CSE 446/546 Spring 2023 Midterm Exam Solutions

May 1, 2023

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

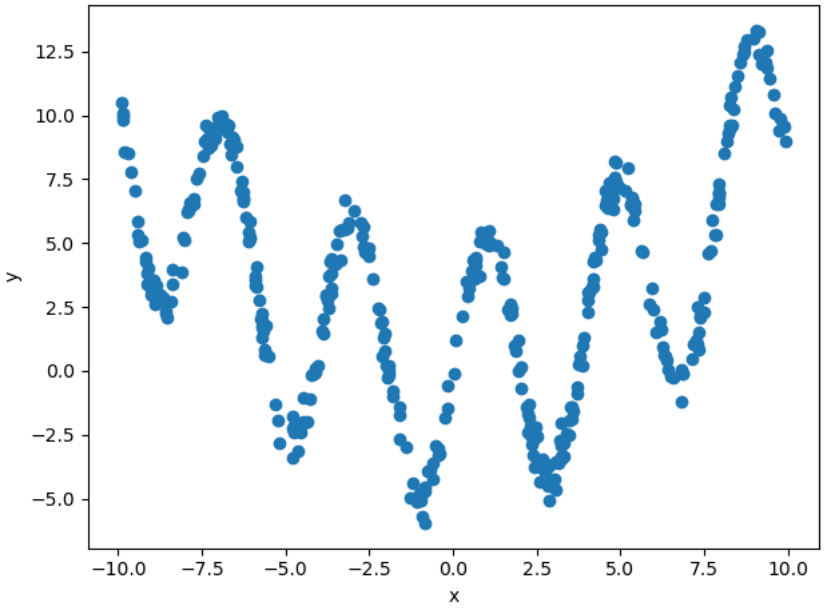
Please write your name and ID on your notes page (if you have one). We will collect this with your exam.

Please take out your Husky Card and have it accessible when you turn in your exam.

Instructions: This exam consists of a set of short questions (True/False, multiple choice, short answer).

- Write your name and ID number in the provided spaces on every page of the exam.
- For each multiple choice and True/False question, clearly indicate your answer by filling in the letter(s) associated with your choice.
- For each short answer question, please write your answer in the provided space.
- If you need to change an answer or run out of space, please very clearly indicate what your final answer is and what you would like graded. Responses where we cannot determine the selected option will be marked as incorrect.
- Please remain in your seats for the last 10 minutes of the exam. If you complete the exam before the last 10 minutes, you may turn in your exam and note sheet by handing them to a TA.

1. We need to fit a function to our dataset $\{(x_i, y_i)\}_{i=1}^n$. Suppose our dataset looks like the following:



We decide to expand our features with general basis functions to improve our estimator: $\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \rightarrow \begin{bmatrix} x_1 & g(x_1) & h(x_1) \\ \vdots & \vdots & \vdots \\ x_n & g(x_n) & h(x_n) \end{bmatrix}$. Which of the following choices of g and h are most likely to produce the best estimator function?

(a) $g(x) = \log(x), h(x) = x^2$

(b) $g(x) = x^4, h(x) = x^2$

(c) $g(x) = \sin(x), h(x) = x^2$

(d) $g(x) = \cos(x), h(x) = x$

Correct answers: (c)

Explanation: The answer is $g(x) = \sin(x), h(x) = x^2$. $g(x) = \log(x)$ does not exist for $x < 0$, so answer (A) is incorrect. A degree-4 polynomial is not complex enough to represent this data so answer (B) is incorrect. A different sinusoidal function like $g(x) = \cos(x)$ could be a good choice, but $h(x) = x$ does not represent the general parabolic shape of the sinusoid as well as $h(x) = x^2$ does, so answer $g(x) = \cos(x), h(x) = x$ is incorrect.

