Data 100 & 200A Principles and Techniques of Data Science

Spring 2019

Midterm 2 Solutions

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.

Last name	
First name	
Student ID number	
CalCentral email (_@berkeley.edu)	
Exam room	
Name of the person to your left	
Name of the person to your right	
All the work on this exam is my own. **(please sign)**	

Terminology and Notation Reference:

exp(x)	e^x
log(x)	$\log_e x$
Linear regression model	$E[Y\|X] = X^T\beta$
Logistic (or sigmoid) function	$\sigma(t) = \frac{1}{1+\exp(-t)}$
Logistic regression model	$P(Y = 1\|X) = \sigma(X^T\beta)$
Squared error loss	$L(y, \theta) = (y - \theta)^2$
Absolute error loss	$L(y, \theta) = \|y - \theta\|$
Cross-entropy loss	$L(y, \theta) = -y \log\theta - (1 - y)\log(1 - \theta)$
Bias	$\text{Bias}[\hat{\theta}, \theta] = E[\hat{\theta}] - \theta$
Variance	$\text{Var}[\hat{\theta}] = E[(\hat{\theta} - E[\hat{\theta}])^2]$
Mean squared error	$\text{MSE}[\hat{\theta}, \theta] = E[(\hat{\theta} - \theta)^2]$

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 β 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) $f(x) = [1 \quad x]^T$

(B) $f(x) = [x \quad 2x]^T$

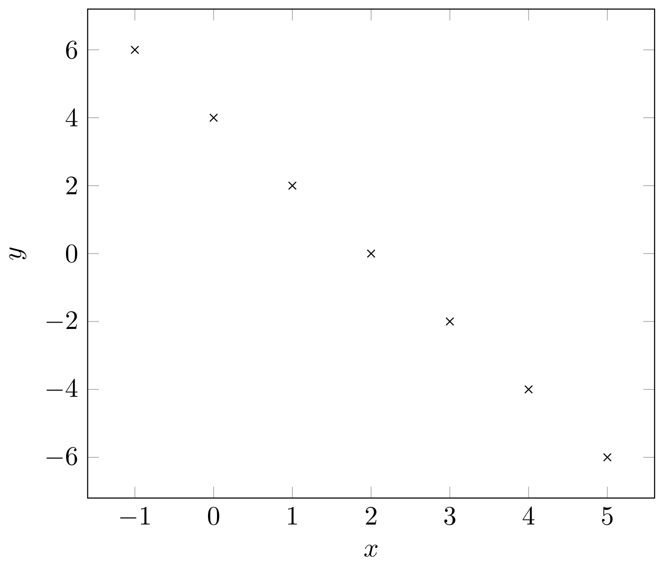
(C) $f(x) = [1 \quad x \quad x^2]^T$

(D) $f(x) = [1 \quad |x|]^T$

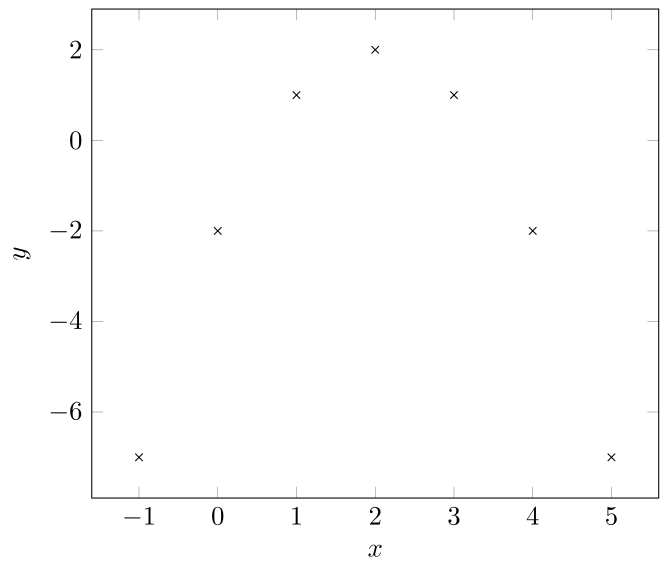
(E) None of the above

Notes: In (D), x=-1 and x=1 must have the same y value, so the V shape cannot be moved horizontally with a linear combination of those features. Dataset (ii) was intended to be parabolic, but the original printed version of the exam had an error in the parabola shape; sorry! Credit was given for C or E.

