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?

Solution: Least Absolute Shrinkage and Selection Operator

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 ☐ $\mathbb{X}$ ☐ $U\mathbb{X}$ ☐ $U\Sigma$ ☐ $\mathbb{X}U$ ☐ $\mathbb{X}\Sigma$ ☐ $\mathbb{X}V$

Solution: Recall that the SVD gives $\mathbb{X} = U\Sigma V^T$ and $\mathcal{P} = U\Sigma = \mathbb{X}V$.

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

rows = ____ # columns = ____

Solution: 1000 rows, 8 columns. 1000 data points, 8 principal components per data point.

(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 = ____

Solution: $\mathcal{V} = 8$

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."

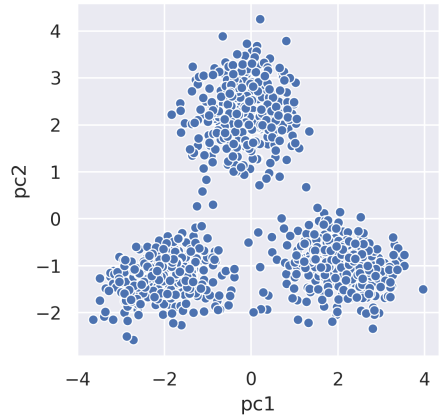
answer =

Solution: $\frac{56^2+53^2}{1000\mathcal{V}}$
The variance described by the first 2 PCs is

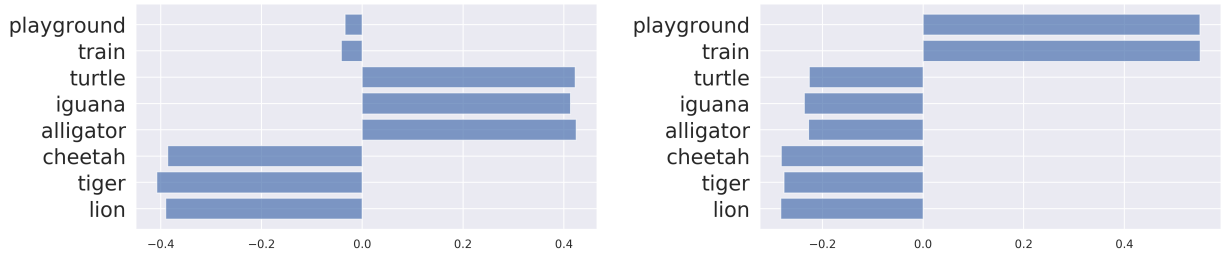
$$\frac{56^2 + 53^2}{1000}$$

and the total variance is $\mathcal{V}$.

(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.

Solution: The group with positive pc2 values and pc1 values around 0 (top group) seems to represent a group of visitors who spend a lot of time at the activites (`train` and `playground`). The group with negative pc1 values and negative pc2 values (bottom left group) seems to represent a group of visitors who spend a lot of time at the mammal exhibits (`cheetah`, `tiger`, and `lion`). The group with positive pc1 values and negative pc2 values (bottom right group) seems to represent a group of visitors who spend a lot of time at the reptile exhibits (`turtle`, `iguana`, and `alligator`).

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 $\hat{y}_1$?

$\hat{y}_1$ = ____

Solution:

$$\hat{y}_1 = 0.1 \cdot (-30) + 12 \cdot 1 + 0.2 \cdot |-30| = -3 + 12 + 6 = \boxed{15}$$

(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 $|e_1|$?

$|e_1|$ = ____

Solution:

$$|e_1| = |y_1 - \hat{y}_1| = |4 - 15| = \boxed{11}$$

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

MSE = ____

Solution: Note, $\vec{e} \cdot \vec{e} = ||\vec{e}||_2^2 = \sum_{i=1}^n e_i^2$. Then, since MSE $= \frac{1}{n} \sum_{i=1}^n e_i^2$, the MSE is $\frac{1}{100} \cdot 9 = \boxed{\frac{9}{100}}$.

