DS 100/200: Principles and Techniques of Data Science Date: April 5, 2019

Discussion #9 Exam Prep Solutions

Name:

1. Of the choices below, why do we prefer to use ridge regression over linear regression (i.e. the normal equation) in certain cases? **Select all that apply.**

   ☐ A. Ridge regression always guarantees an analytic solution, but the normal equation does not.

   ☐ B. Ridge regression encourages sparsity in our model parameters, which is helpful for inferring useful features.

   ☐ C. Ridge regression isn't sensitive to outliers, which makes it preferable over linear regression.

   ☐ D. Ridge regression always performs just as well as linear regression, with the added benefit of reduced variance.

   ☐ E. None of the above

   Solution:

   ☐ A. The regularization term guarantees $(A^T A + \lambda I)$ is invertible, as discussed in discussion 7.

   ☐ B. This is the description for LASSO.

   ☐ C. It is sensitive to outliers.

   ☐ D. Doesn't always perform better.

2. Which of the following are indications that you should regularize? Select all that apply.

   ☐ A. Our training loss is 0.

   ☐ B. Our model bias is too high.

   ☐ C. Our model variance is too high.

   ☐ D. Our weights are too large.

   ☐ E. Our model does better on unseen data than training data.

   ☐ F. We have linearly dependent features.

   ☐ G. We are training a classification model and the data is linearly separable.

3. Suppose we have a data set which we divide into 3 equally sized parts, A, B, and C. We fit 3 linear regression models with L2 regularization (i.e. ridge regression), X, Y, and Z, all on A. Each model uses the same features and training set, the only difference is the λ used by each model. Select all below that are **always true**.

   ☐ A. Suppose Z has the lowest average loss on B. Model Z will have the lowest average loss when evaluated on C.

   ☐ B. If A and B have the same exact mean and variance, the average loss of model Y on B will be exactly equal to the average loss of Y on A.

   ☐ C. If $\lambda = 0$ for model X, $Loss(X, A) \leq Loss(Y, A)$ and $Loss(X, A) \leq Loss(Z, A)$.

   ☐ D. If $\lambda_Y < \lambda_Z$, then $Loss(Y, A) \leq Loss(Z, A)$.

   ☐ E. If $\lambda_Y > \lambda_Z$, then $Loss(Y, B) \geq Loss(Z, B)$.

   ☐ F. None of the above.

Solution:
A: Not guaranteed since we don't know the distributions of B, C.
B: Having the same mean and variance does not imply that the data are the same.
C: Since increasing λ increases bias, the loss of X must be less than or equal to the loss of Y, Z on A.
D: Since Y and Z were trained on A, and Y is less restricted than Z, the loss of Y on A must be less than the loss of Z on A. E: Even though Z is a more restricted (i.e. simpler) model, it is possible that the dataset B is slightly better for Z. In other words, minimizing training error with a regularized model does not guarantee minimized error on unseen datasets.

4. True or False.

   (a) A binary (0/1) classifier that always predicts 1 can get 100% precision, and its recall will be the fraction of ones in the training set.

   ◯ A. True ◯ B. False

   (b) If the training data is linearly separable we expect a logistic regression model to obtain 100% training accuracy.

   ◯ A. True ◯ B. False

   (c) We should use classification if the response variable is categorical.

