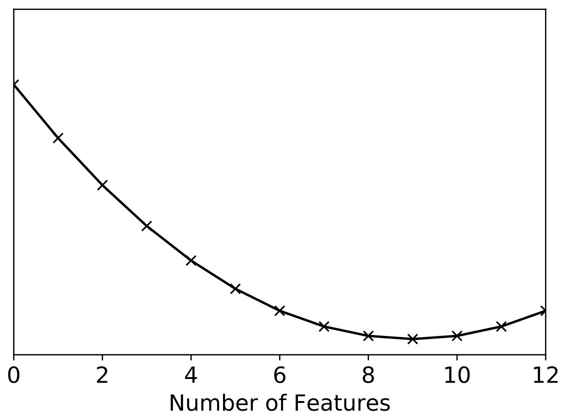
DS 100/200: Principles and Techniques of Data Science Date: March 22, 2019

Discussion #8 Exam Prep Solutions

Name:

1. In the process of training linear models with different numbers of features you created the following plot but forgot to include the Y-axis label.



   (a) The Y-axis might represent the training error: ◯ A. True ◯ B. False

   (b) The Y-axis might represent the bias: ◯ A. True ◯ B. False

   (c) The Y-axis might represent the test error: ◯ A. True ◯ B. False

   (d) The Y-axis might represent the variance. ◯ A. True ◯ B. False

2. Consider the following model training script to estimate the training error:

```
X_train, X_test, y_train, y_test =
    train_test_split(X, y, test_size=0.1)

model = lm.LinearRegression(fit_intercept=True)
model.fit(X_test, y_test)

y_fitted = model.predict(X_train)
y_predicted = model.predict(X_test)

training_error = rmse(y_fitted, y_predicted)
```

(a) **Line 5** contains a serious mistake. Assuming our eventual goal is to compute the *training error*, which of the following corrects that mistake.

◯ A. `model.fit(X_train, y_test)`

◯ B. `model.fit(X_train, y_train)`

◯ C. `model.fit(X, y)`

(b) **Line 10** contains a serious mistake. Assuming we already have corrected the mistake in **Line 5** which of the following corrects the mistake on **Line 10**.

◯ A. `training_error = rmse(y_train, y_predicted)`

◯ B. `training_error = rmse(y_train, y_test)`

◯ C. `training_error = rmse(y_fitted, y_test)`

◯ D. `training_error = rmse(y_fitted, y_train)`

3. Which of the following techniques could be used to reduce over-fitting?

◯ A. Adding noise to the training data
◯ B. Cross-validation to remove features
◯ C. Fitting the model on the test split
◯ D. Adding features to the training data

4. Suppose you are given a dataset $\{(x_i, y_i)\}_{i=1}^n$ where $x_i \in \mathbb{R}$ is a one dimensional feature and $y_i \in \mathbb{R}$ is a real-valued response. To model this data, you choose a model characterized by the following loss function:

$$L(\theta) = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \theta_0 - x_i^3 \theta_1\right)^2 + \lambda |\theta_1| \tag{1}$$

For the following statements, indicate whether it is True or False.

(a) This model includes a bias/intercept term.

◯ A. True ◯ B. False

(b) As λ decreases to smaller values, the model will reduce to a constant θ_0

◯ A. True ◯ B. False

(c) Larger λ values help reduce the chances of overfitting.

◯ A. True ◯ B. False

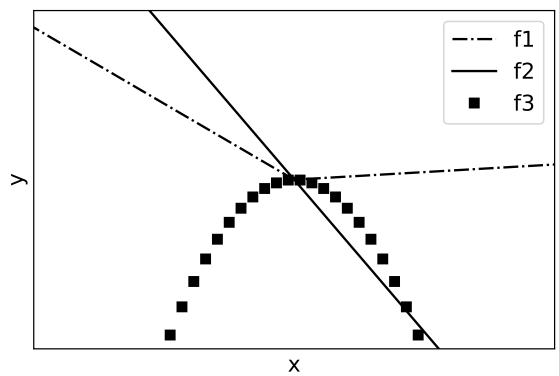
(d) Increasing λ decreases model variance.

◯ A. True ◯ B. False

(e) The training error should be used to determine the best value for λ.

◯ A. True ◯ B. False

5. Use the following plot to answer each of the following questions about convexity:



(a) $f_1(x) = \max(0.01x, -x)$ is convex. ◯ A. True ◯ B. False

(b) $f_2(x) = -2x$ is convex. ◯ A. True ◯ B. False

(c) $f_3(x) = -x^2$ is convex. ◯ A. True ◯ B. False

(d) $f_4(x) = f_1(x) + f_2(x)$ is convex. ◯ A. True ◯ B. False

6. In class, we showed that the expected squared error can be decomposed into several important terms:

$$\mathbb{E}[(Y - f_{\hat{\theta}}(x))^2] = \sigma^2 + (h(x) - \mathbb{E}[f_{\hat{\theta}}(x)])^2 + \mathbb{E}[(\mathbb{E}[f_{\hat{\theta}}(x)] - f_{\hat{\theta}}(x))^2].$$

(a) For which of the following reasons are we taking an expectation? In other words, what are the sources of randomness that we are considering in the derivation of the bias-variance tradeoff?

☐ A. We chose arbitrary features when doing feature engineering.

☐ B. We drew random samples from some larger population when we built our training set.

☐ C. There is some noise in the underlying process that generates our observations Y from our features.

☐ D. Our x values could have had missing or erroneous data, e.g. participants misreading a question on a survey.

☐ E. None of the Above.

(b) Which of the following do we treat as fixed? Select all that apply.