(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

Solution:

i. The first option is not correct – this is only guaranteed when we have an intercept term, which our model does not.

ii. This is also not correct – this would require all our residuals to equal 0 which is usually not the case.

iii. This is correct – at the point at which our empirical risk $\mathcal{R}$ is minimized, the gradient of $\mathcal{R}$ with respect to $\vec{\beta}$ is guaranteed to be 0. *Note: if we use a numerical technique, like gradient descent, the value of our gradient for our estimated value of $\vec{\beta}$ may not be exactly 0, but in this question we're dealing with an* ***exact*** *value for $\hat{\vec{\beta}}$.*

iv. This is also correct – we know that $\vec{e}$ is orthogonal to the span of $\mathbb{X}$. Since $\hat{y} = \mathbb{X}\hat{\vec{\beta}}$, we know that $\hat{y} \in \text{span}\{\mathbb{X}\}$, and thus $\hat{y}$ must also be orthogonal to the residuals.

v. This option is not correct, since there's no direct relationship between $\vec{e}$ and $\hat{\vec{\beta}}$ that doesn't involve $\hat{y}$ or $\mathbb{X}$.

(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.

Solution: We know that $\hat{\vec{\beta}} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\vec{y}$. Thus, this quantity must be a vector, with length equal to the number of features, which in this case is 3.

(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

Solution:

i. This is not correct —- we still don't have an intercept column, and even if we did, we know that the predictions on our training set for $\vec{\beta}_{ridge}$ are worse than those of $\vec{\hat{\beta}}$, and so the residuals (and hence their sum) is larger for the ridge solution than it is for the non-regularized solution.

ii. This is also not correct — $\vec{\hat{\beta}}$ is the unique value of $\vec{\beta}$ that minimizes $\mathcal{R}$. $\vec{\beta}_{ridge}$ doesn't minimize $\mathcal{R}$ (it instead minimizes a regularized risk), and so the gradient of $\mathcal{R}$ is not equal to 0 when evaluated at $\vec{\beta}_{ridge}$.

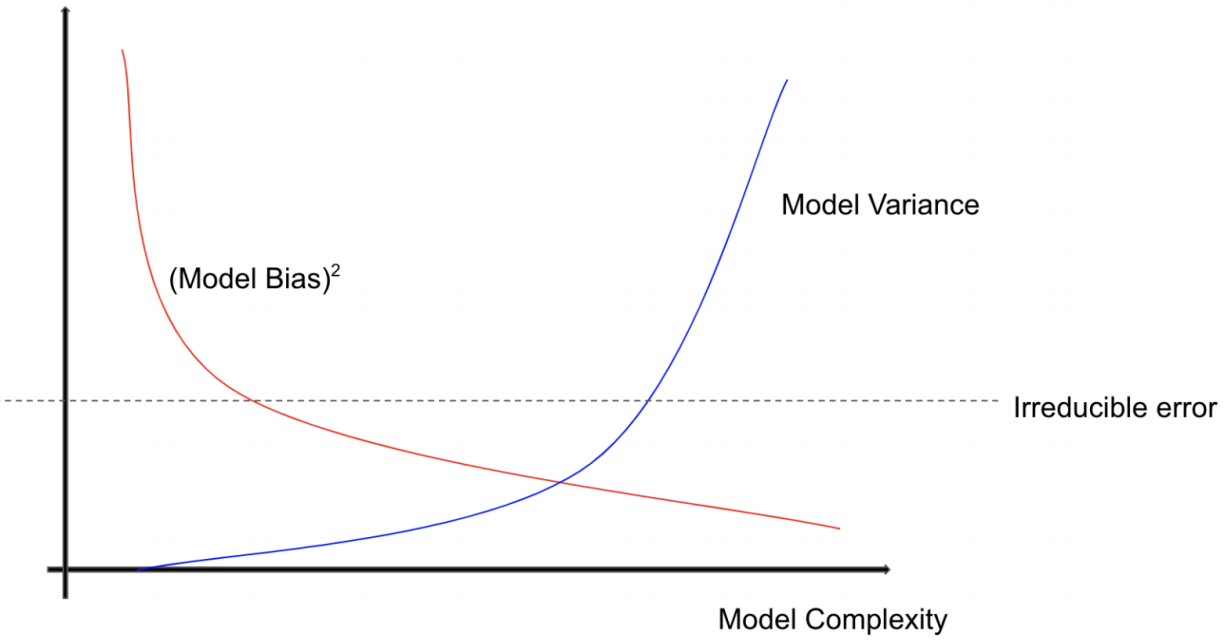
iii. This is true — regularizing our model makes our predictions worse on our training set (in hopes that it generalizes our model to better fit unseen data). As a result, the empirical risk of our regularized model is greater than (or equal to) the empirical risk of our unregularized model.

iv. This is also true — the objective function for ridge regression includes a penalty on the L_2 norm of $\vec{\beta}$, in order to decrease the norm of $\vec{\beta}$. This option follows from that principle.