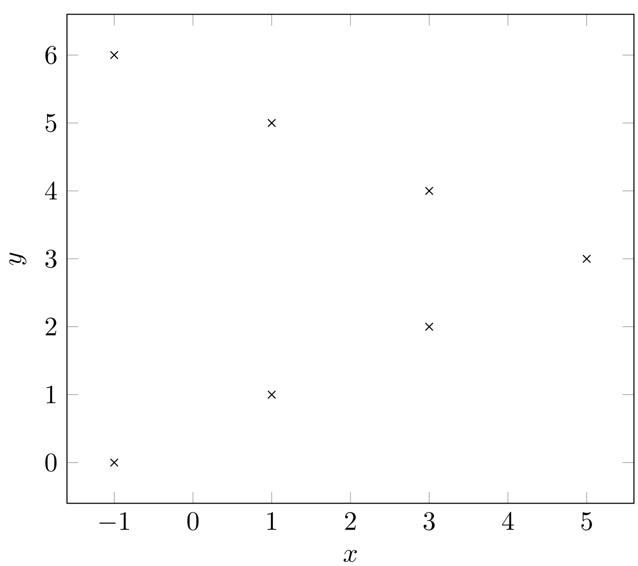
(i) (2 pt) ☐ * A ☐ B ☐ * C ☐ D ☐ E



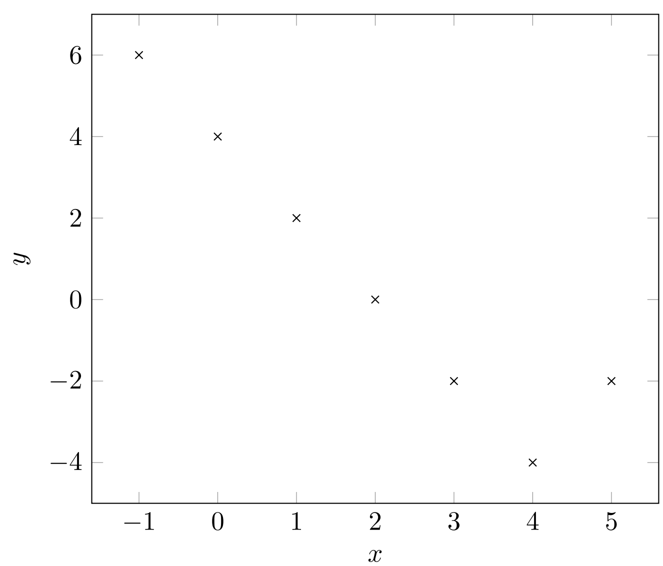
(ii) (2 pt) ☐ A ☐ B ☐ * C ☐ D ☐ E



(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:

x	y	Count
0	0	2
0	1	3
1	0	1
1	1	4

(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 α 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: $-\frac{7}{10} \log \alpha - \frac{3}{10} \log(1 - \alpha)$

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

$$0 = \frac{7}{10\alpha} - \frac{3}{10(1-\alpha)}$$
$$0 = 7(1-\alpha) - 3\alpha$$
$$10\alpha = 7$$
$$\alpha = \frac{7}{10}$$

(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.**

Bias$[\hat{P}(Y = 0|X = 0), P(Y = 0|X = 0)] = E[1 - \hat{\alpha}] - \frac{1}{3}$ or equivalently $\frac{2}{3} - E[\hat{\alpha}]$

Note: the value for $\hat{\alpha}$ computed in part (a) is just for this particular learning set, which is just one sample among many possible samples. We don't know from this one dataset that $E[\hat{\alpha}] = \frac{7}{10}$. Bias does not describe a particular estimate from a particular dataset, but instead refers to the average of estimates obtained from repeated random sampling from the population, i.e., the average of $\hat{\alpha}$ from multiple learning sets.

3. (6 points) Linear Regression

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

$$\begin{aligned}(x_1, y_1) &= (2.5, 3)\\ (x_2, y_2) &= (2, 5)\\ (x_3, y_3) &= (1, 3)\\ (x_4, y_4) &= (3, 5).\end{aligned}$$

You fit a linear regression model $E[Y|X] = \beta_0 + X\beta_1$, where β_0 and β_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.)

$$X_n^T = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 2.5 & 2 & 1 & 3 \end{bmatrix}$$

$$Y_n^T = \begin{bmatrix} 3 & 5 & 3 & 5 \end{bmatrix}$$

$$\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = (X_n^T X_n + \begin{bmatrix} \frac{4}{3} & 0 \\ 0 & \frac{4}{3} \end{bmatrix})^{-1} X_n^T Y_n.$$

Note: the common answer of $\frac{1}{3}$ on the diagonal of the regularization term was given full credit.

(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 $\hat{\beta}_1$.**

$L(y_4, \hat{y}_4) = (5 - (\hat{\beta}_0 + 3\hat{\beta}_1))^2$

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 β needed to express this linear regression model?

◯ 1 ◯ 2 ◯ * 3 ◯ 4 ◯ 5 ◯ 6 ◯ 7 ◯ None of these

Both quantitative features and the intercept are needed.

(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 β needed to express this linear regression model?

◯ 2 ◯ 3 ◯ 4 ◯ 5 ◯ * 6 ◯ 7 ◯ 8 ◯ 9 ◯ 10 ◯ 11 ◯ 12 ◯ 13

Each categorical variable with k outcomes requires $k-1$ features to encode, because an additional feature would be a linear combination of the others and the intercept feature. $(4-1)+(3-1)+1=6$.

(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 β.

☐ the smaller the magnitude of the bias of the estimator of the regression coefficients β.

☐ the larger the variance of the estimator of the regression coefficients β.

☐ the smaller variance of the estimator of the regression coefficients β.

☐ the smaller the true mean squared error of the estimator of the regression coefficients β.

(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.

The cross-validated L_2 risk is a good unbiased estimator for the L_2 risk (average L_2 loss) on unseen data, which is the quantity we care to minimize in the end. The L_1 norm of the regression coefficients in LASSO is a regularization/penalty term that appears only for the purpose of estimating β on the training set.

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 β.

☐ The magnitude of the elements of the estimator of β are increased.

☐ The magnitude of the elements of the estimator of β are decreased.

☐ All elements of the estimator of β are non-negative.

☐ Some elements of the estimator of β are zero.

☐ None of the above.

Note: The first two options were not specific enough, and so the credit for that part of the question was given to all answers. The total magnitude of the estimated β will decrease with an L_1 penalty, but some individual elements of β may stay constant or increase.

(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.

The minimum of β^3 is attained at $\beta = -\infty$, so minimizing empirical risk would always result in a degenerate solution.

(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.**

$$P(Y=0|X=x) \geq \frac{3}{4}$$

$$1 - P(Y=1|X=x) \geq \frac{3}{4}$$

$$P(Y=1|X=x) \leq \frac{1}{4}$$

$$\frac{1}{1+\exp(2+3x)} \leq \frac{1}{4}$$

$$1+\exp(2+3x) \geq 4$$

$$2+3x \geq \log 3$$

$$x \geq \frac{\log 3 - 2}{3}$$