☐ A. $\hat{\theta}$

☐ B. σ^2

☐ C. $h(x)$

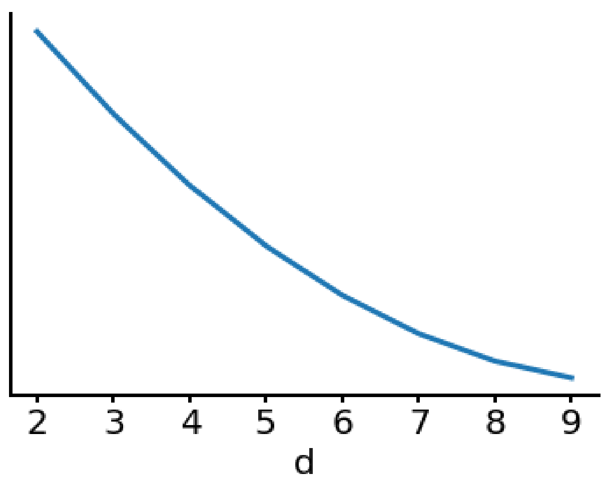
(c) By decreasing model complexity, we are able to decrease σ^2.

◯ A. True

◯ B. False

7. Your team would like to train a machine learning model in order to predict the next YouTube video that a user will click on based on m features for each of the previous d videos watched by that user. In other words, the total number of features is $m \times d$. You're not sure how many videos to consider.

(a) Your colleague generates the following plot, where the value d is on the x axis. However, they forgot to label the y-axis.



Which of the following could the y axis represent? Select all that apply.

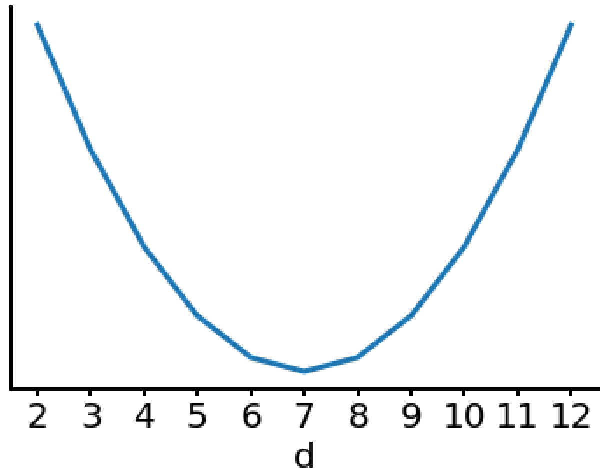
☐ A. Training Error

☐ B. Validation Error

☐ C. Bias

☐ D. Variance

(b) Your colleague generates the following plot, where the value d is on the x axis. However, they forgot to label the y-axis.



Which of the following could the y axis represent? Select all that apply.

☐ A. Training Error

☐ B. Validation Error

☐ C. Bias

☐ D. Variance

8. Elastic Net is a regression technique that combines L_1 and L_2 regularization. It is preferred in many situations as it possesses the benefits of both LASSO and Ridge Regression. Minimizing the L2 loss using Elastic Net is as follows, where $\lambda_1, \lambda_2 >= 0$, $\lambda_1 + \lambda_2 = \lambda$, $\lambda > 0$.

$$\hat{\theta} = \arg\min_{\theta} \frac{1}{n} \sum_{i} (y_i - \theta x)^2 + \lambda_1 \sum_{j=1}^{p} |\theta_j| + \lambda_2 \sum_{j=1}^{p} \theta_j^2$$

Suppose our goal was to get sparse parameters, i.e. we want as many parameters as possible to be zero. Which of the following choices for λ_1, λ_2 are most consistent with this goal, assuming $\lambda = 1$? **There is only one correct answer.**

◯ A. $\lambda_1 = 0, \lambda_2 = 1$

◯ B. $\lambda_1 = 0.5, \lambda_2 = 0.5$

◯ C. $\lambda_1 = 1, \lambda_2 = 0$

Solution: We know that LASSO encourages sparsity in our optimal weights. Setting λ_1 to 1 means we are using LASSO.

9. What happens to bias and variance as we increase the value of λ? Assume $\lambda_2 = \lambda_1$. **There is only one correct answer in each part.** You will be asked to justify why in the next question.

(a) Bias:

◯ A. Bias goes up
◯ B. Bias stays the same
◯ C. Bias goes down

(b) Variance:

◯ A. Variance goes up
◯ B. Variance stays the same
◯ C. Variance goes down

10. Justify why by marking the true statements. **Select all that apply for each part.**

(a) Bias:

☐ A. Bias goes down because increasing λ reduces over fitting.
☐ B. Bias goes down because bias is minimized when $\lambda_2 = \lambda_1$.
☐ C. Bias goes up because increasing λ penalizes complex models, limiting the set of possible solutions.
☐ D. Bias goes up because the loss function becomes non-convex for sufficiently large λ.
☐ E. None of the above

(b) Variance:

☐ A. Variance goes down because increasing λ encourages the value of the loss to decrease.
☐ B. Variance goes down because increasing λ penalizes large model weights.
☐ C. Variance goes up because because increasing λ increases bias.
☐ D. Variance goes up because increasing λ increases the magnitude of terms in the loss function.
☐ E. None of the above

11. What happens to the model parameters $\hat{\theta}$ as $\lambda \to \infty$, i.e. what is $\lim_{\lambda \to \infty} \hat{\theta}$? **Select all that apply.**

☐ A. Converge to 0.
☐ B. Diverge to infinity.
☐ C. Converge to values that minimize the L2 loss.
☐ D. Converge to equal but non-zero values.
☐ E. Converge to a sparse vector.

Solution: The model parameters go to 0.