# Data 100, Final

Summer 2020

## Note:

- This exam was administered over two days in two different formats.
- Day 1 consisted of 12 questions, and was worth a total of 75 points. It appears first in this document.
- Day 2 consisted of 9 questions, and was worth a total of 75 points. It appears after Day 1 in this document.

**1. (6 points)**

**(a) (1 pt)** True or False: Decision trees will always have 100% accuracy on the training set.

○ True

○ False

**(b) (1 pt)** True or False: On a dataset that is linearly separable, decision trees will always return a linear decision boundary.

○ True

○ False

**(c) (2 pt)** Which of the following are reasons that random forests are less prone to overfitting? Select all that apply.

□ Random forests always have access to more training data than a single decision tree.

□ Random forests always institute a node depth limit.

□ A random forest is an example of an ensemble method that involves one or more decision trees.

□ Random forests always have a higher testing accuracy than a single decision tree.

□ None of the above

**(d) (2 pt)** Let's say we have a dataset with 3 classes, labeled A, B, and C, from which we want to build a classifier using a decision tree. The following table displays how many points are in each class:

| Class | Number of Points |
|---|---:|
| A | 4 |
| B | 4 |
| C | 2 |

Recall from lecture that we want to minimize the weighted entropy of our splits. What is the weighted entropy of the following split?

**Node 1:** 4 in class A, 0 in class B, 2 in class C; **Node 2:** 0 in class A, 4 in class B, 0 in class C.

○ 0

○ 10

○ $-\frac{1}{3}\log_2\frac{1}{3}-\frac{2}{3}\log_2\frac{2}{3}$

○ $\frac{2}{5}(-\frac{1}{3}\log_2\frac{1}{3}-\frac{2}{3}\log_2\frac{2}{3})$

○ $\frac{3}{5}(-\frac{1}{3}\log_2\frac{1}{3}-\frac{2}{3}\log_2\frac{2}{3})$

○ None of the above

**2. (3 points)**

The following are questions on K-means and agglomerative clustering.

**(a) (1 pt)** Suppose we run average (average linkage) agglomerative clustering 10 times from start to finish on the same dataset. Assume that tie-breaking is done randomly, but that no two points have the same distance from each other. This will __________ result in the same clustering of points each of the 10 times.

○ Always

○ Sometimes

○ Never

**(b) (1 pt)** Suppose that after an iteration of the K-means clustering algorithm, the cluster centers change locations, but the labels assigned to each point do not change. In such a situation, it is ___________ true that the algorithm has converged because the following iteration will not change the labels assigned to each point.

○ Always

○ Sometimes

○ Never

**(c) (1 pt)** A data point has a negative silhouette score if it is on average closer to the points in its own cluster than the points in the closest neighboring cluster.

○ True

○ False

**3. (7 points)**

Suppose you have data on Berkeley course enrollment for Summer 2020, including the following information for each student:

- last 4 digits of SID number
- course enrolled in (assume each student is only enrolled in one course)
- academic level
- major
- cumulative GPA

**(a) (1 pt)** Select all suitable visualizations for comparing the number of Statistics majors enrolled in each summer course.

□ Scatterplot

□ Boxplot

□ Barplot

□ Histogram

□ Violin plot

□ None of the above

**(b) (1 pt)** You create a boxplot to visualize the distribution of cumulative GPA for each academic level, which is stored numerically (1-5, where 1 = freshman to 5 = graduate).

A boxplot is suitable because academic level is a ____ variable and cumulative GPA is a ____ variable.

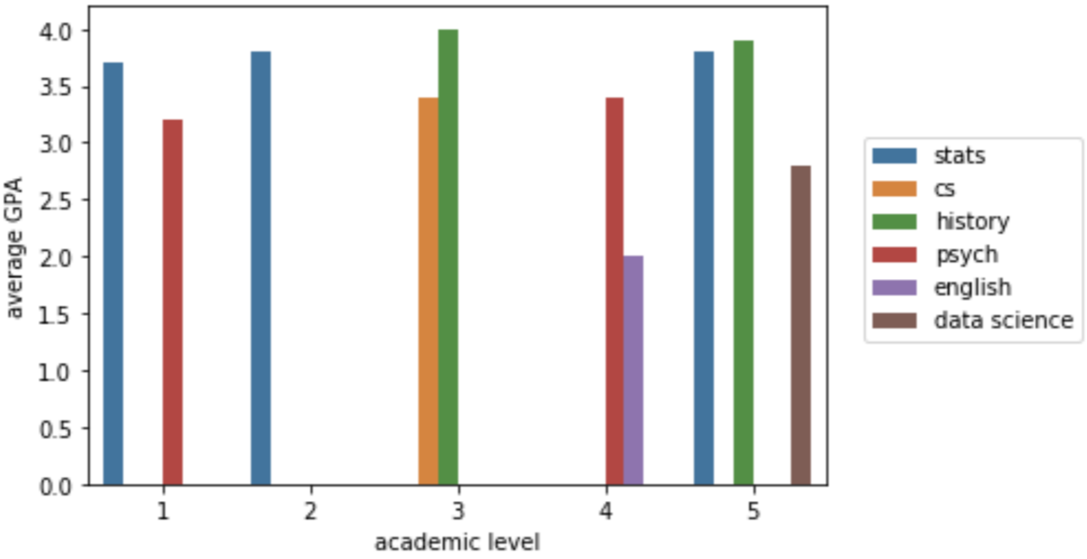
○ Quantitative; quantitative

○ Qualitative; quantitative

○ Quantitative; qualitative

○ Qualitative; qualitative

(c) **(1 pt)** You create the plot below using the first 10 rows of the dataset. Which of the following combinations of encodings is being used in your plot?



