes $R(\theta)$ that you determined in part b?

Solution: One could take the derivative with respect to θ and solve to derive the correct answer,

$$\hat{\theta} = \frac{\sum_{i=1}^{n} g_i \log(t_i)}{\sum_{i=1}^{n}(\log(t_i))^2}$$

However, there's an easier solution. In Homework 5, students solve for the value of θ that minimizes MSE for the model $\hat{y} = \theta x$, and that value is $\hat{\theta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$. The solution here is identical, with $\log(t_i)$ instead of x_i. This works because θ isn't "inside" the $\log$ part of the model.

(d) [3 Pts] Suppose our model is now

$$\hat{g} = \theta_0 + \theta_1 \log(t) + \theta_2(\log(t))^2$$

Is this a linear model? Answer "yes" or "no".

- If yes, write out the design matrix $\mathbb{X}$ for this model using the data in part a.
- If no, explain in one sentence why.

Solution:

Yes, this model is linear because it is linear in terms of the parameters θ.

$$\mathbb{X} = \begin{bmatrix} 1 & 2 & 4 \\ 1 & 3 & 9 \\ 1 & 4 & 16 \end{bmatrix}$$

Fun with Linear Regression [13 Pts]

2. Consider the simple linear regression model without an intercept, $\hat{y} = \theta x$. Assume throughout the entirety of this problem that we are performing least squares linear regression (i.e. that we are choosing parameters that minimize average squared loss), and that we are not using regularization.

(a) [1 Pt] True or False: The point $(\bar{x}, \bar{y})$ always lies on the regression line $\hat{y} = \hat{\theta} x$.

○ A. True

○ **B. False**

Solution: It does not. With an intercept term this statement is true, but without one this property doesn't hold.

A counterexample can be had by picking any set values of x whose mean is 0, and any set of y values whose mean is not 0. Then, $0 \cdot \hat{\theta}$ is 0, but this cannot be equal to $\bar{y}$, which we defined to be non-zero.

(b) [2 Pts] Suppose we create a new feature x' using $x' = 10x$. Let $r(x, y)$ be the correlation coefficient that we discussed in class.

True or False: $r(x, y) = r(x', y)$.

○ **A. True**

○ B. False

Solution: r is a measure of linear association. Multiplying one set of values by 10 doesn't change the strength of their linear association with the other set of values.

(c) [3 Pts] Now suppose we use x' as the independent variable in our simple linear model without an intercept. That is, our model is now $\hat{y} = \beta x'$.

Are the optimal slopes of our two models the same — in other words, is $\hat{\beta}$ for our new model equal to $\hat{\theta}$ for the original model? If yes, explain in one sentence why. If no, give an expression for $\hat{\beta}$ in terms of $\hat{\theta}$.

Solution: No, the optimal slopes are different, and $\hat{\beta} = \frac{1}{10}\hat{\theta}$.

As seen in Homework 5, the optimal $\hat{\theta}$ is $\frac{\sum x_i y_i}{\sum x_i^2}$. Replacing x_i with $10x_i$:

$$
\begin{aligned}
\hat{\theta} &= \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} \\
\hat{\beta} &= \frac{\sum_{i=1}^{n} 10 x_i y_i}{\sum_{i=1}^{n} (10x_i)^2} \\
&= \frac{10 \sum_{i=1}^{n} x_i y_i}{100 \sum_{i=1}^{n} x_i^2} \\
&= \frac{1}{10} \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} \\
&= \frac{1}{10} \hat{\theta}
\end{aligned}
$$

(d) [3 Pts] Suppose we decide to now include an intercept term in our simple linear model. We fit two models. One uses x, and the other uses x' as defined above.

- Model A: $\hat{y} = \theta_0 + \theta_1 x$
- Model B: $\hat{y} = \beta_0 + \beta_1 x'$

Is $\hat{\theta}_1 = \hat{\beta}_1$? If yes, explain in one sentence why. If no, give an expression for $\hat{\beta}_1$ in terms of $\hat{\theta}_1$.

