CSE 446/546: Solutions

Please WAIT to open the exam until you are instructed to begin. You can write your name on this page.

Please write your name and ID on your notes page (if you have one). We will collect this with your exam.

Please take out your student ID and leave it on the corner of your desk, as we will come around and check them while you work on the exam.

Instructions: This exam consists of 40 True/False and multiple choice questions.

Write your name and ID number in the provided spaces on every page of the exam.

For each question, clearly indicate your answer by filling in the letter associated with your choice.

If you need to change an answer, please very clearly indicate what your final answer is. Responses where we cannot determine the selected option will be marked as incorrect.

(1) Scientists tells us that there is a 10% probability that a person will have the flu this winter. Meanwhile, the CDC reports that that 30% of the population will experience flu-like symptoms, and that 60% of people with the flu will be symptomatic. What's the probability that Sarah has the flu given that she has flu-like symptoms?

(A) 60%

(B) 20%

(C) 30%

(D) 10%

Solution:

The solution is (B).

(2) True/False: The variance of a model typically decreases as the number of features increases.

(A) True

(B) False

Solution:

The solution is (B)

(3) Assume you're given two independent random variables X and Y. X is uniformly distributed on the interval $[1, 3]$, whereas Y follows a normal distribution with mean 3 and standard deviation 1. What is $(E[XY])^2 - E[X]E[Y]$?

(A) 3

(B) 30

(C) 6

(D) 0

Solution:

The solution is (B).

(4) True/False: If the columns of A are orthogonal, then $A^T A$ is diagonal.

(A) True

(B) False

Solution:

The answer is (A).

(5) True/False: Assume we train a model on a given dataset. If we were to remove 50% of samples from the dataset and re-train the model from scratch, the new model will be more likely to overfit to its training data than the old one.

(A) True

(B) False

Solution:

The solution is (A)

(6) True/False: If $\{v_1, v_2, \cdots, v_n\}$ and $\{w_1, w_2, \cdots, w_n\}$ are linearly independent, then $\{v_1 + w_1, v_2 + w_2, \cdots, v_n + w_n\}$ are linearly independent.

(A) True

(B) False

Solution:

The answer is (B).

(7) True/False: $\mathbb{E}[\epsilon\epsilon^T] = I$ where $\epsilon_i \sim N(0, \sigma^2)$ such that ϵ is a column vector: $\epsilon \in \mathbb{R}^d$.

(A) True

(B) False

Solution:

The answer is (B).

Figure 1: The following graphic will be used as a representation of bias and variance. Imagine that a true/correct model is one that always predicts a location at the center of each target (being farther away from the center of the target indicates that a model's predictions are worse). We retrain a model multiple times, and make a prediction with each trained model. For each of the targets, determine whether the bias and variance is low or high with respect to the true model.

(8) In Figure 1, subplot I, how are bias and variance related to the true model?

(A) High bias, High variance

(B) High bias, Low variance

(C) Low bias, High variance

(D) Low bias, Low variance

Solution:

The solution is (D).

(9) In Figure 1, subplot II, how are bias and variance related to the true model?

(A) High bias, High variance

(B) High bias, Low variance

(C) Low bias, High variance

(D) Low bias, Low variance

Solution:

The solution is (C)

(10) In Figure 1, subplot III, how are bias and variance related to the true model?

(A) High bias, High variance

(B) High bias, Low variance

(C) Low bias, High variance

(D) Low bias, Low variance

Solution:

The solution is (B)

(11) In Figure 1, subplot IV, how are bias and variance related to the true model?

(A) High bias, High variance

(B) High bias, Low variance

(C) Low bias, High variance

(D) Low bias, Low variance

Solution:

The solution is (A)

(12) Let $x_1, x_2 \in \mathbb{R}$ be sampled from the distribution $\mathcal{N}(\mu, 1)$, where $\mu \in \mathbb{R}$ is an unknown variable. Remember that the PDF of the normal distribution is $f(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{(x-\mu)^2}{2}}$. Using the log-likelihood, find the maximum likelihood estimation of μ in terms of x_1, x_2.

(A) $\frac{2}{x_1+x_2}$

(B) $\log\left(\frac{e^{x_1}+e^{x_2}}{2}\right)$

(C) $\frac{\log(x_1)+\log(x_2)}{2}$

(D) $\frac{x_1+x_2}{2}$

Solution:

The answer is (D).