2. Irreducible error can be completely eliminated by:

   (a) Collecting more training data
   (b) Tuning hyperparameters of the model
   (c) Regularizing the model
   (d) None of the above

   Correct answers: (d)

3. Increasing the regularization of a model would typically:

   (a) Increase its bias and increase its variance
   (b) Increase its bias and decrease its variance
   (c) Decrease its bias and increase its variance
   (d) Decrease its bias and decrease its variance

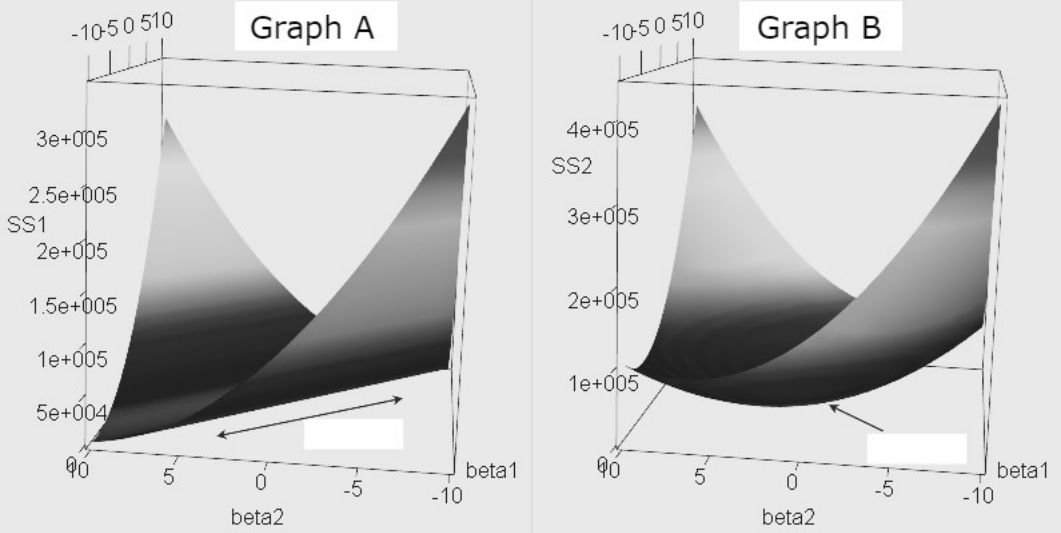
   Correct answers: (b)

4. In a binary classification problem with balanced classes (exactly the same number of positive examples as negative examples), a machine learning model has an accuracy of 85% and misclassifies 10% of positive examples as negative. What is the probability that the model will correctly classify a negative sample?

   Answer: _____

   Explanation: 80%.

5. The below figures are graphs of some loss functions with Loss on the Vertical axis and weight variables on the horizontal axes.



Which graph represents a Ridge Regression Loss function?

(a) Graph A

(b) Graph B

Correct answers: (b)

6. Irreducible error in machine learning is caused by:

(a) Noise in the data

(b) Bias in the model

(c) Variance in the model

(d) Overfitting of the model

Correct answers: (a)

7. Suppose that we are given train, validation, and test sets. Which set(s) should be used to standardize the test data when generating a prediction?

   Answer: _____

   Explanation: We should standardize the input data using the mean and standard deviation from the training data. If we use the mean and standard deviation from the test data, we are using extra information (outside of the training data) to make predictions. Consequently, our predictions fit to, and are dependent on, the test set (e.g. if we use 5 or 10 testing samples, we would generate different predictions), known as "data leakage". (Also accept mean and standard deviation from train and validation sets combined.)

8. Suppose we are performing leave-one-out (LOO) validation and 10-fold cross validation on a dataset of size $100,000$ to pick between 4 different values of a single hyperparameter. How many times greater is the number of models that need to be trained for LOO validation versus 10-fold cross validation?

   Answer: _____

   Explanation: The answer is $10,000$.

9. What are two possible ways to reduce the variance of a model?

   Answer: _____

   _____

   Explanation: Two possible responses: (1) Use more training data. (2) Use a less complex model.

10. Below are a list of potential advantages and disadvantages of stochastic gradient descent(SGD). Select **all** that are true regarding SGD.

    Advantages:

    (a) SGD is more memory-efficient because it takes a smaller number of samples at a time compared to classical gradient descent which takes the entire dataset into weight update

    (b) In SGD, the update on weight w_{t+1} has lower variance because it doesn't take many samples into account at a time

    Disadvantages:

    (c) The noise in the dataset has higher impact on the stability of SGD than on that of the classical gradient descent.

    (d) SGD is more sensitive to learning rate compared to classical gradient descent

    (e) It's more computationally inefficient to use SGD for a large dataset than to use classical gradient descent because it requires more resources to randomly sample a data point for the weight update

    Correct answers: (a), (c), (d)

    Explanation: Note: option (d) (SGD is more sensitive to learning rate compared to classical gradient descent) was deemed unclear and we accepted either (a, c) or (a, c, d) as correct.