○ x, y

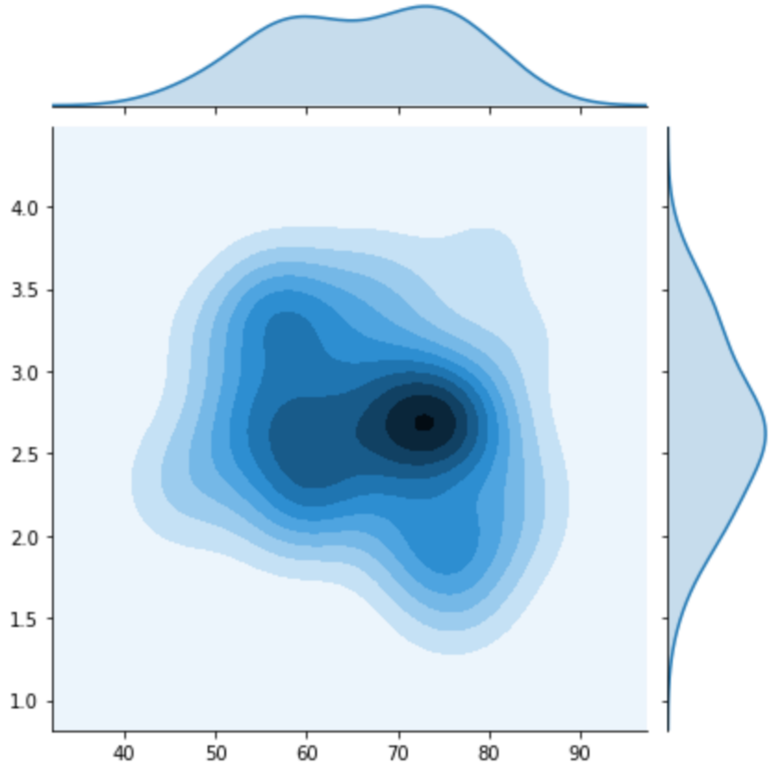
○ x, y, height

○ x, y, area

○ x, y, color

○ x, y, area, color

(d) **(2 pt)** Consider the following contour plot, which depicts a student's final exam score in Data 100 on the $x$-axis and their cumulative GPA on the $y$-axis.



Which of the following conclusions can be drawn from the above plot? Select all that apply.

□ Final exam scores are bimodal

□ Final exam scores are unimodal

□ GPA is bimodal

□ GPA is unimodal

□ The correlation between final exam scores and GPA is greater than 0.3

□ None of the above

**(e) (2 pt)** The triangular kernel is an alternative to the Gaussian and Boxcar kernels that were introduced in Lab 5. The triangular kernel centered at 0 with $\alpha = 1$ is defined by:

$$T(x) = \begin{cases} 1 - |x| & \text{if } |x| \leq 1 \\ 0 & \text{else} \end{cases}$$

Which of the following would be the equation of a triangular kernel centered at $z$ with $\alpha = 1$?

A:

$$T_\alpha(x, z) = \begin{cases} |z - 1| - |x| & \text{if } |x - z| \leq 1 \\ 0 & \text{else} \end{cases}$$

B:

$$T_\alpha(x, z) = \begin{cases} 1 - |x - z| & \text{if } |x - z| \leq 1 \\ 0 & \text{else} \end{cases}$$

C:

$$T_\alpha(x, z) = \begin{cases} 1 - |x - z| & \text{if } |x| \leq |1 - z| \\ 0 & \text{else} \end{cases}$$

D:

$$T_\alpha(x, z) = \begin{cases} |x - z| & \text{if } |x| \leq |1 - z| \\ 0 & \text{else} \end{cases}$$

○ A

○ B

○ C

○ D

**4. (10 points)**

Suppose we are trying to fit a linear model $\hat{\mathbb{Y}} = \mathbb{X}\theta$. Assume that our design matrix $\mathbb{X}$ contains $n$ rows and $p + 1$ columns, the first of which contains the value 1 for each observation.

Let $\dot{\mathbb{X}}$ be the mean-centered version of $\mathbb{X}$ (that is, each column **except the intercept column** is scaled to have a mean of 0), and let $\dot{\mathbb{Y}} = [\dot{y}_1, \dot{y}_2, ..., \dot{y}_n]$ be the mean-centered version of $\mathbb{Y}$, our true response vector. Assume that $p \geq 5$.

In the first few subparts of this question, we will explore the models that result from un-regularized linear regression, depending on whether or not we center our data.

$$\hat{\beta} = \arg\min_\theta \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \mathbb{X}_i^T \theta\right)^2$$

$$\hat{\gamma} = \arg\min_\theta \frac{1}{n} \sum_{i=1}^{n} \left(\dot{y}_i - \dot{\mathbb{X}}_i^T \theta\right)^2$$

**(a) (3 pt)** Assuming that both $\mathbb{X}$ and $\dot{\mathbb{X}}$ have full column rank, which of the following **are guaranteed** to be true?

□ $\hat{\beta}_0 = \hat{\gamma}_0$

□ $\hat{\beta}_1 = \hat{\gamma}_1$

□ $\hat{\beta}_2 = \hat{\gamma}_2$

□ $\hat{\beta}_p = \hat{\gamma}_p$

□ $\hat{\gamma}_0 = 0$

□ The MSE of the original model (corresponding to $\hat{\beta}$) on the original data is the same as the MSE of the new model (corresponding to $\hat{\gamma}$) on the centered data.

□ None of the above

**(b) (2 pt)** Which of following **are guaranteed** to be true regarding un-regularized ordinary least squares linear regression with an intercept term?

□ The mean of the residuals is 0.

□ The mean of the fitted values $\hat{y}$ equals the mean of the observed values $y$.

□ The weight corresponding to the intercept term is positive.

□ There is a unique solution to the normal equations.

□ None of the above

**(c)** **(1 pt)** In the remaining subparts of this question, we explore the models that result from ridge regression, depending on whether or not we regularize the intercept term ($\theta_0$) of our model.

$$\hat{\theta_A} = \arg\min_{\theta} \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \mathbb{X}_i^T\theta\right)^2 + \lambda\sum_{k=0}^{p}\theta_i^2$$

$$\hat{\theta_B} = \arg\min_{\theta} \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \mathbb{X}_i^T\theta\right)^2 + \lambda\sum_{k=1}^{p}\theta_i^2$$

Note that the only difference between $\hat{\theta_A}$ and $\hat{\theta_B}$ is that to determine $\hat{\theta_A}$, we regularize the intercept term, and in $\hat{\theta_B}$ we do not.