Solution:
No, the optimal slopes are different, and $\hat{\beta}_1 = \frac{1}{10}\hat{\theta}_1$.
We know from lecture that the optimal $\hat{\theta}_1$ is $r \cdot \frac{\sigma_y}{\sigma_x}$. The correlation coefficient doesn't change when we multiply a set of values by a constant (as established in part b), and $\sigma_{x'} = 10\sigma_x$.

(e) [4 Pts] Is $\hat{\theta}_0 = \hat{\beta}_0$? If yes, explain in one sentence why. If no, give an expression for $\hat{\beta}_0$ in terms of $\hat{\theta}_0$.

Solution: Yes. Recall from lecture, $\hat{\theta}_0 = \bar{y} - \hat{\theta}_1 \bar{x}$. We established in the above part that $\hat{\beta}_1 = \frac{1}{10}\hat{\theta}_1$, and $\bar{x'} = 10\bar{x}$. $\bar{y}$ is the same for both models. Putting these facts together:

$$
\begin{aligned}
\hat{\theta}_0 &= \bar{y} - \hat{\theta}_1 \bar{x} \\
\hat{\theta}_0 &= \bar{y} - \frac{1}{10} \cdot \hat{\theta}_1 \cdot (10\bar{x}) \\
&= \bar{y} - \hat{\beta}_1 \bar{x'} \\
&= \hat{\beta}_0
\end{aligned}
$$

OLS vs. Ridge Regression [10 Pts]

3. Consider the least squares regression model, $\hat{\mathbb{Y}} = \mathbb{X}\theta$. Assume that $\mathbb{X}$ and $\mathbb{Y}$ refer to the design matrix and true response vector for our training data.

Let $\hat{\gamma}$ be the parameter vector that minimizes mean squared error without regularization. Specifically:

$$\hat{\gamma} = \arg\min_{\gamma} \frac{1}{n}||\mathbb{Y} - \mathbb{X}\gamma||_2^2$$

Let $\hat{\beta}$ be the parameter vector that minimizes mean squared error with L_2 regularization, using a non-negative regularization hyperparameter λ (i.e. ridge regression). Specifically:

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{n}||\mathbb{Y} - \mathbb{X}\beta||_2^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

For each of the following problems, fill in the blank.

(a) [1 Pt] If we set $\lambda = 0$, then $||\hat{\gamma}||_2^2$ is __________ $||\hat{\beta}||_2^2$.

○ A. less than

○ **B. equal to**

○ C. greater than

○ D. impossible to tell

(b) [2 Pts] For each of the remaining parts, you can assume that λ is set such that the predicted response vectors for our two models ($\hat{\mathbb{Y}} = \mathbb{X}\hat{\gamma}$ and $\hat{\mathbb{Y}} = \mathbb{X}\hat{\beta}$) is different.

The **training** RMSE of the model $\hat{\mathbb{Y}} = \mathbb{X}\hat{\gamma}$ is __________ than the model $\hat{\mathbb{Y}} = \mathbb{X}\hat{\beta}$.

○ **A. less than**

○ B. equal to

○ C. greater than

○ D. impossible to tell

(c) [2 Pts] Now, assume we've fit both models using our training data, and evaluate both models on some unseen testing data.

The **test** RMSE of the model $\hat{\mathbb{Y}} = \mathbb{X}\hat{\gamma}$ is __________ than the model $\hat{\mathbb{Y}} = \mathbb{X}\hat{\beta}$.

○ A. less than

○ B. equal to

○ C. greater than

○ **D. impossible to tell**

(d) [2 Pts] Assume that our design matrix $\mathbb{X}$ contains a column of all ones. The sum of the residuals of our model $\hat{\mathbb{Y}} = \mathbb{X}\hat{\beta}$ is ____________.

○ A. equal to 0

○ **B. not necessarily equal to 0**

(e) [1 Pt] As we increase λ, the bias of the model $\hat{\mathbb{Y}} = \mathbb{X}\hat{\beta}$ tends to ____________.

