INSTRUCTIONS

• You have 70 minutes to complete the exam.
• The exam is closed book, closed notes, closed computer, closed calculator, except for two 8.5" × 11" crib sheets of your own creation.
• Mark your answers on the exam itself. We will not grade answers written on scratch paper.

<table>
<thead>
<tr>
<th>Last name</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>First name</td>
<td></td>
</tr>
<tr>
<td>Student ID number</td>
<td></td>
</tr>
<tr>
<td>CalCentral email</td>
<td>(<a href="mailto:_@berkeley.edu">_@berkeley.edu</a>)</td>
</tr>
<tr>
<td>Exam room</td>
<td></td>
</tr>
<tr>
<td>Name of the person to your left</td>
<td></td>
</tr>
<tr>
<td>Name of the person to your right</td>
<td></td>
</tr>
</tbody>
</table>

All the work on this exam is my own. (please sign)

Terminology and Notation Reference:

<table>
<thead>
<tr>
<th>exp(x)</th>
<th>$e^x$</th>
</tr>
</thead>
<tbody>
<tr>
<td>log(x)</td>
<td>$\log_e x$</td>
</tr>
<tr>
<td>Linear regression model</td>
<td>$E[Y</td>
</tr>
<tr>
<td>Logistic (or sigmoid) function</td>
<td>$\sigma(t) = \frac{1}{1+\exp(-t)}$</td>
</tr>
<tr>
<td>Logistic regression model</td>
<td>$P(Y = 1</td>
</tr>
<tr>
<td>Squared error loss</td>
<td>$L(y, \theta) = (y - \theta)^2$</td>
</tr>
<tr>
<td>Absolute error loss</td>
<td>$L(y, \theta) =</td>
</tr>
<tr>
<td>Cross-entropy loss</td>
<td>$L(y, \theta) = -y \log \theta - (1 - y) \log(1 - \theta)$</td>
</tr>
<tr>
<td>Bias</td>
<td>Bias[\hat{\theta}, \theta] = E[\hat{\theta}] - \theta</td>
</tr>
<tr>
<td>Variance</td>
<td>Var[\hat{\theta}] = E[(\hat{\theta} - E[\hat{\theta}])^2]</td>
</tr>
<tr>
<td>Mean squared error</td>
<td>MSE[\hat{\theta}, \theta] = E[(\hat{\theta} - \theta)^2]</td>
</tr>
</tbody>
</table>
1. (8 points) Feature Engineering

For each dataset depicted below in a scatterplot, fill in the squares next to all of the letters for the vector-valued functions $f$ that would make it possible to choose a column vector $\beta$ such that $y_i = f(x_i)^T \beta$ for all $(x_i, y_i)$ pairs in the dataset. The input to each $f$ is a scalar $x$ shown on the horizontal axis, and the corresponding $y$ value is shown on the vertical axis.

(A) $f(x) = [1 \ x]^T$
(B) $f(x) = [x \ 2x]^T$
(C) $f(x) = [1 \ x \ x^2]^T$
(D) $f(x) = [1 \ |x|]^T$
(E) None of the above

(i) (2 pt) □ A □ B □ C □ D □ E

(ii) (2 pt) □ A □ B □ C □ D □ E

(iii) (2 pt) □ A □ B □ C □ D □ E

(iv) (2 pt) □ A □ B □ C □ D □ E
2. (6 points) Estimation

A learning set \((x_1, y_1), \ldots, (x_{10}, y_{10})\) is sampled from a population where \(X\) and \(Y\) are both binary.

The learning set data are summarized by the following table of row counts:

<table>
<thead>
<tr>
<th>(x)</th>
<th>(y)</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
</tbody>
</table>

(a) (4 pt) You decide to fit a constant model \(P(Y = 1|X = 0) = P(Y = 1|X = 1) = \alpha\) using the cross-entropy loss function and no regularization. What is the formula for the empirical risk on this learning set for this model and loss function? What estimate of the model parameter \(\hat{\alpha}\) minimizes empirical risk? You must show your work for finding the estimate \(\hat{\alpha}\) to receive full credit.

Recall: Since \(Y\) is binary, \(P(Y = 0|X) + P(Y = 1|X) = 1\) for any \(X\).

Empirical Risk:

Estimate \(\hat{\alpha}\) (show your work):

(b) (2 pt) The true population probability \(P(Y = 0|X = 0)\) is \(\frac{1}{3}\). Provide an expression in terms of \(\hat{\alpha}\) for the bias of the estimator of \(P(Y = 0|X = 0)\) described in part (a) for the constant model. You may use \(E[\ldots]\) in your answer to denote an expectation under the data generating distribution of the learning set, but do not write \(P(\ldots)\) in your answer.

\[
\text{Bias}[\hat{P}(Y = 0|X = 0), P(Y = 0|X = 0)] =
\]
3. (6 points) Linear Regression

A learning set of size four is sampled from a population where $X$ and $Y$ are both quantitative:

$$(x_1, y_1) = (2.5, 3)$$
$$(x_2, y_2) = (2, 5)$$
$$(x_3, y_3) = (1, 3)$$
$$(x_4, y_4) = (3, 5).$$

You fit a linear regression model $E[Y|X] = \beta_0 + X \beta_1$, where $\beta_0$ and $\beta_1$ are scalar parameters, by ridge regression, minimizing the following objective function:

$$\frac{1}{4} \sum_{i=1}^{4} (y_i - (\beta_0 + x_i \beta_1))^2 + \frac{\beta_0^2 + \beta_1^2}{3}.$$ 