11. Which of the following is not a convex function?

    (a) $f(x) = x$

    (b) $f(x) = x^2$

    (c) $f(x) = e^x$

    (d) $f(x) = \frac{1}{1+e^{-x}}$

    Correct answers: (d)

12. Recall the loss function used in ridge regression,

$$f(w) = \sum_{i=1}^{n} \left(y_i - x_i^T w\right)^2 + \lambda ||w||_2^2$$

What happens to the weights as $\lambda \to \infty$?

(a) Weights approach positive infinity.
(b) Weights approach 0.
(c) Weights approach negative infinity.
(d) Not enough information.

Correct answers: (b)

13. Why is it important to use a different test set to evaluate the final performance of the model, rather than the validation set used during model selection?

(a) The model may have overfit to the validation set
(b) The test set is a better representation of new, unseen data
(c) Both a and b
(d) None of the above

Correct answers: (c)

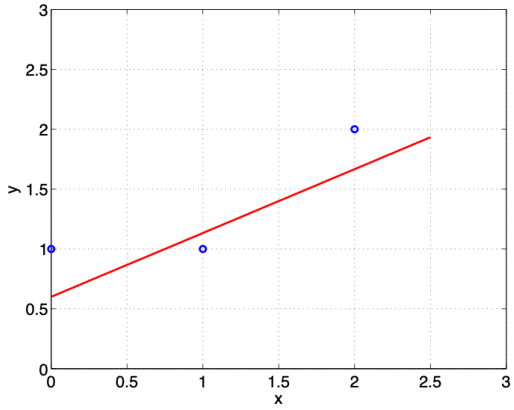
14. What is cross-validation **not** used for?
(a) To evaluate the performance of a machine learning model on unseen data.
(b) To select a model's hyperparameters.
(c) To determine the generalization of a machine learning model.
(d) To train multiple machine learning models on different datasets.

Correct answers: (d)

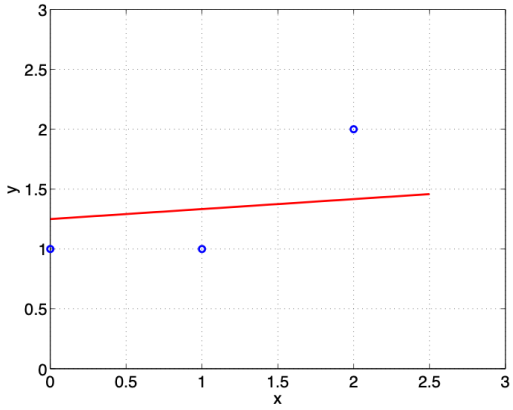
Explanation: The answer "to train multiple ML models on different datasets" is the correct one. We could argue that CV trains the same machine learning model on different partitions of the same dataset, but **not** *multiple* ML models on *different* datasets

15. The plots below show linear regression results on the basis of only three data points.

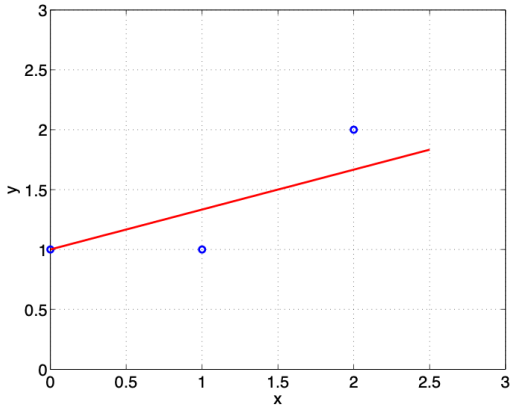
a)



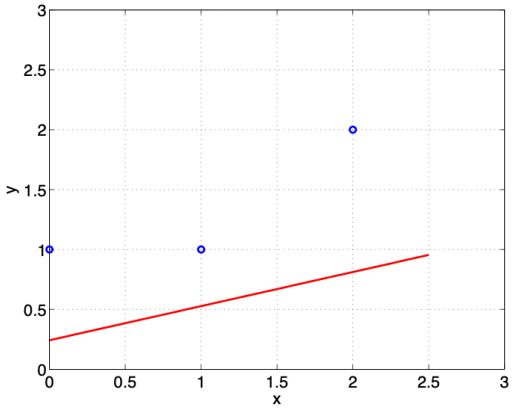
b)



c)



d)



Which plot would result from using the following objective, where $\lambda = 10$?

$$f(w) = \sum_{i=1}^{3} (y_i - wx_i - b)^2 + \lambda w^2$$

(a) Plot A

(b) Plot B

(c) Plot C

(d) Plot D

Correct answers: (b)

Explanation: The answer is B. The slope is strongly regularized making the regression function flat. Since we

don't regularize the offset parameter it is still possible to lift the flat function in the middle of the responses.