Assume we fix $\lambda = 2$ in both models. If we use $\dot{\mathbb{X}}$ instead of $\mathbb{X}$ above (i.e. if all columns of our design matrix, other than the intercept column, are centered), then $\hat{\theta_A} = \hat{\theta_B}$.

○ Always true

○ Sometimes true

○ Never true

**(d)** **(1 pt) In the remaining subparts of this problem, we are not assuming that our design matrix is mean-centered, i.e. we are not using $\dot{\mathbb{X}}$.**

Again, assume we fix $\lambda = 2$ in both models. Let $\text{MSE}_A$ be the training MSE using $\hat{\theta_A}$ and $\text{MSE}_B$ be the training MSE using $\hat{\theta_B}$.

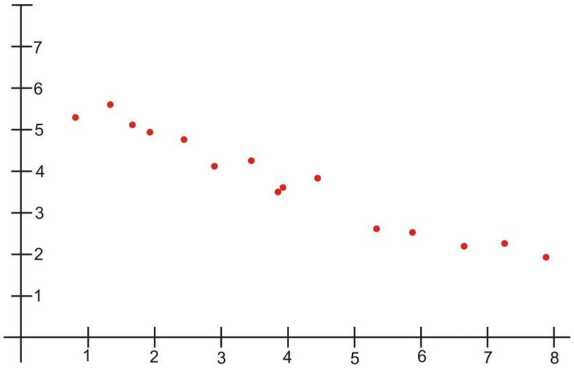
$$\text{MSE}_A \geq \text{MSE}_B$$

○ Always true

○ Sometimes true

○ Never true

**(e)** **(1 pt)** Again, assume we fix $\lambda = 2$ in both models. Suppose, for just this subpart, that our model is the simple linear model. Below, we've plotted our test data, with our sole feature on the $x$-axis and true responses on the $y$-axis.



Which parameter will achieve a lower test MSE?

○ $\hat{\theta_A}$

○ $\hat{\theta_B}$

**(f) (1 pt)** Now, let's consider the predictions made by the two models.

As $\lambda$ approaches positive infinity, what value do the predictions of the first model $(\hat{\theta_A})$ approach?

○ 0

○ the mean of the true $y$ values

○ $\frac{1}{\lambda}$

○ $\lambda$

○ $p+1$

○ Impossible to tell

**(g) (1 pt)** As $\lambda$ approaches positive infinity, what value do the predictions of the first model $(\hat{\theta_B})$ approach?

○ 0

○ the mean of the true $y$ values

○ $\frac{1}{\lambda}$

○ $\lambda$

○ $p+1$

○ Impossible to tell

**5. (6 points)**

Suppose we have a design matrix $\mathbb{X}$ and observed response vector $\mathbb{Y}$ that we assume are generated from some true linear model. Specifically, we assume $Y_i = f_{\theta^*}(\mathbb{X}_i) + \epsilon_i$, where

- $\mathbb{X}_i$ represents the $i$th row of our design matrix, and $Y_i$ represents the $i$th true response value
- $f_{\theta^*}(x)$ represents the underlying true linear model, parametrized by $\theta^*$
- Each $\epsilon_i$ is a random variable with $\mathbb{E}[\epsilon_i] = 0$ and $\text{Var}[\epsilon_i] = \sigma^2$

We want to use a linear model $\hat{\mathbb{Y}} = \mathbb{X}\theta$ to fit our data.

Assume that $\mathbb{X}$ and $\mathbb{Y}$ refer to our entire dataset. After performing a train-test split, we use squared loss and $L_2$ regularization with a positive regularization hyperparameter $\lambda$ (i.e. ridge regression) to determine our parameter estimate $\hat{\theta}$, using just our training data.

For each of following questions, select all options that are true.

**(a) (2 pt)** Which of following quantities are **fixed** in our true model to generate $\mathbb{X}$ and $\mathbb{Y}$?

□ $f_{\theta^*}(x)$

□ $\epsilon$

□ $\sigma^2$

□ $\hat{\theta}$

□ $\theta^*$

□ $\lambda$

□ None of the above

**(b) (2 pt)** As we increase $\lambda$, our estimated parameter $\hat{\theta}$ changes. What other values also change as we increase $\lambda$?

□ $\epsilon$

□ $\sigma^2$

□ $\theta^*$

□ Model bias

□ Model variance

□ Training error

□ Test error

□ None of the above

**(c)** **(2 pt)** In the bias-variance decomposition, our model's risk was defined as $\mathbb{E}[(Y - f_{\hat{\theta}}(x))^2]$.

Which of the following is equal to $\mathbb{E}[Y - f_{\hat{\theta}}(x)]$?

○ 0

○ The negative of model bias

○ Model bias

○ Model bias squared

○ Model variance

○ Model standard deviation

○ Observation variance

○ None of the above

**6. (14 points)**

Throughout this question, we are dealing with pandas DataFrame and Series objects. All code for this question, where applicable, must be written in Python. You may assume that pandas has been imported as pd.

The following DataFrame `consoles` contains the names of all major video game consoles. The `name` column is the primary key of the table. The `type` column contains the console type, which can only take on the values `handheld` and `home`. The `sales` column indicates the number of units sold in millions (rounded to the nearest million). The `year` column indicates the release year of the console. The data in the `Generation`, `Sales`, `Year` columns are all stored as `int`. All data come from Wikipedia, and the first six rows are shown below.

| name | generation | manufacturer | type | year | sales |
|---|---|---|---|---|---|
| Playstation 2 | 6 | Sony | home | 2000 | 155 |
| Xbox 360 | 7 | Microsoft | home | 2005 | 84 |
| Nintendo DS | 7 | Nintendo | handheld | 2004 | 154 |
| Atari 2600 | 2 | Atari | home | 1977 | 30 |
| Dreamcast | 6 | Sega | home | 1998 | 9 |
| Playstation 4 | 8 | Sony | home | 2013 | 110 |

**(a) (2 pt)** Which of the following lines of code correctly finds the names of all sixth generation consoles that were released before the year 2000?