(13) Suppose our data distribution has the property that $y_i = \beta x_i + c + \epsilon_i$ for $x_i, \beta \in \mathbb{R}^d$, $c \in \mathbb{R}$, $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. Suppose we learn a model

$$\hat{\beta} = \text{argmin}_\gamma ||\gamma X - y||_2^2.$$

True/False: : $\hat{\beta}$ is an unbiased estimate of β.

(A) True

(B) False

Solution:

(B)

(14) How will regularizing the weights in a linear regression model change the bias and variance (relative to the same model with no regularization)?

(A) Increase bias, increase variance

(B) Increase bias, decrease variance

(C) Decrease bias, increase variance

(D) Decrease bias, decrease variance

Solution:

The solution is (B).

(15) True/False: Given a fixed training set, the training loss is never larger in a polynomial regression of degree $d+1$ than in one of degree d, where $(d \geq 1)$.

(A) True

(B) False

Solution:

The solution is (A)

(16) True/False: Given both a train and test set, the test loss is always lower in a polynomial regression of degree $d+1$ than in one of degree d, where $(d \geq 1)$.

(A) True

(B) False

Solution:

The solution is (B)

(17) On which factor does the value of irreducible error depend in linear regression?

(A) n, the number of observations in the training set

(B) m, the dimension of features in the training set

(C) σ^2, the variance of the noise

(D) None of those

Solution:

The solution is (C)

(18) How does the irreducible error change if we increase the regularization coefficient λ in ridge regression?

(A) Increase

(B) Decrease

(C) Not change

(D) The answer depends on the dataset X and true weights w^*.

Solution:

The solution is (C)

(19) True/False: k-fold cross-validation with $k = 100$ is computationally more expensive (slower) than "leave-one-out" cross validation. (Assume that there are enough data points to divide the dataset evenly by k.)

(A) True

(B) False

Solution:

The solution is (B)

(20) Assume we have a data matrix X. Which of the following is a true statement when comparing leave-one-out cross validation (LOOCV) error with the true error?

(A) LOOCV error is typically a slight underestimation of the true error of a model trained on X.

(B) LOOCV error is typically a slight overestimation of the true error of a model trained on X.

(C) LOOCV error is an unbiased estimator of the true error of a model trained on X.

Solution:

The solution is (B)

(21) True/False: LASSO is a *convex* optimization problem.

(A) True

(B) False

Solution:

The answer is (A).

(22) In LASSO regression, if the regularization parameter $\lambda = 0$, then which of the following is true?

(A) This LASSO model can be used for feature selection.

(B) The loss function is as same as the ridge regression loss function.

(C) The loss function is as same as the ordinary least square loss function.

(D) Large coefficients are penalized.

Solution:

The solution is (C).

(23) In a LASSO Regression, if the regularization parameter λ is very high, which of the following is true?

(A) The model can shrink the coefficients of uninformative features to exactly 0

(B) The loss function is as same as the ordinary least square loss function.

(C) The loss function is as same as the ridge regression loss function

(D) The bias of the model is no lower than the bias of the model with a smaller λ.

Solution:

The solution is (A).

(24) True/False: In LASSO regression, if the regularization parameter λ is very large and two informative features are highly collinear (i.e., that there exists an α such that $x_{ij} \approx \alpha x_{ij'}$ for all $i \in [n]$) , then LASSO will assign one of those coefficients to zero while ridge regression never will.

(A) True

(B) False

Solution:

The solution is (A).

(25) For ridge regression, if the regularization parameter is too large, which of the following is true?

(A) Large coefficients will not be penalized

(B) The model will underfit the data

(C) The loss function will be the as same as the ordinary least square loss function

(D) The model will overfit the data

Solution:

The solution is (B).

(26) True/False: For any convex function $f : \mathbb{R} \to \mathbb{R}$, for any $x \in \mathbb{R}$, any $\lambda \in (0, 1)$ we have that:

$$f(\lambda x_1 + (1 - \lambda)x_2) \leq (1 - \lambda)f(x_1) + \lambda f(x_2)$$

(A) True

(B) False

Solution:

The answer is (B).

(27) True/False: All local minimizers for a convex function f are global minimizers for f.

(A) True

(B) False

Solution:

The answer is (A).

(28) Which function is not a convex function?

(A) Sigmoid/Logistic function: $f(x) = 1/(1 + e^{-x})$.

(B) Linear function: $f(x) = 3x$.

(C) Square function: $f(x) = x^2$.

(D) ReLU function: $f(x) = \max\{x, 0\}$.

Solution:

The answer is (A).

(29) Which of the following is not a convex set?