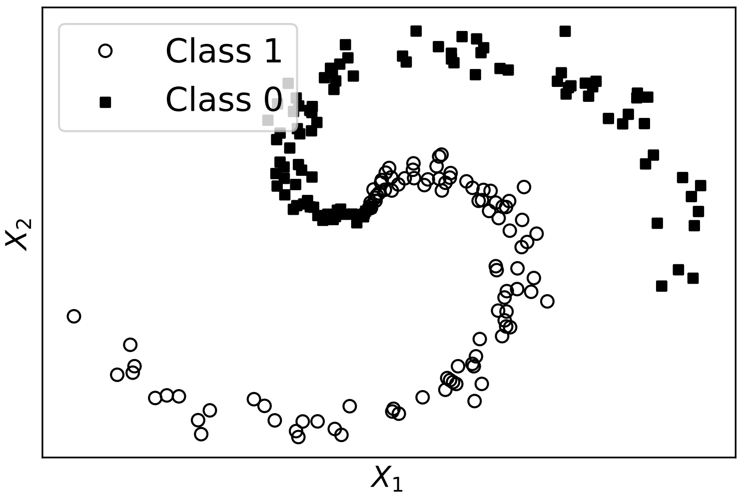
   ◯ A. True ◯ B. False

   (d) A binary classifier that only predicts class 1 may still achieve 99% accuracy on some prediction tasks.

   ◯ A. True ◯ B. False

5. The plot below is a scatter plot of a dataset with two dimensional features and binary labels (e.g., Class 0 and Class 1). Without additional feature transformations, is the this dataset linearly separable?

   ◯ A. Yes. ◯ B. No. ◯ C. We cannot tell that from this plot.



6. We perform a 4-fold cross validation on 4 different hyper-parameters, the mean square error are shown in the table below. Which λ should we select?

Fold Num	$\lambda = 0.1$	$\lambda = 0.2$	$\lambda = 0.3$	$\lambda = 0.4$	Row Max	Row Min	Row Avg
1	80.2	84.1	70.1	91.2	91.2	70.1	83.36
2	76.8	77.3	83.3	88.8	88.8	76.8	83
3	81.5	74.5	81.6	86.5	86.5	74.5	82.12
4	79.4	75.2	79.2	85.4	85.4	75.2	80.92
Col Avg	79.475	77.775	78.55	87.975			

   ◯ A. $\lambda = 0.1$ ◯ B. $\lambda = 0.2$ ◯ C. $\lambda = 0.3$ ◯ D. $\lambda = 0.4$

7. Answer **true** or **false** for each of the following statements about logistic regression:

   (a) If no regularization is used and the training data is linearly separable, the optimal model parameters will tend towards positive or negative infinity.

   ◯ A. True ◯ B. False

   (b) After using L^2 regularization, the optimal model parameter will be the mean of the data, since L^2 regularization is similar to the square loss.

   ◯ A. True ◯ B. False

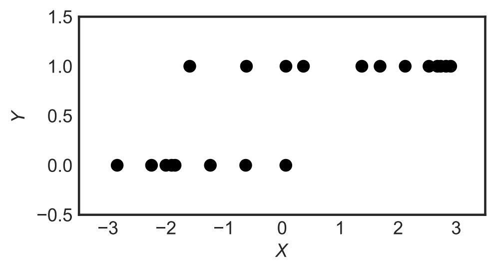
(c) L^1 regularization can help us select a subset of the features that are important.

◯ A. True ◯ B. False

(d) After using the regularization, we expect the training accuracy to increase and the test accuracy to decrease.

◯ A. True ◯ B. False

8. Suppose you are given the following dataset $\{(x_i, y_i)\}_{i=1}^n$ consisting of x and y pairs where the covariate $x_i \in \mathbb{R}$ and the response $y_i \in \{0, 1\}$.



Given this data, the value $\mathbb{P}(Y = 1 \mid x = -1)$ is likely closest to:

◯ A. 0.95 ◯ B. 0.50 ◯ C. 0.05 ◯ D. -0.95

9. Suppose we train a binary classifier on some dataset. Suppose y is the set of true labels, and $\hat{y}$ is the set of predicted labels.

y	0	0	0	0	0	1	1	1	1	1
$\hat{y}$	0	1	1	1	1	1	1	0	0	0

Determine each of the following quantities.

(a) The number of true positives

Solution: 2

(b) The number of false negatives

Solution: 3

(c) The precision of our classifier. Write your answer as a simplified fraction.