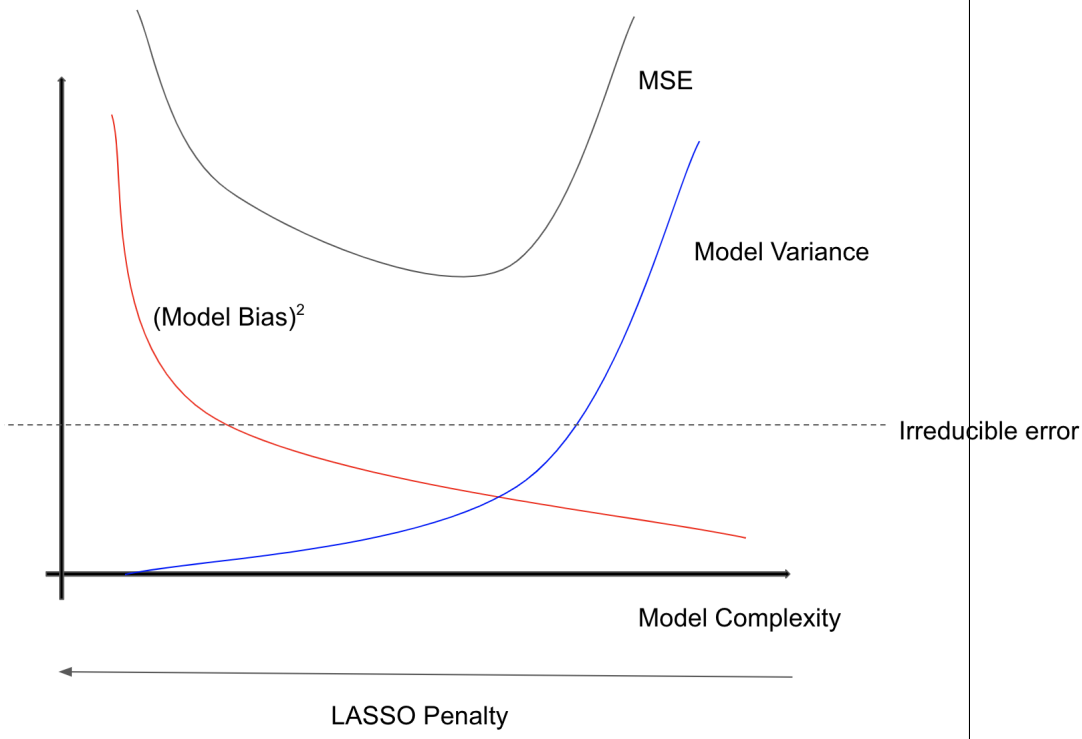
3 Bias-Variance Tradeoff

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.



Solution:
MSE = (Model Bias)^2 + Model Variance + Irreducible error

Irreducible error is also known as σ^2, i.e. the variance of the noise term ϵ.

An approximately U-shaped curve should be drawn where each point on the curve is the sum of the three curves/lines. The minimum occurs at the minimum of this drawn curve.

(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.**

Solution: A smaller λ value means higher model complexity. Remember that a zero λ value means a model with no regularization.

(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.**

Solution:

- ☐ **A higher λ value means more regularization which reduces model variance.**
- ☐ **Including a large number of uninformative features may lead to overfitting, which in turn with increase the model variance.**
- ☐ The learning rate α in gradient descent won't impact the model's variance bias tradeoff, it is simply a numerical method that is used to fit the model to data.
- ☐ **Generally, a larger training set will reduce model variance.**

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.