○ **A. increase**

○ B. stay the same

○ C. decrease

(f) [1 Pt] As we increase λ, the model variance of the model $\hat{\mathbb{Y}} = \mathbb{X}\hat{\beta}$ tends to ____________.

○ A. increase

○ B. stay the same

○ **C. decrease**

(g) [1 Pt] As we increase λ, the observation variance of the model $\hat{\mathbb{Y}} = \mathbb{X}\hat{\beta}$ tends to ____________.

○ A. increase

○ **B. stay the same**

○ C. decrease

Feature Engineering [5 Pts]

4. Suppose we have one qualitative variable that that we convert to numerical values using one-hot encoding. We've shown the first four rows of the resulting design matrix below:

a	**b**	**c**
1	0	0
1	0	0
0	0	1
0	1	0

(a) [2 Pts] Say we train a linear model m_1 on these data. Then, we replace all of the 1 values in column **a** with 3's and all of the 1 values in column **b** with 2's and train a new linear model m_2. Neither m_1 nor m_2 have an intercept term. On the training data, the average squared loss for m_1 will be ___ that of m_2.

○ A. greater than

○ B. less than

○ **C. equal to**

○ D. impossible to tell

Solution: Note that we can just re-scale our weights accordingly. Any model we can get with m_1 we can also get with m_2 (and vice versa).

(b) [2 Pts] To account for the intercept term, we add a column of all ones to our design matrix from part a. That is, the resulting design matrix has four columns: **a** with 3's instead of 1's, **b** with 2's instead of 1's, **c**, and a column of all ones. What is the rank of the new design matrix with these four columns?

○ A. 1

○ B. 2

○ **C. 3**

○ D. 4

Solution: Note that the column $\mathbf{c} =$ intercept column $- \frac{1}{3}\mathbf{a} + \frac{1}{2}\ \mathbf{b}$. Hence, there is a linear dependence relationship, meaning that one of the columns is redundant and that the rank of the new design matrix is 3.

(c) [1 Pt] Suppose we divide our sampling frame into three clusters of people, numbered 1, 2, and 3. After we survey people, along with our survey results, we save their cluster number as a new feature in our design matrix. Before training a model, what should we do with the cluster column? (Note: This part is independent of parts a and b.)

○ A. Leave as is

○ **B. One-hot encode it**

○ C. Normalize it

○ D. Use bag of words

Solution: The cluster number is a categorical variable, so it should be one-hot encoded.

(Cross-)validate me [6 Pts]

5. Suppose we have a design matrix $\mathbb{X}$ comprising of n observations, d features, and an additional intercept term. We decide to use $\mathbb{X}$ to create, tune, and evaluate a regularized linear regression model with two regularization hyperparameters, λ_1 and λ_2. We have i different choices for λ_1 and j different choices for λ_2, so there are $i \cdot j$ possible combinations of λ_1 and λ_2.

   We set aside 20% of our data to use as a test set and perform k-fold cross-validation to tune our hyperparameters. We compute the cross-validation error of each of the $i \cdot j$ possible combinations of λ_1 and λ_2 values, and our goal is to find the combination of values with the lowest cross-validation error. All following answers should be expressed in terms of i, j, n, d, k, and/or constants only (except for part c).

   (a) [2 Pts] For a single combination of hyperparameter values, how many model parameters do we fit?

   Solution: $k \cdot (d + 1)$

   (b) [2 Pts] How many observations are in the validation set of each fold?

   Solution: $\frac{0.8n}{k}$

   (c) [2 Pts] How many model parameters do we calculate in total? Your answer can be in terms of A, your answer to part a.

   Solution: $i \cdot j \cdot k \cdot (d + 1)$

Gradient Descent [6 Pts]

6. We define the ”sigmoid loss” to be $L(y, \hat{y}) = \sigma(\hat{y} - y)$. Suppose we decide to use the constant model $\hat{y} = \theta$.