□ `consoles[(consoles['generation'] == 6) & (consoles['year'] < 2000)]['name']`

□ `consoles[consoles['generation'] == 6].loc[:2000, 'name']`

□ `consoles[consoles['year'] < 2000].groupby('generation').loc[6]['name']`

□ `consoles[consoles['year'] < 2000].groupby('generation').sum().loc[6]['name']`

□ None of the above

**(b) (2 pt)** Which of the following lines of code correctly finds the number of seventh generation consoles?

□ `consoles[consoles['generation'] == 7].shape[0]`

□ `consoles.groupby('generation').size().loc[7]`

□ `consoles.groupby('generation').size().iloc[7]`

□ `consoles.groupby('name')['generation'].count().loc[7]`

□ None of the above

**(c) (2 pt)** Fill in the blank to find the latest generation that Atari participated in. Write Pandas code. Your code should return only the latest generation.

```
latest_gen = consoles[_________]["generation"].max()
```

**(d)** **(4 pt)** Create a DataFrame containing data only for those consoles whose manufacturers have sold at least 30 million units of **each** of their consoles. Write Pandas code. The resulting DataFrame must have the same structure and format as `consoles`. You may only use one line of code. The resulting DataFrame should be assigned to the variable `df`.

**(e)** **(4 pt)** Find the best-selling home console and the best-selling handheld console of all time. Write Pandas code. The result can have any number of columns, but it must only have two rows corresponding to the best-selling home console and the best-selling handheld console. Assume that are no ties. You may only use one line of code. The resulting DataFrame must be assigned to the variable `best_selling`.

**7. (4 points)**

Let us consider the *mean nearest squared error* for a dataset and regression line, which is calculated as the mean of the squared distances from each observation to the **nearest** point on our regression line.

**(a) (1 pt)** The mean nearest squared error will be _________ the mean squared error for any dataset. Assume that both the mean squared error and the slope of the regression line are not zero.

○ less than

○ equal to

○ greater than

○ not enough information

**(b) (1 pt)** The mean nearest squared error will be ________ the mean absolute error for any dataset. Again, assume that both the mean absolute error and the slope of the regression line are not zero.

○ less than

○ equal to

○ greater than

○ not enough information

**(c) (2 pt)** Kevin is trying to fit a simple linear regression model $\hat{y} = \theta_0 + \theta_1 x$ on a dataset with 900 observations.

After determining $\hat{\theta_0}$ and $\hat{\theta_1}$ by minimizing mean squared error, Kevin found the following:

- For 250 inputs, his predicted value $\hat{y}$ was 1 more than the actual value $y$.
- For another 500 inputs, his predicted value $\hat{y}$ was 0.5 less than the actual value $y$.
- All other points were predicted perfectly.

Calculate the **mean squared error** of Kevin's model. Select the closest answer. (Note that this question is not asking about mean nearest squared error.)

○ 0

○ 0.25

○ 0.35

○ 0.42

○ 0.55

○ 1

**8. (5 points)**

**(a) (1 pt)** Suppose that during the process of finding the SVD of our centered design matrix $X$, we lose information of the $\Sigma$ matrix. We still have $X$, $U$, and $V^T$. Can we still find our data's principal components?

○ Yes

○ No

**(b) (2 pt)** Let $\vec{u}_1$ be the first column and let $\vec{u}_2$ be the second column of the matrix $U$ given from the SVD. Which of the following combinations of values of $\vec{u}_1$ and $\vec{u}_2$ are **not valid**?

□ $\vec{u}_1 = \begin{bmatrix}1 & 0 & 0 & 0 & 0 & 0\end{bmatrix}^T, \vec{u}_2 = \begin{bmatrix}0 & 0 & 1 & 0 & 0 & 0\end{bmatrix}^T$

□ $\vec{u}_1 = \begin{bmatrix}1 & 0 & 0 & 0 & 0 & 0\end{bmatrix}^T, \vec{u}_2 = \begin{bmatrix}0 & 1 & 0 & 0 & 0 & 0\end{bmatrix}^T$

□ $\vec{u}_1 = \begin{bmatrix}\frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0\end{bmatrix}^T, \vec{u}_2 = \begin{bmatrix}\frac{1}{2} & -\frac{1}{2} & 0 & 0 & 0 & 0\end{bmatrix}^T$

□ $\vec{u}_1 = \begin{bmatrix}1 & 0 & 0 & 0 & 0 & 0\end{bmatrix}^T, \vec{u}_2 = \begin{bmatrix}\frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0\end{bmatrix}^T$

□ $\vec{u}_1 = \begin{bmatrix}0.8 & -0.6 & 0 & 0 & 0 & 0\end{bmatrix}^T, \vec{u}_2 = \begin{bmatrix}0.6 & 0.8 & 0 & 0 & 0 & 0\end{bmatrix}^T$

□ None of the above

**(c) (2 pt)** We have some centered design matrix $X$ with 20 observations and 3 features. The variance of the three features are 125, 20, and 5 respectively. Suppose we use the SVD on $X$ to get $U$, $\Sigma$ and $V^T$. Which of the following is the tightest possible lower bound for $\sigma_1$ (the largest singular value)?

*Note: For example, if $x = 5$, $x \geq 4$ is a tighter lower bound than $x \geq 0$, since 4 is closer to 5 than 0 is.*

○ $\sigma_1 \geq 125$

○ $\sigma_1 \geq 50$

○ $\sigma_1 \geq 20$

○ $\sigma_1 \geq 5$

○ $\sigma_1 \geq 0$

**9. (4 points)**

Consider the following table `classes`, which contains information regarding what classes students took at a particular high school. The SID uniquely identifies a student, but it is not a primary key in the `classes` table because a student can take multiple classes.

```
CREATE TABLE classes (
    sid INT // the SID of a student,
    class TEXT // the name of a class,
    grade_option TEXT // PNP (pass/no pass) or GRD (letter grade)
);
```

Here are the first few rows of the classes table:

| sid | class | grade_option |
|---|---|---|
| 53 | IB Biology | PNP |
| 53 | English 11 | GRD |
| 23 | IB Calculus | GRD |
| 23 | IB Biology | GRD |
| 23 | Spanish 4 | PNP |
| 10 | IB United States History | GRD |
| 10 | Physics | GRD |

**(a) (4 pt)** IB classes are designated by a `class` field that starts with 'IB'. IB classes are worth a maximum of 5 grade points while all other classes are worth a maximum of 4 grade points. Only classes taken for a letter grade (GRD) count towards a student's grade point average (GPA). Fill in the blank in the SQL query to find the maximum possible GPA each student could have. The output should include the SID and the maximum possible GPA.