Solution:

i. With a (relatively small) dataset of 90 points, 10-fold CV is reasonable. We will be computing (10 folds) * (3 choices of λ) = 30 validation errors, each of which is obtained by training a ridge regression model on some portion of the 90 training data points and testing on the remainder of the 90 points we didn't use for training. This answer must also be the solution because the other statements are not correct/logical.

ii. In general, there is no rule saying that we have to use the same number of folds as there are choices of hyperparameters. The number of folds is completely separate from the number of hyperparameters.

iii. The test data is not considered at all for CV, so there is no relationship between any property of the test data and the number of folds in CV.

iv. CV is the only method taught in this class for selecting hyperparameters, so this statement is incorrect.

(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.

Solution:

i. Note that computing $(\mathbb{X}^T\mathbb{X} + n\lambda I)^{-1}\mathbb{X}^T\vec{y}$ is training (or "fitting") a model with data matrix $\mathbb{X}$ and a regularization parameter λ. In CV, we train a model on each fold for each value of λ. Thus, we have to train (3 folds) * (3 choices of λ) = 9 models.

ii. Since we are doing 3-fold CV, we split our training data into 3 parts of equal size. For each fold, 2 of these parts will be used for training the model and 1 part will be used for validation. Since our training data has 90 points, each part will have 30 points. Since 2 parts are used to train each model, the $\mathbb{X}$ matrix will have 60 points and therefore 60 rows.

(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.

Solution:

i. Note that computing the MSE of a model on some data is evaluating the model's error on that data. In CV, we are interested in knowing each model's error on each fold. Remember that we have a different model for each of 3 choices of λ and that we have 3 folds. Thus, we will be computing the MSE (3 choices of λ) * (3 folds) = 9 times.

ii. Since we are doing 3-fold CV, we split our training data into 3 parts of equal size. For each fold, 2 of these parts will be used for training the model and 1 part will be used for validation. Validation is the process of computing the error of a model on a particular fold. Since our training data has 90 points, each part will have 30 points. Since 1 part is used for validation, the $\mathbb{X}$ matrix will have 30 points and therefore 30 rows.

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 $x^{(1)}, y^{(1)}$?

$x^{(1)}$ = _____ $y^{(1)}$ = _____

Solution: The gradient is = [2*1 + 5*2, 2*2 + 5*1] = [12, 9] so next guess is [1, 2] - 0.1 * [12, 9] = -0.2, 1.1

(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` = ☐

Solution: For this worker, we have that $x_{edu} = 10$, $x_m = 0$, and $x_b = 1$. When we plug these values in to the fitted model, we get:

$$\hat{y} = -8 + 3 \times 10 + 6 \times 0 - 3 \times 1 = 19$$

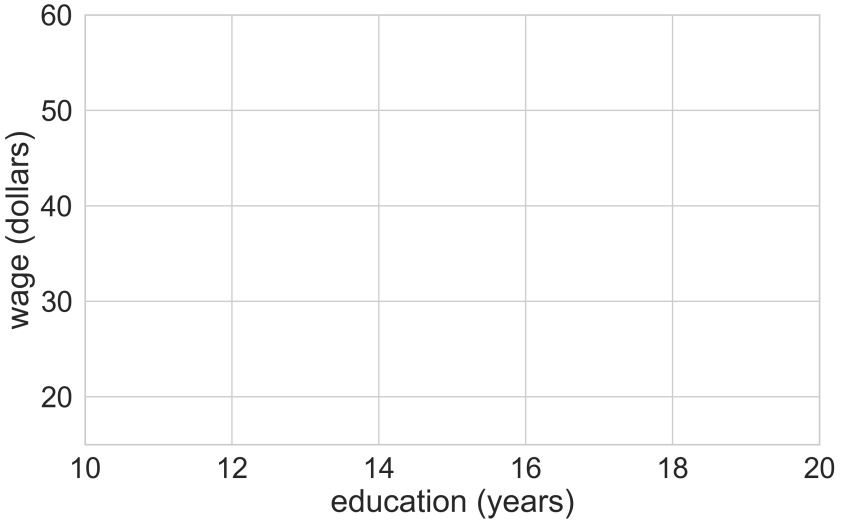
(b) [2 Pts] For a white collar worker with 10 years of education, what is the predicted value of `wage` according to our model?