(A) Unit ball: $\{x \in \mathbb{R}^2 | \ \|x\|_2^2 \leq 1\}$.

(B) Unit sphere: $\{x \in \mathbb{R}^2 | \ \|x\|_2^2 = 1\}$.

(C) Unit cube: $\{x \in \mathbb{R}^2 | \ 0 \leq x_1 \leq 1, 0 \leq x_2 \leq 1\}$.

(D) Line: $\{x \in \mathbb{R}^2 | \ x_1 + x_2 = 1\}$.

Solution:

The answer is (B).

(30) True/False: We use stochastic gradient descent instead of gradient descent in order to speed up per-iteration computation at the expense of more variance.

(A) True

(B) False

Solution:

The answer is (A).

(31) Which of the shapes shown in Figure 2 is a convex shape?

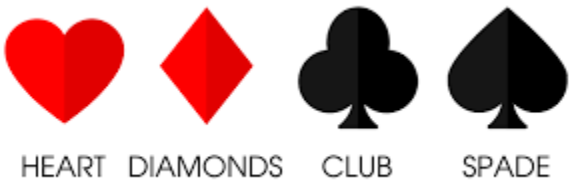

Figure 2: Shapes for question (31).

(A) Heart

(B) Diamonds

(C) Club

(D) Spade

Solution:

The answer is (B).

(32) Which of the following is *not* a true statement about gradient descent (GD) vs. stochastic gradient descent (SGD)?

(A) Both provide unbiased estimates of the true gradient at each step.

(B) The memory and compute requirements of a single update step for both methods scales linearly with the number of features.

(C) The memory and compute requirements of a single update step for both methods scales linearly with the number of data points.

(D) GD is likely to converge in fewer updates/iterations than SGD, with a properly selected learning rate.

Solution:

The solution is (C).

(33) Which of the following is a true statement about gradient descent (GD)?

(A) When training, we should not update the bias (aka offset, or intercept) term using GD.

(B) Decreasing the learning rate, keeping all other hyperparameters fixed, guarantees that the error of our estimated parameters will decrease.

(C) GD can be expensive to run on datasets with a large number of samples.

(D) An advantage of GD over SGD is that GD requires only a single update step to converge.

Solution:

The solution is (C).

(34) True/False: The bias of a model is defined as the expected difference between the prediction $\hat{y}$ and the true value y.

(A) True

(B) False

Solution:

The solution is (B). However, we accepted either answer due to ambiguity in the question.

(35) True/False: Consider the sets of features $S \subseteq S'$. True or false: the bias of the model trained on features in S' is no larger than the bias of the model trained on features in S.

(A) True

(B) False

Solution:

The solution is (A).

(36) True/False: The cross-validation error is a better estimate of the true error than the training error.

(A) True

(B) False

Solution:

The solution is (A).

(37) True/False: If a model is trained with "leave-one-out" cross validation, then the expected error of the model on unseen data is equal to the training error of the model.

(A) True

(B) False

Solution:

The solution is (B).

(38) Write down a closed form solution for the optimal parameters w that minimize the loss function

$$L(w) = \sum_{i=1}^{n} (y_i - x_i^T w)^2$$

in terms of the $n \times d$ matrix X whose i-th row is x_i^T and the n by 1 vector y whose i-th entry is y_i. (You may assume that any relevant matrix is invertible.)

(A) $\widehat{w} = 2(X^T X)^{-1} X^T y$

(B) $\widehat{w} = (X^T X)^{-1} X^T y$

(C) $\widehat{w} = (X^T X)^{-1} X y$

(D) $\widehat{w} = (X X^T)^{-1} X^T y$

Solution:

The solution is (B).

(39) True/False: Let $x_1, \ldots, x_n \in \mathbb{R}^+$ be sampled i.i.d. from the distribution $\text{Exp}(\theta) = \theta e^{-\theta x}$, where $\theta \in \mathbb{R}^+$ is an unknown variable. By analyzing the log-likelihood, what is the maximum likelihood estimation of θ (in terms of the samples)?

(A) $\frac{1}{n} \prod_{i=1}^{n} x_i$

(B) $\frac{1}{n} \sum_{i=1}^{n} x_i$

(C) $n/(\sum_{i=1}^{n} x_i)$

(D) $-n/(\sum_{i=1}^{n} x_i)$

Solution:

The answer is (C).

(40) True/False: In the least-squares linear regression setting, if we double the data matrix X, we double the resulting least squares solution $\hat{w}$.

(A) True

(B) False

Solution:

The answer is (B).