If we assume that the table only contains the 7 rows above, then the output should be:

| sid | max_gpa |
|---|---|
| 53 | 4 |
| 23 | 5 |
| 10 | 4.5 |

```
SELECT sid, ____________ AS max_gpa
FROM classes
WHERE grade_option = 'GRD'
GROUP BY sid;
```

**10. (7 points)**

**(a) (2 pt)** Which of the following are true regarding fact and dimension tables?

□ Primary and foreign keys are useful in determining the relationship between fact and dimension tables.

□ The primary key for fact tables usually only consist of a single column.

□ Joins can help us combine information from fact and dimension tables.

□ Looking at a fact table alone tells us all the information we need regarding our data.

□ None of the above

**(b) (1 pt)** In this question, assume that we are using the logistic regression model $\hat{y} = \sigma(x^T\theta)$.

Suppose we want to modify cross-entropy loss to penalize predictions for observations that are truly positive twice as much as we penalize predictions for observations that are truly negative. Which of the following loss functions could we use? Recall that the average cross-entropy loss is:

$$R(\theta) = -\frac{1}{n}\sum_{i=1}^{n}\left(y_i \log(\hat{y_i}) + (1-y_i)\log(1-\hat{y_i})\right)$$

○ $R(\theta) = -\frac{2}{n}\sum_{i=1}^{n}\left(y_i \log(\hat{y_i}) + (1-y_i)\log(1-\hat{y_i})\right)$

○ $R(\theta) = -\frac{1}{n}\sum_{i=1}^{n}\left(2y_i \log(\hat{y_i}) + (1-y_i)\log(1-\hat{y_i})\right)$

○ $R(\theta) = -\frac{1}{n}\sum_{i=1}^{n}\left(y_i \log(\hat{y_i}) + 2(1-y_i)\log(1-\hat{y_i})\right)$

○ $R(\theta) = -\frac{1}{n}\sum_{i=1}^{n}\left((y_i+2) \log(\hat{y_i}) + (1-y_i)\log(1-\hat{y_i})\right)$

**(c) (2 pt)** Which of these models is linear? Select all that apply.

□ $f_\theta(x) = \theta_0 + \theta_1 x^n$

□ $f_\theta(x) = \theta_0 + \theta_1 x^{\theta_2}$

□ $f_\theta(x) = \theta_2 x_1 x_2 + \theta_0 x_2 x_3 + \theta_1 x_1^{x_2}$

□ $f_\theta(x) = \theta_1 x_1^2 + \theta_1 x_2^2 + \theta_1 x_3^2$

□ $f_\theta(x) = \theta_0 + \sum_{i=1}^{s} \theta_i \cos(ix)$

□ $f_\theta(x) = \theta_0 + \sum_{i=1}^{s} i \cos(\theta_i x)$

□ None of the above

**(d) (2 pt)** Suppose we want to use the bootstrap to find a percentile confidence interval for one of the parameters in a model. Which of the following statements are true? Select all that apply.

□ For a particular 90% confidence interval, there is a 90% chance that the true population parameter is inside the interval.

□ Roughly 90% of several 90% confidence intervals created using the same process should contain the true population parameter.

□ Exactly 90% of several 90% confidence intervals created using the same process should contain the true population parameter.

□ A 90% confidence interval is generally wider than a 95% confidence interval.

□ A 90% confidence interval is generally narrower than a 95% confidence interval.

□ The bootstrapped sampling distribution of an estimator must be normal in order for us to create a bootstrap percentile confidence interval.

□ None of the above

**11. (5 points)**

Fill in the blanks below to complete the implementation of `create_folds`, a function that takes in a value `k` and returns `train_folds` and `valid_folds` that we can use for cross-validation. Assume that `X` is a design matrix already defined in our notebook environment.

```
# Helper function to create k folds
  def create_folds(k):
      fold_size = ____(i)____
      shuffled_idx = np.random.permutation(X.shape[0])
      train_folds = []
      valid_folds = []
      for i in range(k):
          train_idx = np.append(shuffled_idx[0 : i*fold_size], shuffled_idx[(i + 1)*fold_size :])
          valid_idx = ____(ii)____
          train_folds.append(train_idx)
          valid_folds.append(valid_idx)

      return train_folds, valid_folds
```

**(a) (1 pt)** What goes in blank `(i)`?

○ `X.shape[0]`

○ `k`

○ `X.shape[1] // k`

○ `X.shape[0] // k`

**(b) (2 pt)** What goes in blank `(ii)`?

○ `np.append(shuffled_idx[i*fold_size], shuffled_idx[(i + 1)*fold_size])`

○ `shuffled_idx[i*fold_size : (i + 1)*fold_size]`

○ `i*fold_size : (i + 1)*fold_size`

○ `shuffled_idx - train_idx`

**(c)** **(2 pt)** Now, fill in the blank to complete the implementation of `k_fold_ridge_CV`, which takes in a value `k` and a list of `lambda_choices`, and returns the value in `lambda_choices` that minimizes cross-validation RMSE for a ridge regression model, trained on `X` and `y`.

```
lambda_choices = [1000, 100, 10, 1, 0.1, 0.01, 0.001]

def k_fold_ridge_CV(k, lambda_choices):
    train_folds, valid_folds = create_folds(k)
    all_errors = []

    for lamb in lambda_choices:
        lambda_errors = []
        for train_idx, valid_idx in zip(train_folds, valid_folds):
            model = lm.Ridge(alpha = lamb)
            model.fit(X[train_idx], y[train_idx])
            rmse = ____(iii)____
            lambda_errors.append(rmse)
        all_errors.append(np.mean(lambda_errors))

    return lambda_choices[np.argmin(errors)]
```

What goes in blank `(iii)`?