`wage` = ☐

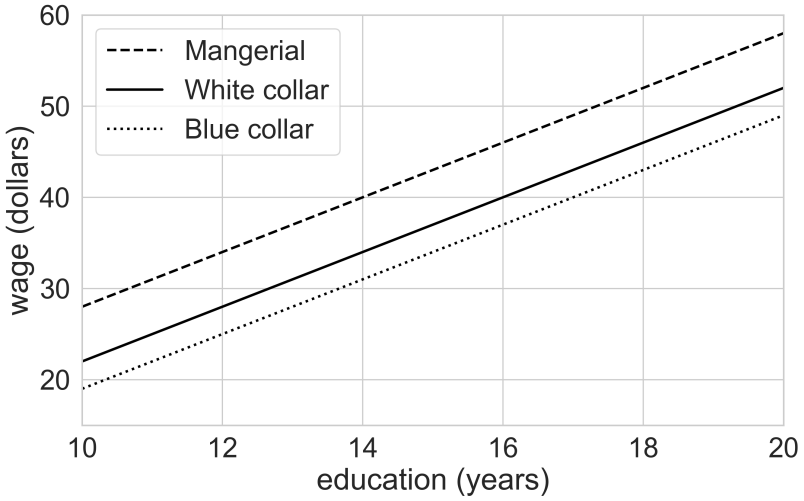
Solution: For this worker, we have that $x_{edu} = 10$, $x_m = 0$ and $x_b = 0$. When we plug these values in to the fitted model, we get:

$$\hat{y} = -8 + 3 \times 10 + 6 \times 0 - 3 \times 0 = 22$$

(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.



Solution: The fitted model yields three parallel lines. The slope of the lines is 30. The intercept depends on job type. The intercept is -8 for the white collar workers, $-8-3=-11$ for blue collar workers, and $-8+6=-2$ for managerial workers. We can use the points determined in parts (a) and (b) to draw the lines. Specifically, for a white collar worker we found that the line goes through the point $(10, 22)$. It must also go through the point $(20, 52)$. Similarly, the blue collar line goes through the points $(10, 19)$ and $(20, 49)$ and the managerial worker line goes through the points $(10, 28)$ and $20, 58)$. The figure below shows these three lines.



(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.

Solution: The model that was fitted has a constant (bias) term and dummy variables for two of the three categories (managerial and blue collar workers). The dummy variable x_m is 1 for observations where job.type is 3 (i.e. managerial).

bias	x_{edu}	x_m	x_b
1	10	0	1
1	14	0	0
1	12	0	1
1	16	1	0

(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.

Solution: To allow the slopes to be different for the different job types, we augment to design from the previous problem to include variables that allow education to have a different slope. We can do this by adding two additional features that contain the education for subgroups of the data as shown below.

bias	x_{edu}	x_m	x_b	$x_{edu,m}$	$x_{edu,b}$
1	10	0	1	0	10
1	14	0	0	0	0
1	12	0	1	0	12
1	16	1	0	16	0

Now, our model looks like

$$\hat{y} = \beta_0 + \beta_1 \cdot x_{edu} + \beta_2 \cdot x_m + \beta_3 \cdot x_b + \beta_4 \cdot x_{edu_m} + \beta_5 \cdot x_{edu_b}$$

Another approach to encapsulating these three separate models (one for each job type) into one model is to create three pairs of education levels and biases for each of the job types.

x_{edu_b} and $bias_b$ will only have values in that column if the original datapoint was of `job_type` 1. Otherwise, both values in these columns will be 0.

x_{edu_b}	$bias_b$	x_{edu_w}	$bias_w$	x_{edu_m}	$bias_m$
10	1	0	0	0	0
0	0	14	1	0	0
12	1	0	0	0	0
0	0	0	0	16	1

Now, our model looks like

$$\hat{y} = \beta_1 \cdot x_{edu_b} + \beta_2 \cdot bias_b + \beta_3 \cdot x_{edu_w} + \beta_4 \cdot bias_w + \beta_5 \cdot x_{edu_m} + \beta_6 \cdot bias_m$$

For a given observation, if the original `job_type` value was 1 (i.e. the person was a blue collar worker), then all features other than x_{edu_b} and $bias_b$ are set to 0, so we have $\hat{y} = \beta_1 \cdot x_{edu_b} + \beta_2 \cdot bias_b$.

