Data 100, Midterm 2

Fall 2019

Name: _____

Email: _____@berkeley.edu

Student ID: _____

Exam Room: _____

*All work on this exam is my own **(please sign)***: _____

Instructions:

- This midterm exam consists of **100 points** and must be completed in the **80 minute** time period ending at **9:30**, unless you have accommodations supported by a DSP letter.
- Note that some questions have circular bubbles to select a choice. This means that you should only **select one choice**. Other questions have boxes. This means you should **select all that apply**.
- When selecting your choices, you must **fully shade** in the box/circle. Check marks will likely be mis-graded.
- You may use two cheat sheets each with two sides.
- Please show your work for computation questions as we may award partial credit.

Reference Table

$\exp(x)$	e^x
$\log(x)$	$\log_e(x)$ or $\ln(x)$
Linear regression model	$\hat{y} = f_{\vec{\beta}}(\vec{x}) = \vec{x}^T\vec{\beta}$
Logistic (or sigmoid) function	$\sigma(t) = \frac{1}{1+\exp(-t)}$
Logistic regression model	$\hat{y} = f_{\vec{\beta}}(\vec{x}) = P(Y = 1\|\vec{x}) = \sigma(\vec{x}^T\vec{\beta})$
Squared error loss	$L(y, \hat{y}) = (y - \hat{y})^2$
Absolute error loss	$L(y, \hat{y}) = \|y - \hat{y}\|$
Cross-entropy loss	$L(y, \hat{y}) = -y\log(\hat{y}) - (1 - y)\log(1 - \hat{y})$
Model Bias	$E[f_{\vec{\beta}}(\vec{x})] - g(x)$
Model Variance	$E[(f_{\vec{\beta}}(\vec{x}) - E[f_{\vec{\beta}}(\vec{x})])^2]$

0 Howdy

[0 pts] In LASSO regression, LASSO is an acronym. What does it stand for?

1 PCA

A children's zoo collects data about how much time 1000 visitors spend at each of 8 selected exhibits and stores them in a dataframe `df_zoo`. These exhibits include 6 animals and 2 activities (`train` and `playground`). An example row of `df_zoo` is given below.

	lion	tiger	cheetah	alligator	iguana	turtle	train	playground
0	11.749806	7.276999	7.173714	1.973740	6.653379	0.222565	11.381617	10.057392

(a) [3 Pts] Suppose we **center and scale** `df_zoo` (as we learned about in class) to form the design matrix $\mathbb{X}$. $\mathbb{X}$ has 1000 rows and 8 columns exactly corresponding to the dataframe described above, except that it has been centered and scaled. Suppose we then use SVD to decompose $\mathbb{X}$ into U, Σ, and V^T. Suppose that we want to compute the principal component matrix $\mathcal{P}$, where the 1st column of $\mathcal{P}$ is the 1st principal component, the 2nd column of P is the 2nd principal component, etc. Which of the following expressions are equal to $\mathcal{P}$? Select all that apply.

☐ U ☐ Σ ☐ V^T ☐ $\mathbb{X}$ ☐ $U\mathbb{X}$ ☐ $U\Sigma$ ☐ $\mathbb{X}U$ ☐ $\mathbb{X}\Sigma$ ☐ $\mathbb{X}V$

(b) [2 Pts] How many rows and columns are in $\mathcal{P}$?

rows = ____ # columns = ____

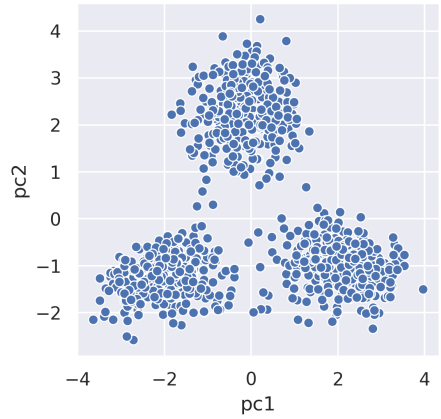
(c) i. [3 Pts] What is the total variance $\mathcal{V}$ of our centered and scaled design matrix $\mathbb{X}$? If there is not enough information provided in the problem statement, write "not enough information."