○ `np.sqrt(np.mean((model.predict(X[valid_idx]) - y[valid_idx])**2))`

○ `np.mean((model.predict(X[valid_idx]) - y[valid_idx])**2)`

○ `np.sqrt(np.mean((model.predict(X[train_idx]) - y[train_idx])**2))`

○ `np.sqrt(np.mean((y[valid_idx] - X[valid_idx] @ lm.coef_ - lm.intercept_)**2) + lamb * np.sum(lm.coef_**2))`

**12. (4 points)**

(a) **(4 pt)** The Pappy's ordering machine is broken! Every time a customer orders, the machine adds gibberish that make it harder to read and sometimes adds fake orders, causing long wait times. A real order is a string that starts with the number of items followed by the item name, and the order ends at the first "|". Each order only has one quantity and one item name (an order cannot have multiple different items).

Below, write a **regular expression** to extract valid orders. Assume customers cannot order more than 99 items in one order, and orders that are single digits show up as a single number (e.g. 9 instead of 09). If a customer tries to order more than 99 items, your regular expression should only capture the last two digits. Also, don't hardcode the answer or Pappy's will fire you and you will receive no credit.

Your answer must be a regular expression.

```
orders = ['9burgers|', '40fries|fhw|', 'sevensprites|', '0000009nuggets|mj1i2', \
'sofjsd35milkshakes|1quarterpounder|', '534salads|burgerkingisking']
pattern = r'__________________'

>>> for order in orders:
...    print(re.findall(pattern, order))
['9burgers']
['40fries']
[]
['9nuggets']
['35milkshakes']
['34salads']
```

# Data 100, Final Day 2

Summer 2020

## Instructions:

- This exam consists of 9 questions (numbered 0 through 8), worth a total of 75 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**.
- You must write this exam on paper; you may not use a tablet or any other digital writing utensil.
- **Please show your work. Answers without work shown may not receive full credit.**
- Please write your initials on the top of every page.

# 0 Statement of Academic Integrity [1 Pt]

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.** Note that this question is worth a point.

*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.*

Write down the time at which you are taking this exam. You should choose from the following options:

- Thursday, 7:00-8:45PM PDT (main exam)
- Thursday, 7:00-9:30PM PDT (DSP 150%)
- Thursday, 7:00-10:15PM PDT (DSP 200%)
- Friday, 8:00-9:45AM PDT (alternate)
- Friday, 8:00-10:30AM PDT (alternate DSP 150%)
- Friday, 8:00-11:15AM PDT (alternate DSP 200%)
- Other

You **must** submit the exam to Gradescope by the ending time listed above.

# 1 Great Expectations [8 Pts]

1. Suppose you roll a fair six-sided die 10 times.

   - Let $X_1$ be the number of 1s you see in 10 rolls.
   - Let $X_2$ be the number of 3s or 5s you see in 10 rolls.
   - Let $X_3$ be the number of even numbers you see in 10 rolls.

   For example, $[1, 3, 5, 5, 5, 1, 2, 4, 6, 2]$ is one of many possible sequences when $X_1 = 2$, $X_2 = 4$, and $X_3 = 4$.

   (a) [2 Pts] Determine $\mathbb{E}[X_2]$. Simplify your answer.

   (b) [2 Pts] Determine $\mathbb{E}[X_1 + X_2 + X_3]$. Simplify your answer.

   (c) [2 Pts] Are $X_1$ and $X_2$ independent? Answer yes or no, and justify your answer. (Do not say "yes, because $X_1$ depends on $X_2$" or "no, because $X_1$ doesn't depend on $X_2$"; credit will only be given to rigorous answers.)

   (d) [2 Pts] Find $P(X_1 = 2, X_2 = 3, X_3 = 5)$. You may leave your answer as an unsimplified fraction.

# 2 I Scream for Ice Cream! [10 Pts]

2. The local ice cream shop has just hired you as a data scientist. Your first task is to build a simple linear model, $\hat{y} = \theta_0 + \theta_1 x$, where you will predict the total amount of sales (in dollars) using the daily high temperature (in degrees Celsius). Given their data set, you compute the following summary statistics:

| | | | | |
|---|---|---|---|---|
| $r = 0.7$ | $x_{min} = 15°C$ | $x_{max} = 38°C$ | $\bar{x} = 20°C$ | $\sigma_x = 2°C$ |
| $n = 50$ | $y_{min} = \$0$ | $y_{max} = \$98.50$ | $\bar{y} = \$7.50$ | $\sigma_y = \$1$ |

To fit your model, you decide to use squared loss.

(a) [2 Pts] Find the optimal parameters $\hat{\theta}_0$ and $\hat{\theta}_1$. Simplify your answer.

(b) [2 Pts] One observation in the dataset had a daily high temperature of $30°C$, and had a total sales amount of $\$19.00$. What is the residual for this observation?

(c) [3 Pts] What is the variance of the fitted values in your regression model?

(d) [3 Pts] What is the mean squared error of your regression model? *Hint: You may consider using the fact that* $\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$.

# 3 More Regression? Really? [4 Pts]

3. In both parts of this question, assume that we are fitting a linear model $\hat{\mathbb{Y}} = \mathbb{X}\theta$ by minimizing mean squared error without regularization. Also assume that our design matrix $\mathbb{X}$ has $n$ rows and $p + 1$ columns.

   (a) [2 Pts] Let $M$ be the MSE of our ordinary least squares regression model. Let $S = ||e||_2$, i.e. let $S$ be the $L_2$ norm of the residual vector of our model.

   Write a mathematical expression that expresses $M$ in terms of $S$. *Hint: You may use $n$ and/or $p$ in your answer.*

   (b) [2 Pts] Suppose our design matrix is square (i.e. $n = p + 1$) and full rank. What is the MSE of our resulting linear model? Select from one of the following three answers, and **explain your choice**.

   - Some number less than 0
   - 0
   - Some number greater than 0

# 4 Please Be Mine [6 Pts]

4. The Data 100 Mining Company is ethically and sustainably mining mineral ores. They record the mass, volume, and density of the $n$ ores they mine in Table A, which consists only of those three columns. Note that density can be expressed as a combination of mass and volume using the formula density $=$ mass $\cdot$ volume$^{-1}$.