Solution: $\frac{2}{2+4} = \frac{1}{3}$

10. You have a classification data set, where x is some value and y is the label for that value:

x	y
2	1
3	0
0	1
1	0

Suppose that we're using a logistic regression model to predict the probability that $Y = 1$ given x:

$$\mathbb{P}(Y = 1|x) = \sigma(\phi^T(x)\theta)$$

(a) Suppose that $\phi(x) = [\phi_1 \quad \phi_2 \quad \phi_3]^T = \begin{bmatrix}1 & x & x^2\end{bmatrix}^T$ and our model parameters are $\theta^* = \begin{bmatrix}1 & 0 & -2\end{bmatrix}^T$. For the following parts, leave your answer as an expression (do not numerically evaluate log, e, π, etc).

i. Compute $\hat{\mathbb{P}}(y = 1|x = 0)$.

> **Solution:** $\frac{1}{1+\exp(-1)}$

ii. What is the loss for this single prediction $\hat{\mathbb{P}}(y = 1|x = 0)$, assuming we are using KL divergence as our loss function (or equivalently that we are using the cross entropy as our loss function)?

> **Solution:** $\log(1 + \exp(-1))$

(b) Suppose $\phi(x) = \begin{bmatrix}1 & x & x\%2\end{bmatrix}^T$, where % is the modulus operator. Are the data from part a linearly separable with these features? If so, give the equation for a separating plane, e.g. $\phi_2 = 3\phi_3 + 1$. Use 1-indexing, e.g. we have ϕ_1, ϕ_2, and ϕ_3. If not, just write "no".

> **Solution:** Yes, they can be separated by the hyperplane $\phi_3 = 0.5$.

11. Suppose we have the dataset below.

x	y
1	1
-1	0

Suppose we have the feature set $\phi(x) = [\phi_1 \quad \phi_2]^T = [1 \quad x]^T$. Suppose we use gradient descent to compute the θ which minimizes the KL divergence under a logistic model without regularization, i.e.

$$\arg\min_\theta -\frac{1}{n}\sum_{i=1}^{n}(y_i\phi(x_i)^T + log(\sigma(-\phi(x_i)^T\theta)))$$

Select all that are true regarding the data points and the optimal theta value θ.

☐ A. The data is linearly separable.

☐ B. The optimal θ yields an average cross entropy loss of zero.

☐ C. The optimal θ diverges to $-\infty$

☐ D. The optimal θ diverges to $+\infty$

☐ E. The equation of the line that separates the 2 classes is $\phi_2 = 0$.

☐ F. None of the above.

Solution:

☐ A. True. When drawn in the 2-D feature space, the points are linearly separable.

☐ B. True. If the data is linearly separable, we can achieve an average cross entropy loss of zero and our parameter value θ will diverge.

☐ C. False. The optimal theta value θ diverges to $+\infty$

☐ D. True. The optimal theta value θ diverges to $+\infty$

☐ E. True. If we draw the line $\phi_2 = 0$ in the 2-D feature space, this separates the points.

☐ F. False. 4 choices were true above.

12. Suppose we have the dataset below.

x	y
-3	1
-1	0
1	0
3	1

Suppose we have the feature set $\phi(x) = \begin{bmatrix}1 & x^2\end{bmatrix}^T$. Suppose we use gradient descent to compute the θ which minimizes the KL divergence under a logistic model without regularization, i.e.

$$\arg\min_{\theta} -\frac{1}{n}\sum_{i=1}^{n}(y_i\phi(x_i)^T + log(\sigma(-\phi(x_i)^T\theta)))$$

(a) Explain in 10 words or fewer why the magnitudes of θ_1 and θ_2 will be very large.

> **Solution:** Because the data is linearly separable.

(b) Will the sign of θ_2 be negative or positive?

- ◯ A. Could be either, it depends on where our gradient descent starts
- ◯ B. Positive
- ◯ C. Negative
- ◯ D. Neither, θ_2 will be zero

(c) If we use L_1 regularization, which of our θ values would you expect to be zero?

- ◯ A. Neither of them
- ◯ B. θ_1
- ◯ C. θ_2
- ◯ D. Both θ_1 and θ_2