answer = ____

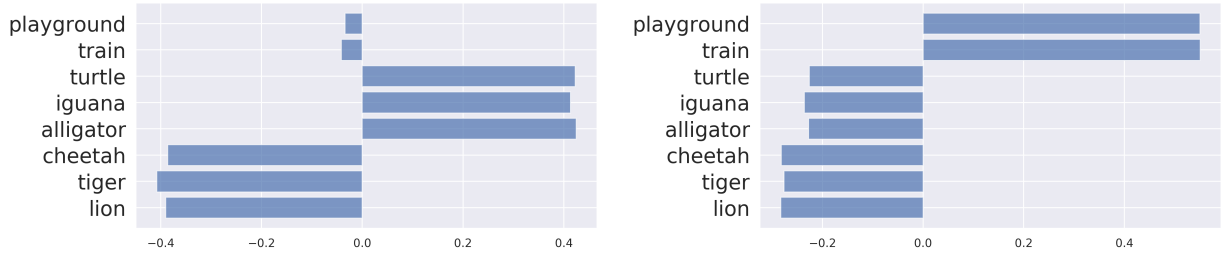
ii. [3 Pts] Suppose our first 6 singular values are $56, 53, 21, 20, 20, 19$. What fraction of the variance is captured by the first two principal components? Do not carry out any arithmetic operations; just give us a numerical expression that could be evaluated into the correct answer. Regardless of your answer to the previous problem, you may assume that you know $\mathcal{V}$, and may give your answer for this problem in terms of $\mathcal{V}$. If there is not enough information, write "not enough information."

answer = ____

(d) [6 Pts] Below is a 2D scatterplot of the first two principal components. We see that there appear to be 3 types of visitors, grouped on the top, bottom-left, and bottom-right.



Below are plots of the first and second rows of V^T.



Use these plots to describe the characteristics of each of the 3 groups in the scatterplot above. Your explanations should only be a sentence or two.

Top group description:

Bottom-left group description:

Bottom-right group description:

2 Linear Regression

Suppose we have a data set of 100 points whose first few rows are shown below, and that we'd like to predict $\vec{y}$ from $\vec{v}$ and $\vec{w}$. Suppose we create a design matrix $\mathbb{X}$ whose first column is $\vec{v}$, second column is $\vec{w}$, and third column $\vec{u}$ is a new feature $u_i = |v_i|$. The resulting model is $\hat{y}_i = \beta_1 v_i + \beta_2 w_i + \beta_3 |v_i|$. **The top row is row 1**, e.g. $y_1 = 4$.

y	v	w
4	-30	1
6	-40	2
5	20	3

(a) [3 Pts] For the data above, suppose we arbitrarily pick $\vec{\beta} = [0.1, 12, 0.2]^T$. What is $\hat{y}_1$?

$\hat{y}_1$ = ☐

(b) [2 Pts] For the data above, let $\vec{e}$ be the residual vector if $\vec{\beta} = [0.1, 12, 0.2]^T$. What is $|e_1|$?

$|e_1|$ = ☐

(c) [3 Pts] For the data above, suppose that $\vec{e} \cdot \vec{e} = 9$. What is the MSE?

MSE = ☐

(d) [3 Pts] Let $\hat{\vec{\beta}}$ be the **exact** parameter vector that minimizes the empirical L_2 risk, where we write this risk as $\mathcal{R}(\vec{\beta}, \mathbb{X}, \vec{y})$. Also, let $\vec{e}$ be the residuals for the optimal parameter vector $\hat{\vec{\beta}}$. Which of the following quantities are guaranteed to be zero?