   **For each of the following parts, justify your answer. Answers without sufficient justification will not receive full credit.**

   (a) [1 Pt] What is the rank of Table A?

   (b) [2 Pts] In some new table, Table B, they record the rarity of the $n$ ores as "Abundant", "Common", "Scarce", or "Rare", after which they one-hot encode the rarity feature. (To be clear, there are only 4 columns in Table B.) The company then merges Tables A and B to form Table C. What is the rank of Table C? *You may assume there are no issues in merging Tables A and B.*

   (c) [3 Pts] They then record whether the ore is magnetic or not, one hot-encode this feature, and add it to Table C along with an intercept term to form Table D. What is the rank of Table D? *As before, you may assume there are no issues with merging the tables.*

# 5 GPA Descent [7 Pts]

5. Consider the following non-linear model with two parameters:

$$f_\theta(x) = \theta_0 \cdot 0.5 + \theta_0 \cdot \theta_1 \cdot x_1 + \sin(\theta_1) \cdot x_2$$

For some nonsensical reason, we decide to use the residuals of our model as the loss function. That is, the loss for a single observation is

$$L(\theta) = y_i - f_\theta(x_i)$$

We want to use gradient descent to determine the optimal model parameters, $\hat{\theta}_0$ and $\hat{\theta}_1$.

(a) [3 Pts] Suppose we have just one observation in our training data, $(x_1 = 1, x_2 = 2, y = 4)$. Assume that we set the learning rate $\alpha$ to 1.

An incomplete version of the gradient descent update equation for $\theta$ is shown below. $\theta_0^{(t)}$ and $\theta_1^{(t)}$ denote the guesses for $\theta_0$ and $\theta_1$ at timestep $t$, respectively.

$$\begin{bmatrix} \theta_0^{(t+1)} \\ \theta_1^{(t+1)} \end{bmatrix} = \begin{bmatrix} \theta_0^{(t)} \\ \theta_1^{(t)} \end{bmatrix} - \begin{bmatrix} A \\ B \end{bmatrix}$$

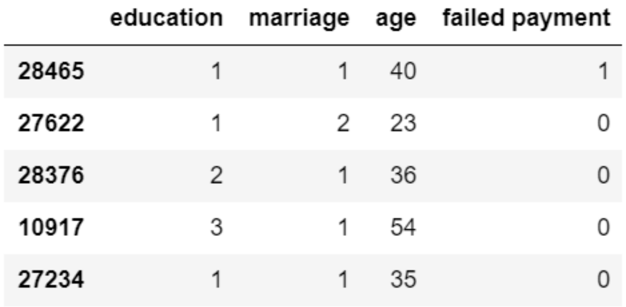
Express both $A$ and $B$ in terms of $\theta_0^{(t)}$, $\theta_1^{(t)}$, and any necessary constants.

(b) [2 Pts] Assume we initialize both $\theta_0^{(0)}$ and $\theta_1^{(0)}$ to 0. Determine $\theta_0^{(1)}$ and $\theta_1^{(1)}$ (i.e. the guesses for $\theta_0$ and $\theta_1$ after one iteration of gradient descent).

(c) [2 Pts] What happens to $\theta_0^{(t)}$ as $t \to \infty$ (i.e. as we run more and more iterations of gradient descent)?

# 6 Wall Street Logistics [18 Pts]

6. You have been tasked with performing an analysis on customers of a credit card company. Specifically, you will be developing a classification model to classify whether or not specific customers will fail to pay their next credit card payment. You decide to approach this problem with a logistic regression classifier. The first 5 rows of our data are shown below.



| | education | marriage | age | failed payment |
|---|---|---|---|---|
| **28465** | 1 | 1 | 40 | 1 |
| **27622** | 1 | 2 | 23 | 0 |
| **28376** | 2 | 1 | 36 | 0 |
| **10917** | 3 | 1 | 54 | 0 |
| **27234** | 1 | 1 | 35 | 0 |

The numerical data in the education and marriage columns correspond to the following categories:

`education`: 1 - graduate school; 2 - university; 3 - high school; 4 - other
`marriage`: 1 - married; 2 - single; 3 - other

Our response variable, labeled as failed payment, can have values of 0 (makes their next payment) or 1 (fails to make their next payment).

You use the logistic regression model $\hat{y} = P(Y = 1|x) = \sigma(x^T\theta)$ . Assume that the following value of $\hat{\theta}$ minimizes un-regularized mean cross-entropy loss for this data set:

$$\hat{\theta} = [-1.30, 0.08, -0.08, 0.001]^T$$

Here, -1.30 is the intercept term, 0.08 corresponds to `education`, -0.08 corresponds to `marriage` status, and 0.001 corresponds to `age`.

(a) [2 Pts] Consider a customer who is 50 years old, married, and only has a high school education. Compute the chance that they fail to pay their next credit card payment. Give your answer as a probability in terms of $\sigma$.

(b) [2 Pts] This specific customer fortunately made their next payment on time! Compute the cross-entropy loss of the prediction in part a. Leave your answers in terms of $\sigma$.

(c) [2 Pts] How does a one-unit increase in `age` impact the **log-odds** of making a failed payment? Give a precise, numerical answer, not just "it increases" or "it decreases."

(d) [3 Pts] Let's consider all customers who are married and whose highest level of education is high school. What is the **minimum age** of such a customer, such that they more likely to fail their next payment than make their next payment, under our logistic regression model?

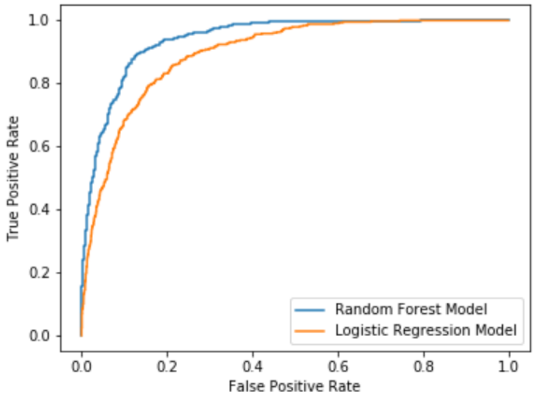
(e) [3 Pts] Suppose you choose a threshold $T = 0.8$. The decision boundary of the resulting classifier is of the form

$$A \cdot \texttt{education} + B \cdot \texttt{marriage} + C \cdot \texttt{age} + D = 0$$

What are the values of $A$, $B$, $C$, and $D$? Your answers may contain a $\log$, but should not contain $\sigma$. Show your work.