(a) (4 pt) Fill in all blanks below to compute the parameter estimates that minimize this regularized empirical risk. (You do not need to compute their values; just fill in the matrices appropriately.)

\[
\begin{bmatrix}
X^T_n &=& \begin{bmatrix}
\_ & \_ & \_ & \_ & \_ \\
2.5 & 2 & 1 & 3
\end{bmatrix}
\end{bmatrix}
\]

\[
\begin{bmatrix}
Y^T_n &=& \begin{bmatrix}
\_ & \_ & \_ & \_ & \_ \\
\_ & \_ & \_ & \_ & \_ & \_ & \_ & \_ & \_ & \_ & \_
\end{bmatrix}
\end{bmatrix}
\]

\[
\begin{bmatrix}
\hat{\beta}_0 \\
\hat{\beta}_1
\end{bmatrix} = (X^T_n X_n + \begin{bmatrix}
\_ & \_ & \_ & \_ & \_
\end{bmatrix})^{-1} X^T_n Y_n.
\]

(b) (2 pt) Without computing values for $\hat{\beta}_0$ and $\hat{\beta}_1$, write an expression for the squared error loss of the learning set observation $(x_4, y_4)$ in terms of $\hat{\beta}_0$ and $\hat{\beta}_1$ and any relevant numbers. Your solution should not contain any of $\hat{y}_4$, $x_4$, or $y_4$, but instead just numbers and $\hat{\beta}_0$ and $\hat{\beta}_1$.

\[
L(y_4, \hat{y}_4) =
\]
4. (8 points) Model Selection

(a) (2 pt) You have a quantitative outcome $Y$ and two quantitative covariates ($X_1, X_2$). You want to fit a linear regression model for the conditional expected value $E[Y|X]$ of the outcome given the covariates, including an intercept. Bubble in the minimum dimension of the parameter vector $\beta$ needed to express this linear regression model?

○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ None of these

(b) (2 pt) You have a quantitative outcome $Y$ and two qualitative covariates ($X_1, X_2$). $X_1 \in \{a, b, c, d\}$, $X_2 \in \{e, f, g\}$, and there is no ordering to the values for either variable. You want to fit a linear regression model for the conditional expected value $E[Y|X]$ of the outcome given the covariates, including an intercept. Bubble in the minimum dimension of the parameter vector $\beta$ needed to express this linear regression model?

○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ 8 ○ 9 ○ 10 ○ 11 ○ 12 ○ 13

(c) (2 pt) Bubble all true statements: In ridge regression, when the assumptions of the linear model are satisfied, the larger the shrinkage/penalty parameter,
- the larger the magnitude of the bias of the estimator of the regression coefficients $\beta$.
- the smaller the magnitude of the bias of the estimator of the regression coefficients $\beta$.
- the larger the variance of the estimator of the regression coefficients $\beta$.
- the smaller variance of the estimator of the regression coefficients $\beta$.
- the smaller the true mean squared error of the estimator of the regression coefficients $\beta$.

(d) (2 pt) Bubble all true statements: A good approach for selecting the shrinkage/penalty parameter in LASSO is to:
- minimize the learning set risk for the squared error ($L_2$) loss function.
- minimize the learning set risk for the absolute error ($L_1$) loss function.
- minimize the cross-validated regularized risk for the squared error ($L_2$) loss function.
- minimize the cross-validated risk for the squared error ($L_2$) loss function.
- minimize the variance of the estimator of the regression coefficients.
5. **(12 points) Logistic Regression**

(a) **(2 pt)** Bubble the expression that describes the odds ratio \( \frac{P(Y=1|X)}{P(Y=0|X)} \) of a logistic regression model.

*Recall:* \( P(Y = 0|X) + P(Y = 1|X) = 1 \) for any \( X \).

- \( X^T \beta \)
- \( -X^T \beta \)
- \( \exp(X^T \beta) \)
- \( \sigma(X^T \beta) \)
- None of these

(b) **(2 pt)** Bubble the expression that describes \( P(Y = 0|X) \) for a logistic regression model.

- \( \sigma(-X^T \beta) \)
- \( 1 - \log(1+\exp(X^T \beta)) \)
- \( 1 + \log(1+\exp(-X^T \beta)) \)
- None of these

(c) **(2 pt)** Bubble all of the following that are typical effects of adding an \( L_1 \) regularization penalty to the loss function when fitting a logistic regression model with parameter vector \( \beta \).

- The magnitude of the elements of the estimator of \( \beta \) are increased.
- The magnitude of the elements of the estimator of \( \beta \) are decreased.
- All elements of the estimator of \( \beta \) are non-negative.
- Some elements of the estimator of \( \beta \) are zero.
- None of the above.

(d) **(3 pt)** What would be the primary disadvantage of a regularization term of the form \( \sum_{j=1}^{J} \beta_j^3 \) rather than the more typical ridge penalty \( \sum_{j=1}^{J} \beta_j^2 \) for logistic regression? Answer in one sentence.

(e) **(3 pt)** For a logistic regression model \( P(Y = 1|X) = \sigma(-2 - 3X) \), where \( X \) is a scalar random variable, what values of \( x \) would give \( P(Y = 0|X = x) \geq \frac{3}{4} \)? **You must show your work for full credit.**