☐ $\sum e_i$ ☐ The MSE ☐ $\nabla_{\vec{\beta}}(\mathcal{R}(\hat{\vec{\beta}}, \mathbb{X}, \vec{y}))$ ☐ $\vec{e} \cdot \hat{\vec{y}}$ ☐ $\vec{e} \cdot \hat{\vec{\beta}}$ ☐ None of these

(e) [1 ½ Pts] For the data above, the matrix $\mathbb{X}$ has full rank (i.e. no columns are linear combinations of any others). Suppose we compute $\mathcal{Z} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\vec{y}$. What is $\mathcal{Z}$? **Select one and fill in its blank.**

◯ It is a vector of length _____.
◯ It is a matrix with _____ rows and _____ columns.
◯ It does not exist because $|v_i|$ is not differentiable.

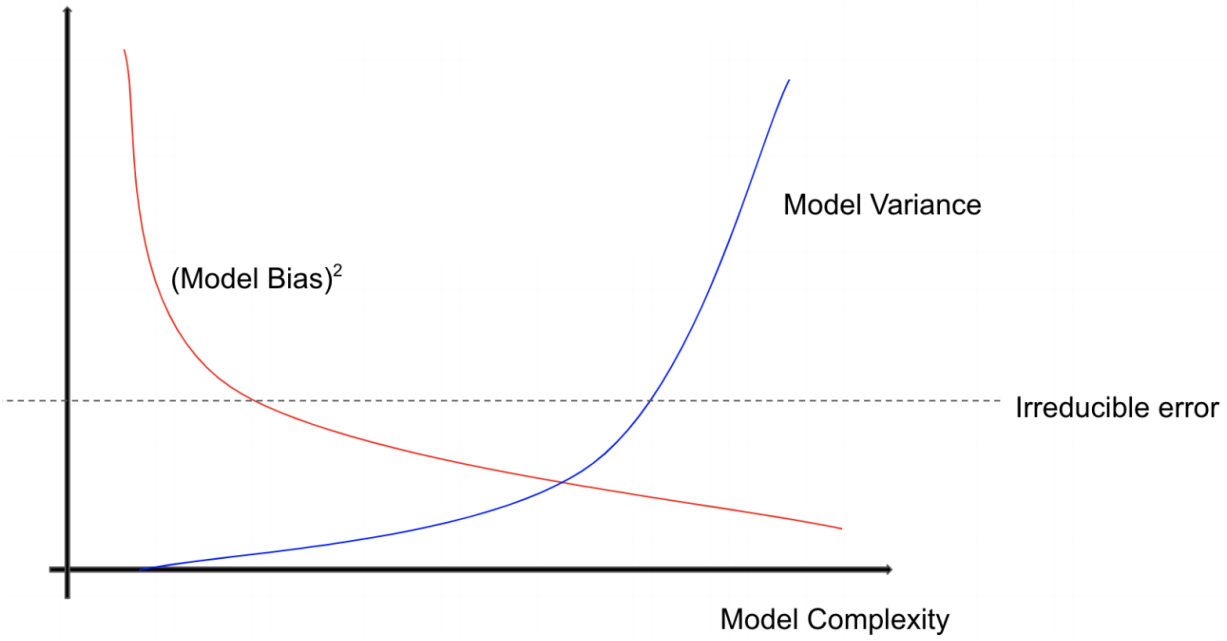
(f) [5 Pts] Let $\vec{\beta}_{ridge}$ be the $\vec{\beta}$ that minimizes the sum of the MSE plus an L_2 regularization term for a positive λ. Let $\vec{e}$ be the residuals for the parameter vector $\vec{\beta}_{ridge}$. Which of the following are true? Recall that $||\vec{\beta}||_2^2$ is the sum of the squares of the components of $\vec{\beta}$ and $\mathcal{R}$ is the empirical L_2 risk defined in (d).