   (a) [3 Pts] The loss for a single observation i is $\sigma(\theta - y_i)$. Determine the partial derivative of this loss with respect to θ. Recall from Homework 1 that $\frac{\partial}{\partial x}\sigma(x) = \sigma(x)(1 - \sigma(x))$. You can have σ in your answer.

   Solution:

   $$\frac{\partial}{\partial \theta}\sigma(\theta - y_i) = \sigma(\theta - y_i)\big(1 - \sigma(\theta - y_i)\big)$$

   (b) [1 Pt] For the constant model $\hat{y} = \theta$, which of the following values of θ would result in the lowest sigmoid loss for a single observation i?

   ○ A. $\theta = y_i + 100$

   ○ B. $\theta = y_i$

   [✓ ○ **C.** $\theta = y_i - 100$]

   (c) [2 Pts] The average sigmoid loss across our entire dataset is

   $$R(\theta) = \frac{1}{n}\sum_{i=1}^{n}\sigma(\theta - y_i)$$

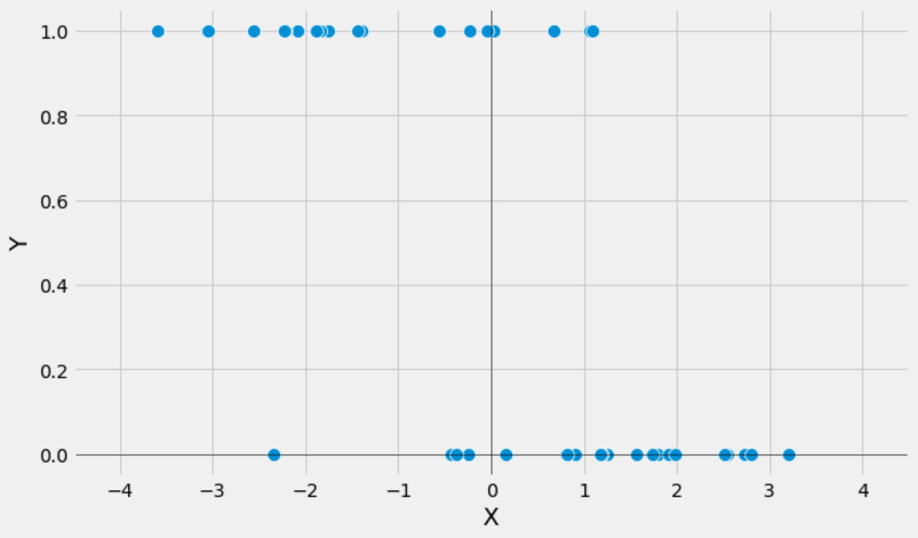
   Write out the standard gradient descent update equation for this average loss. Your answer should include $\theta^{(t+1)}$, $\theta^{(t)}$, α (the learning rate), and your answer from part a, for which you may use the letter A as a substitute.

   Solution:

   $$\theta^{(t+1)} = \theta^{(t)} - \alpha \cdot \frac{1}{n}\sum_{i=1}^{n}\big[\sigma(\theta - y_i)\big(1 - \sigma(\theta - y_i)\big)\big]$$

Linear Separability [7 Pts]

7. Our goal is to fit a logistic regression model with x (a continuous quantitative variable) as the sole feature and y (a binary class label) as the response. A scatter plot of x and y is shown below.



(a) [1 Pt] Is this dataset linearly separable?

○ A. Yes

○ **B. No**

○ C. Impossible to tell

(b) [2 Pts] We now wish to fit a logistic regression model with an intercept term to our data. Remember, such a model takes the form $\hat{y} = P(Y = 1|x) = \sigma(\theta_0 + \theta_1 x)$.

Which of the following is *closest* to the scale parameter θ_1?