Note: since b is not regularized, the sum of positive and negative errors will be exactly zero at the optimal setting of the parameters $\sum_{i=1}^{3} \left(y_i - \hat{w}x_i - \hat{b}\right)^2$

16. Let $a \in \mathbb{R}^n$ and $b \in \mathbb{R}^m$. Let $C = ab^\top$. What are the dimensions of the matrix C, what is the range of C, and what is the null space of C?

Answer:
Dimensions of C: _____

Range of C: _____

Null space of C (provide the set of vectors): _____

Explanation:
- $C \in \mathbb{R}^{n \times m}$.
- $range(C) = span(\{a\})$ (if $b \neq 0$ and $\{0\}$ otherwise).
- $null(C) = \{v \in \mathbb{R}^m | v^T b = 0\}$ (if $a \neq 0$ and $\mathbb{R}^m$ otherwise).

Full credit for getting the main cases right

17. What is the objective of least squares regression?

(a) To minimize the sum of the absolute differences between predicted and actual values.
(b) To minimize the sum of the squared differences between predicted and actual values.
(c) To maximize the number of points on the line of best fit.

Correct answers: (b)

18. An unbiased machine learning model is trained on a dataset with noisy features and achieves a prediction accuracy of 75%. If the irreducible error due to noise in the features is estimated to be 10%, what is the estimated variance of the model?

Answer: _____

Explanation: The answer is 15%.

19. Convexity is a desirable property in machine learning because it:

(a) guarantees gradient descent finds a global minimum in optimization problems for functions that have a global minimum
(b) helps to avoid the model overfitting
(c) speeds up model training
(d) reduces model complexity

Correct answers: (a)

20. True/False: Stochastic gradient descent typically results in a smoother convergence plot (loss vs. epochs) as compared to gradient descent.

(a) True
(b) False

Correct answers: (b)

21. Consider the univariate function $f(x) = x^2$. This function has a unique minimum at $x^* = 0$. We're using gradient descent (GD) to find this minimum and at time t we arrive at the point $x_t = 2$. What is the step size that would bring us to x^* at time $t + 1$?

Answer: _____

Explanation: Using the definition of GD and the requirements in the problem statement we are looking for η such that:

$$x_* = x_t - \eta \nabla f(x_t) \iff 0 = 2 - \eta \nabla f(2) \iff \eta \cdot (2 \cdot 2) = 2 \iff \eta = \frac{1}{2}$$

22. Let $X \in \mathbb{R}^{m \times n}$, $w \in \mathbb{R}^n$, $Y \in \mathbb{R}^m$, and $R(w)$ be some regularization function from $\mathbb{R}^n \to \mathbb{R}$. Consider mean squared error with regularization $L(w) = \|Xw - Y\|_2^2 + \lambda R(w)$. What is $\nabla_w L(w)$?

(a) $2Y^\top (Xw - Y) + \lambda$

(b) $2X^\top (X^\top Xw - Y) + \lambda \nabla_w R(w)$

(c) $2X^\top (Xw - Y) + \lambda \nabla_w R(w)$

(d) $2Y^\top (X^\top Xw - Y) + \lambda R(w)$

Correct answers: (c)

23. Suppose we have n Gaussian distributions $\mathcal{D}_1, \ldots, \mathcal{D}_n$, where each $\mathcal{D}_i = \mathcal{N}(\mu_i, \sigma^2)$. In other words, each Gaussian distribution shares the same variance σ^2, but may have different means μ_i. For each distribution, we draw a single data point $X_i \sim \mathcal{D}_i$. Given $X = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}$, we want to predict $\begin{bmatrix} \mu_1 \\ \vdots \\ \mu_n \end{bmatrix}$.

Solution 1 is to predict X, and Solution 2 is to predict $\frac{7}{8}X$. Why might Solution 2 produce lower mean squared error than Solution 1?

Answer: _____

Explanation: This is Stein's Paradox. Answer must explain what that means though. Solution 2 introduces a small amount of bias to decrease variance. Depending on the value of σ^2, this may produce lower error overall.

This page is intentionally left blank as scratch paper.