☐ $\sum e_i = 0$
☐ $\nabla_{\vec{\beta}}(\mathcal{R}(\vec{\beta}_{ridge}, \mathbb{X}, \vec{y})) = 0$
☐ $\mathcal{R}(\hat{\vec{\beta}}, \mathbb{X}, \vec{y}) \leq \mathcal{R}(\vec{\beta}_{ridge}, \mathbb{X}, \vec{y})$
☐ $||\vec{\beta}_{ridge}||_2^2 \leq ||\hat{\vec{\beta}}||_2^2$
☐ None of these

3 Bias-Variance Tradeoff

We obtain n data points (n is some large fixed integer) which have been generated from the true model $Y = f(x) + \epsilon$, where ϵ is random noise ($\mathbb{E}[\epsilon] = 0, \text{Var}(\epsilon) = \sigma^2$).

We fit linear models of varying complexity to our data, and plotted the bias, variance, and irreducible error below.



(a) [1 ½ Pts] Sketch the MSE on the above graph. Where does its minimum occur? Draw a star on your MSE plot where the minimum occurs.

(b) [1 Pt] Suppose we control the complexity of the linear models using a Ridge penalty term $\lambda \sum \beta_i^2$. Which of the following is true?

- ◯ The left side of the graph represents small λ.
- ◯ The right side of the graph represents small λ.

(c) [3 Pts] Which of the following can impact our model variance? Select all that apply.

- ☐ The regularization coefficient λ.
- ☐ The choice of features to include in our design matrix.
- ☐ The learning rate α in gradient descent.
- ☐ The size of the training set.

4 Cross Validation

Suppose we have a training dataset of 90 points, and a test set of 30 points, and want to know which λ value is best for a ridge regression model. Our candidate hyperparameters are $\lambda = 0.1$, $\lambda = 1$, and $\lambda = 10$.

(a) [2 ½ Pts] A DS100 student suggests performing 10-fold cross validation to find the optimal λ. Is the choice of 10-fold CV reasonable?

- ◯ Yes.
- ◯ No, since we have 3 candidate hyperparameters we should use 3-fold cross validation.
- ◯ No, since we have 30 test points, we should use 30-fold cross validation.
- ◯ No, CV should never be used for selecting hyperparameters.

(b) Suppose we select the best choice of λ from the three choices available using **3-fold** cross validation. As mentioned in class, we can compute the optimal parameters for a ridge regression model with the expression $\vec{\beta} = (\mathbb{X}^T\mathbb{X} + n\lambda I)^{-1}\mathbb{X}^T\vec{y}$. Assume that we use this closed equation to fit the parameters for our model.

i. [2 Pts] During the entire process of selecting our best λ, how many total times will we evaluate the expression $(\mathbb{X}^T\mathbb{X} + n\lambda I)^{-1}\mathbb{X}^T\vec{y}$?

◯ 1 ◯ 2 ◯ 3 ◯ 6 ◯ 9 ◯ 30 ◯ 60 ◯ 90 ◯ 270

ii. [2 Pts] How many rows will be in $\mathbb{X}$ each time this expression is evaluated?

◯ 1 ◯ 2 ◯ 3 ◯ 6 ◯ 9 ◯ 30 ◯ 60 ◯ 90 ◯ 120
◯ It will vary each time. ◯ Not enough information.

(c) As in the previous part, suppose we want to select the best λ from the three choices above using **3-fold** cross validation. To evaluate the MSE for a given $\vec{\beta}$, we use the sum of squares: $||\vec{y} - \mathbb{X}\vec{\beta}||_2^2$. Reminder that this expression is just another way of writing $\sum(\vec{y_i} - \vec{x_i}^T\vec{\beta})^2$.

i. [2 Pts] During the entire process of selecting our best λ, how many times will this expression get evaluated?

◯ 1 ◯ 2 ◯ 3 ◯ 6 ◯ 9 ◯ 30 ◯ 60 ◯ 90

ii. [2 Pts] How many rows will be in $\mathbb{X}$ each time this expression is evaluated?

◯ 1 ◯ 2 ◯ 3 ◯ 6 ◯ 9 ◯ 30 ◯ 60 ◯ 90 ◯ 120
◯ It will vary each time. ◯ Not enough information.