The same principle applies to the other two job types as well.

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?

Solution: 7 of our predictions were correct, out of 10 total. Thus, our accuracy is $\boxed{\frac{7}{10}}$.

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

Solution: The number of true positives, TP, is 6. The number of false positives, FP, is 3. Then, the precision is $\frac{TP}{TP+FP} = \frac{6}{9} = \boxed{\frac{2}{3}}$.

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

Solution: From the solution to the previous part, we know that $TP = 6$. The number of false negatives, FN, here is 0 (we only predicted 0 once, and in that case the true value was actually 0). Thus, the recall is $\frac{TP}{TP+FN} = \frac{6}{6+0} = \boxed{1}$.

(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.**

Solution: As we increase our classification threshold, the number of false positives decreases, but the number of false negatives (i.e. undetected points) increases. As a result, our precision increases (more of the points we say are positive will actually be positive), but our recall decreases (there will be more points that are actually positive that we don't detect). However, in some cases precision can also decrease, when increasing a threshold lowers the number of true positives but keeps the number of true negatives the same. As seen in lecture, accuracy may increase or decrease – there typically exists an optimal threshold that maximizes accuracy, and if we increase or decrease our threshold from that point, accuracy decreases.

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".

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

Solution: To be explicit, our observation vector here is $\vec{x} = [1, 20, 1]^T$. Then, $\vec{x}^T\vec{\beta} = 1(-1.2) + 20(-0.005) + 1(2.5) = 1.2$.
Then, $P(Y = 1|\vec{x}) = \sigma(\vec{x}^T\vec{\beta}) = \boxed{\sigma(1.2)}$.

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 = $\boxed{\quad}$

Solution: Here, $y = 1$ and $\hat{y} = \sigma(1.2)$. Then,

$$\text{cross entropy loss} = -y\log(\hat{y}) - (1-y)\log(1-\hat{y}) = \boxed{-\log(\sigma(1.2))}$$

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.

$f =$ [ ]

Solution: We start by finding a simple relationship between the odds and $\sigma(t) = \sigma(\vec{x}^T\vec{\beta})$.
In logistic regression $p = \sigma(\vec{x}^T\vec{\beta})$.
The odds are defined as odds $= \frac{p}{1-p}$. Substituting in, we have:

$$\text{odds} = \frac{\frac{1}{1+e^{-t}}}{1 - \frac{1}{1+e^{-t}}}$$

$$\text{odds} = \frac{\frac{1}{1+e^{-t}}}{\frac{1+e^{-t}-1}{1+e^{-t}}}$$

$$\text{odds} = \frac{1}{e^{-t}}$$

$$\text{odds} = e^t$$

Thus, we have that odds $= e^{\vec{x}^T\vec{\beta}}$.
From here, the problem is relatively straightforward, as we have that

$$f/m = \frac{e^{-1.2-0.005\text{age}+2.5}}{e^{-1.2-0.005\text{age}}}$$

$$f/m = e^{2.5}$$

Solution: Recall, the assumption upon which we derived the logistic model was that the log-odds of our probability was linear. That is,

$$\log \frac{P(Y=1|\vec{x})}{P(Y=0|\vec{x})} = \vec{x}^T\vec{\beta}$$

Exponentiating both sides and substituting in the model and provided weights:

$$\frac{P(Y=1|\vec{x})}{P(Y=0|\vec{x})} = e^{\vec{x}^T\vec{\beta}} = e^{-1.2-0.005\cdot(\text{age})+2.5\cdot(\text{sex})}$$

We're told to consider the odds for a fixed age. So,

$$m = e^{-1.2-0.005\cdot(\text{age})+2.5\cdot 0} = e^{-1.2-0.005\cdot(\text{age})}$$

$$f = e^{-1.2-0.005\cdot(\text{age})+2.5\cdot 1} = e^{-1.2-0.005\cdot(\text{age})} \cdot e^{2.5}$$

Thus, we can say that $\boxed{f = m \cdot e^{2.5}}$.