○ A. -100

○ **B. -1**

○ C. 1

○ D. 100

(c) [2 Pts] Which of the following is *closest* to $P(Y = 1|x = 1)$? (*Hint: You may want to sketch what this model looks like before answering.*)

○ A. 0

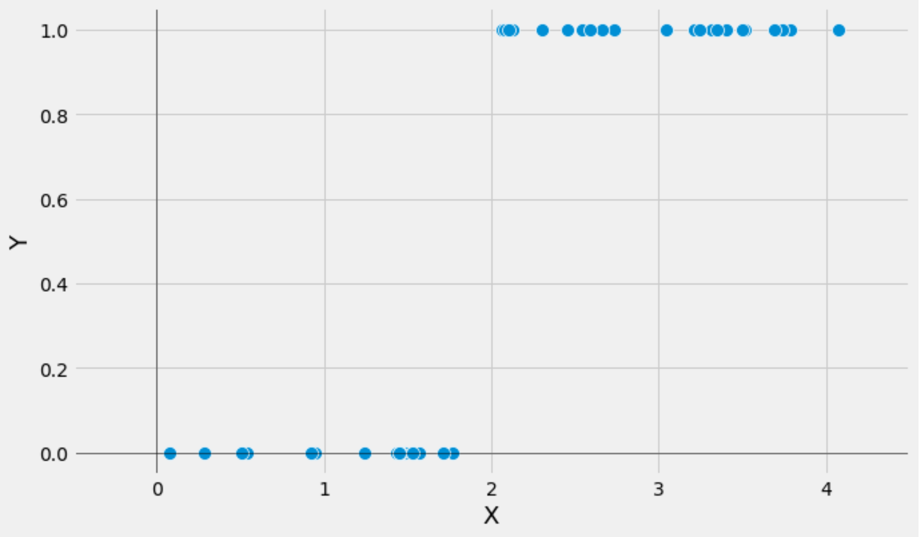
○ **B.** $\frac{1}{4}$

○ C. $\frac{1}{2}$

○ D. $\frac{3}{4}$

○ E. 1

Now, we have a different set of x and y values, as shown below:



(d) [1 Pt] Is this dataset linearly separable?

- ○ **A. Yes**
- ○ B. No
- ○ C. Impossible to tell

(e) [1 Pt] As with the previous dataset, we now wish to fit a logistic regression model with an intercept to the data. Which of the following is *closest* to $P\left(Y=1|x=4\right)$?

- ○ A. 0
- ○ B. $\frac{1}{4}$
- ○ C. $\frac{1}{2}$
- ○ D. $\frac{3}{4}$
- ○ **E. 1**

Evaluating Classifiers [7 Pts]

8. Suppose we trained a logistic regression classifier for some binary classification task. The true labels y and predicted probabilities $P(Y = 1|x)$ are given below.

y	$P(Y = 1\|x)$
1	0.9
1	0.6
0	0.6
0	0.55
1	0.4
0	0.3

In order to make predictions, we need to threshold $P(Y = 1|x)$. We use our thresholds as follows:

$$\hat{Y} = \begin{cases} 1 & P(Y = 1|x) \geq T \\ 0 & P(Y = 1|x) < T \end{cases}$$

We want to decide between three possible thresholds for $P(Y = 1|x)$.

- Model A uses a threshold of $T = 0.25$.
- Model B uses a threshold of $T = 0.5$.
- Model C uses a threshold of $T = 0.75$.

(a) [1 Pt] What is the precision of Model C?

Solution: If 0.75 is the threshold, there is only one true positive, and only one predicted positive. Thus, the precision is $\frac{1}{1} = \boxed{1}$.

(b) [1 Pt] What is the recall of Model C?

Solution: If 0.75 is the threshold, there is only one true positive, while there are three positives in the original dataset. Thus, the recall is $\boxed{\frac{1}{3}}$.

(c) [2 Pts] Suppose our model was built to predict whether or not an individual has some disease. If we predict that they do, we call them in for further testing. Which of the three models is best suited for this task, and why?

Solution: Model A, since we want to minimize false negatives.

(d) [3 Pts] Let T^* be the threshold of a model whose precision and recall are equal. For our given labels and predicted probabilities, what is a possible value of T^*?

Solution: In order for the precision and recall to be equal, we need the number of false positives and false negatives to be equal. Values of T^* in the interval (0.55, 0.6] satisfy this requirement.