CSE 446/546: Practice Midterm Exam Questions

The exam consists of True/False and multiple choice questions.

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

Please note that these examples are designed to give you a sense of the general format and topics that could be covered on the exam, but it is NOT a comprehensive study guide and should not be considered a substitute for studying the rest of the material covered in the course so far.

(1) Assume there are a 2 bags of 10 marbles each, which can be either blue or red. One of the bags contains 10 blue marbles, whereas the other bag contains 5 blue and 5 red marbles. You can only look at marbles once you draw them from a bag. You randomly select a bag with 50% probability without looking inside, and draw a blue marble from it. What's the probability that the second marble you draw from that same bag will also be blue?

(A) 70%

(B) 45%

(C) 40%

(D) 85%

(2) What is the minimum number of people in a room needed to have at least a 50% probability that at least 2 people have the same month of birth (Assuming that everyone is equally likely to be born in any of the twelve months of the year)? Note: we won't have something that requires a calculator on the exact exam, but this sort of probability question is fair game.

(A) 4

(B) 5

(C) 6

(D) 7

(3) Suppose we consider $\tilde{X} = X_{\cdot,1:d-1}$, the first $d-1$ columns of X. Suppose we learn a model (ignoring the last column)

$$\hat{\tilde{\beta}}, c = \text{argmin}_{\gamma,c} ||\gamma\tilde{X} + c - Y||_2^2.$$

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

(A) True

(B) False

(4) True/False: The bias of a model always decreases as the amount of training data available increases.

(A) True

(B) False

(5) Which modification reduces the irreducible error?

(A) Get a hold of a larger training set.

(B) Add additional functions to your hypothesis class (e.g., switch from linear functions to cubic polynomials.)

(C) Add a new feature but keep the distribution of the noise unchanged

(D) None of those

(6) True/False: Imagine we are creating a model to predict Seattle air quality. In order to improve our model performance, we should create an validation dataset from test dataset for hyperparameter tuning.

(A) True

(B) False

(7) To prevent overfitting of our linear model, we would apply regularization on the model. Which of the following datasets should we evaluate using our model in order to decide the right amount of regularization?

(A) Train

(B) Validation

(C) Test

(D) All of the above

(8) The L1 penalty in a LASSO regression is equivalent to what prior on its weights?

(A) Gaussian

(B) Laplacian

(C) Uniform

(9) If we believe only a small number of available features are useful in building a linear predictor for y, what regularization penalty should we use?

(A) No regularization

(B) L_1

(C) L_2

(D) Constrain our weights to have $||w||_2 \leq \lambda$.

(10) You estimate a ridge regression model with some data taken from an experiment, and find (using cross-validation) an optimal ridge penalty λ_1. You then buy a new sensor which has noise with 1/4 the variance as before. Using the same number of observations as before you collect new data, you find a new optimal ridge penalty λ_2. Which of the following will be closest to true?

(A) $\lambda_1/\lambda_2 = 1/4$

(B) $\lambda_1/\lambda_2 = 1/2$

(C) $\lambda_1/\lambda_2 = 1$

(D) $\lambda_1/\lambda_2 = 2$

(E) $\lambda_1/\lambda_2 = 4$

(11) True/False: Ridge regression is "scale invariant" in the sense that the test set prediction accuracy is unchanged if one rescales the features.

(A) True

(B) False

(12) The L2 penalty in a ridge regression is equivalent to what prior on its weights?

(A) Gaussian

(B) Laplace

(C) Uniform

(13) True/False: Ridge regression can shrink all coefficients to **exactly** 0 if the regularization parameter λ is large enough.

(A) True

(B) False

(14) In a ridge regression model parameterized by w, what is the penalty term?

(A) The square of the magnitude of w's coefficients

(B) The square root of the magnitude of w's coefficients

(C) The absolute sum of w's coefficients

(D) The sum of w's coefficients

(15) Which of the following is *not* a true statement about stochastic gradient descent (SGD)?

(A) It is possible to use the same data example across different iterations.

(B) Stochastic gradient descent is faster because it has a higher convergence rate in optimization.

(C) Stochastic gradient descent can create vastly different gradients in the first few steps comparing to those in gradient descent.