5 Gradient Descent

(a) [3 Pts] The learning rate can *potentially* affect which of the following? Select all that apply. Assume nothing about the function being minimized other than that its gradient exists. You may assume the learning rate is positive.

- ☐ The speed at which we converge to a minimum.
- ☐ Whether gradient descent converges.
- ☐ The direction in which the step is taken.
- ☐ Whether gradient descent converges to a local minimum or a global minimum.

(b) [3 Pts] Suppose we run gradient descent with a fixed learning rate of $\alpha = 0.1$ to minimize the 2D function $f(x, y) = 5 + x^2 + y^2 + 5xy$.

The gradient of this function is

$$\nabla_{x,y} f(x, y) = \begin{bmatrix} 2x + 5y \\ 2y + 5x \end{bmatrix}$$

If our starting guess is $x^{(0)} = 1, y^{(0)} = 2$, what will be our next guess $x^{(1)}, y^{(1)}$?

$x^{(1)} =$ $\qquad$ $y^{(1)} =$

(c) [2 Pts] Suppose we are performing gradient descent to minimize the empirical risk of a linear regression model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2$ on a dataset with 100 observations. Let $\mathcal{D}$ be the number of components in the gradient, e.g. $\mathcal{D} = 2$ for the equation in part b. What is $\mathcal{D}$ for the gradient used to optimize this linear regression model?

◯ 2 ◯ 3 ◯ 4 ◯ 8 ◯ 100 ◯ 200 ◯ 300 ◯ 400 ◯ 800

6 One Hot Encoding and Feature Engineering

A Canadian study of workers in the 1980s collected the following information:

- `wage` (hourly in dollars)
- `edu` (years)
- `job_type` (1 for blue collar, 2 for white collar, and 3 for managerial)

A data scientist fitted a model with `wage` as the response, and the other two variables as features (`job_type` was one-hot encoded). The resulting fitted model was $\hat{y} = \vec{x} \cdot \vec{\hat{\beta}}$, where $\vec{\hat{\beta}} = \begin{bmatrix} -8 & 3 & 6 & -3 \end{bmatrix}^T$, i.e.

$$\hat{y} = -8 + 3x_{edu} + 6x_m - 3x_b,$$

where y is the hourly wage, x_{edu} is years of education, and the other two variables are the dummies for managerial and blue collar workers, respectively.

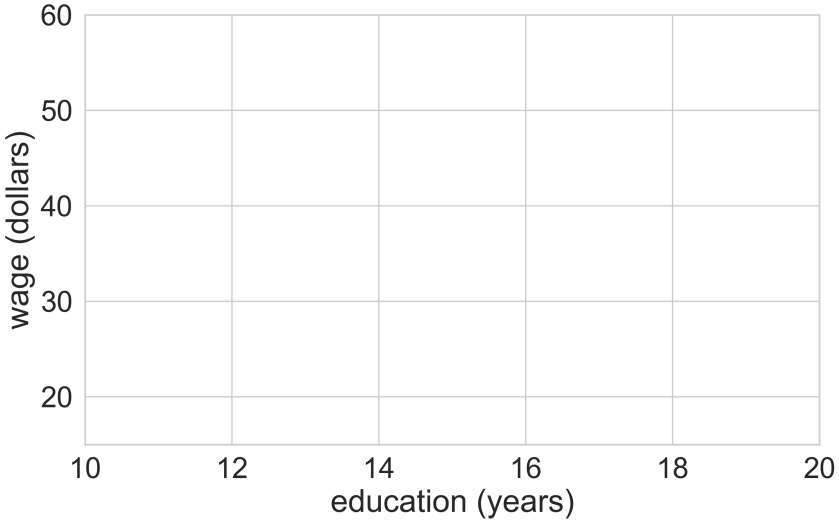
(a) [2 Pts] For a blue collar worker with 10 years of education, what is the predicted value of `wage` (the predicted hourly wage) according to our model?