(f) [2 Pts] Suppose with the above threshold you achieve a training accuracy of 100%. Can you conclude your training data was linearly separable in the feature space? Answer yes or no, and explain in one sentence.

(g) [2 Pts] To further your analysis, you also create a random forest classifier. To compare classifiers you generate a ROC curve for both models. Which of the two models would you choose to use, based on the ROC curve? Explain in one sentence. (Do not worry about the implementation details of how ROC curves are created for random forests.)



(h) [2 Pts] For whatever reason, we decide to multiply the `education` feature by 4 and `age` feature by 0.5 in our data. What is the new value of $\hat{\theta}$ that minimizes un-regularized mean cross-entropy loss?

If you don't believe it's possible to tell, say so. For your convenience, the value of $\hat{\theta}$ that minimized un-regularized mean cross-entropy loss on the original data was $\hat{\theta} = [-1.30, 0.08, -0.08, 0.001]^T$; here, -1.30 is the intercept term, 0.08 corresponds to `education`, -0.08 corresponds to `marriage` status, and 0.001 corresponds to `age`.

Note: This question is independent of parts e and f, i.e. do not assume that we achieved a training accuracy of 100%.

# 7 Pacific Coast Academy [15 Pts]

7. Suppose you have some design matrix $D$ with 51 rows and 5 columns. Below are the first 3 rows of $D$:

| | | | | |
|---|---|---|---|---|
| 3 | 3 | 12 | 5 | 1 |
| 3 | 4 | 8 | 2 | 1 |
| 1 | 2 | 12 | 7 | 2 |

You are also given the following information regarding the mean and variance of each of the 5 columns:

| | Column 1 | Column 2 | Column 3 | Column 4 | Column 5 |
|---|---|---|---|---|---|
| Mean | 2 | 3 | 10 | 5 | 1 |
| Variance | 0.3 | 0.3 | ? | 0.3 | 0.3 |

You compute the singular value decomposition of $D$ using the following Python code:

```
X = D - np.mean(D, axis=0) # X is the centered version of D
u, s, vt = np.linalg.svd(X, full_matrices=False)
```

The value of `s` that results is `array([9, 4, 2, 1, 0])`.

**For all parts of this question, you must justify your answer for full credit. If the answer to any part cannot be found with the information given, write "not enough information."**

(a) [1 Pt] What is the rank of matrix $D$?

(b) [2 Pts] What is the variance accounted for by the second principal component? Your answer should not contain any variables.

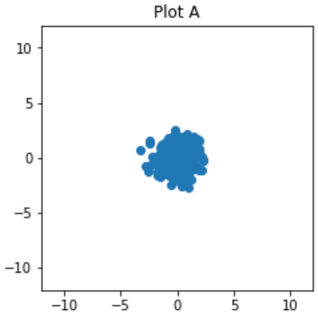
(c) [2 Pts] We want to choose the first $k$ principal components, such that at least 95% of the variation in $D$ is retained. What is the smallest possible value of $k$ that we can choose?

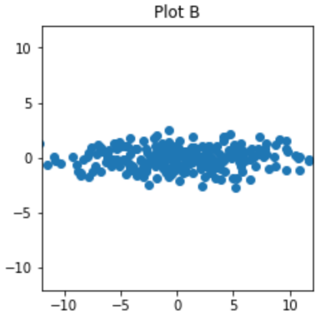
(d) [3 Pts] Unfortunately, the data you received got corrupted, so they do not include the variance of column 3. Given all of the information above, what is the variance of column 3?

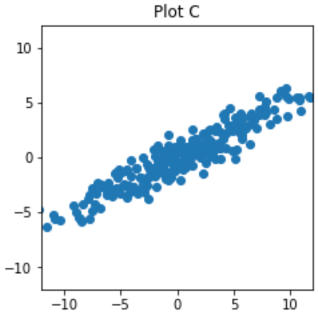
(e) [3 Pts] You are now given that the output of `vt[2]` is `array([0.8, 0.6, 0, 0, 0])`. Note that this is the third row of $V^T$ because of 0-indexing. Given this information, what is the output of `u[0, 2]` (the first entry in the third column of $U$)? *Hint: Recall that* $X\vec{v}_i = \sigma_i \vec{u}_i$.

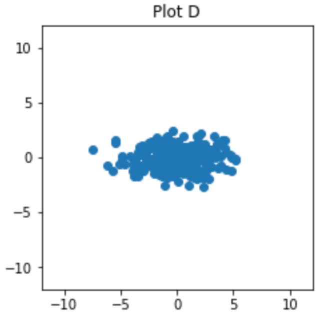
(f) [2 Pts] What is the output of `np.mean(u[:, 0])`?

(g) [2 Pts] Which of the following four scatter plots depicts PC 2 plotted against PC 1? There is only one correct answer, and one of the four answers is correct.









# 8 Computational and Inferential Thinking [6 Pts]

8. Suppose we are trying to estimate the true tip percentage $\theta^*$ given to waiters and waitresses at restaurants. Instead of collecting data and fitting a model, we use another approach.

   We use a random number generator to select an integer between 1 and 20 (inclusive of both endpoints), uniformly at random. We repeat this process $n$ times, giving us a total of $n$ numbers. You may assume that each selection is independent.

   Our estimator $\hat{\theta}$ is the number of prime numbers in our list of $n$ numbers. For reference, there are 8 prime numbers between 1 and 20, inclusive.

   Suppose $\theta^* = 16$.

   (a) [3 Pts] What is the bias of the estimator $\hat{\theta}$? *Hint: The only variable your answer should contain is $n$.*

   (b) [1 Pt] What should $n$ be so that $\hat{\theta}$ is an unbiased estimator of $\theta^*$?

   (c) [2 Pts] What is the variance of the estimator $\hat{\theta}$ when using the value of $n$ from part b?