`wage` = ☐

(b) [2 Pts] For a white collar worker with 10 years of education, what is the predicted value of `wage` according to our model?

`wage` = ☐

(c) [6 Pts] Sketch the fitted model on the graph below. Hint: What you did in parts (a) and (b) is useful here. When grading we will only look at y-values for $x = 10$ and $x = 20$, so don't worry about exact values other than these. Don't worry about exact shape.



(d) [5 Pts] The first four rows of the original data frame appear below on the left.

wage	edu	job.type
15	10	1
28	14	2
20	12	1
35	16	3

Create the design matrix $\mathbb{X}$ used to fit the model on the previous page by filling in the table below. Put the variable name in the first row and fill the remaining 4 rows with the corresponding data. You may not need all columns. Use the top row to name your columns.

(e) [6 Pts] Suppose we believe that the slope of the relationship between education level and wage is different for each of our 3 job types, e.g. perhaps white collar workers have salaries that are 2x their years of education, but blue collar workers only 1.5x. Create a design matrix below that will yield a model with different slopes and y-intercepts for each job type. Use the top row to name your columns. You may not need all columns.

Warning: This is a very challenging problem. Move on if you're stuck.

7 Logistic Regression

Suppose we want to build a classifier to predict whether a person survived the sinking of the Titanic. The first 5 rows of our dataset are given below.

	age	survived	female
0	22.0	0	0
1	38.0	1	1
2	26.0	1	1
3	35.0	1	1
4	35.0	0	0

(a) For a given classifier, suppose the first 10 predictions of our classifier and 10 true observations are as follows:

prediction	1	1	1	1	1	0	1	1	1	1
true label	0	1	1	1	0	0	0	1	1	1

i. [1 Pt] What is the accuracy of our classifier on these 10 predictions?

ii. [1 ½ Pts] What is the precision on these 10 predictions?

iii. [1 ½ Pts] What is the recall on these 10 predictions?

(b) [4 ½ Pts] In general (not just for the Titanic model), if we increase the threshold for a classification model, what of the following can happen to our precision, recall, and accuracy? We have not included the option "X can stay the same", because this is trivially true (e.g. if we increase the threshold by some tiny number, it will have no effect).

- ☐ Precision can increase.
- ☐ Precision can decrease.
- ☐ Recall can increase.
- ☐ Recall can decrease.
- ☐ Accuracy can increase.
- ☐ Accuracy can decrease.

For convenience, we repeat the figure from the previous page below.

	age	survived	female
0	22.0	0	0
1	38.0	1	1
2	26.0	1	1
3	35.0	1	1
4	35.0	0	0

(c) Suppose after training our model we get $\vec{\beta} = \begin{bmatrix} -1.2 & -0.005 & 2.5 \end{bmatrix}^T$, where -1.2 is an intercept term, -0.005 is the parameter corresponding to passenger's age, and 2.5 is the parameter corresponding to sex.

i. [3 Pts] Consider Sīlānah Iskandar Nāsīf Abī Dāghir Yazbak, a 20 year old female. What chance did she have to survive the sinking of the Titanic according to our model? Give your answer as a probability in terms of σ. If there is not enough information, write "not enough information".

$P(Y = 1 | \text{age} = 20, \text{female} = 1) =$ [ ]

ii. [3 Pts] Sīlānah Iskandar Nāsīf Abī Dāghir Yazbak actually survived. What is the cross-entropy loss for our prediction in part i? If there is not enough information, write "not enough information."

cross entropy loss = [ ]

iii. [6 Pts] Let m be the odds of a given male passenger's survival according to our model, i.e. if the passenger had an 80% chance of survival, m would be 4, since their odds of survival are $0.8/0.2 = 4$. It turns out we can compute f, the odds of survival for a female of the same age, even if we don't know the age of the two people. What is this relationship? *Hint: How are the odds related to* $t = \vec{x}^T\vec{\beta}$ *for a given observation?*

Warning: This is a very challenging problem. Move on if you're stuck.

$f =$ [ ]