DS-GA-1003: Machine Learning (Spring 2023)

Midterm Exam (4:55pm–6:35pm, March 7)

Answer the questions in the spaces provided. If you run out of room for an answer, use the blank page at the end of the test.

Name: _____

NYU NetID: _____

Question	Points	Score
Generalization	15	
Optimization	15	
Regularization	10	
SVM	13	
Kernels	15	
Total:	68	

1. **Generalization and risk decomposition.** Sara is a data scientist who works for a hospital, and she is tasked with building a model to predict which patients are likely to develop diabetes. She has a dataset that contains information about the patient's age, BMI, blood pressure, glucose level, and other relevant factors.

   To begin her work, Sara must decide which machine learning algorithm to use. She knows that there are two main types of machine learning: supervised and unsupervised learning.

   (a) (3 points) What are the differences between supervised and unsupervised learning, and in what situations would each type be appropriate?

After some research, Sara decided to use a supervised learning algorithm. Before training her model, Sara splits her dataset into training, validation, and test sets.

(b) (4 points) What is the purpose of splitting a dataset into training, validation, and test sets, and how does this affect the estimation of the generalization error? **Briefly** explain using the definition of generalization error.

Next, Sara wants to explore her model and its error.

(c) (4 points) Explain the concept of a hypothesis class, how it relates to approximation error in machine learning, and how it relates to the model's ability to fit the true underlying function.

With all of these considerations in mind, Sara is ready to train her model and predict which patients are likely to develop diabetes.

(d) (4 points) What is estimation error, and how is it influenced by the complexity of the model and the amount of available training data?

2. **Optimization.**

(a) (3 points) What is the difference between batch gradient descent and stochastic gradient descent, and when should one be used over the other?

(b) (3 points) How does the learning rate affect the convergence of the gradient descent algorithm, and what are the advantages and disadvantages of using a higher learning rate?

(c) (3 points) What is the tradeoff between using a larger or smaller mini-batch size in stochastic gradient descent, in terms of gradient estimation quality and optimization speed?

(d) (3 points) How does the differentiability of a loss function affect the optimization process, and what methods can be used to handle non-differentiable loss functions?

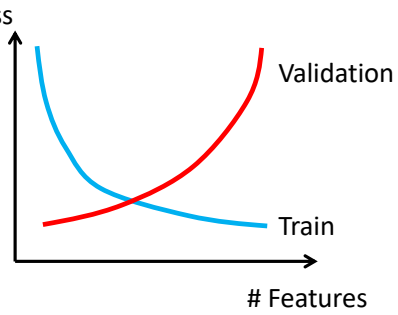
(e) (3 points) How do outliers affect the selection of a loss function, and how can this issue be addressed?

3. **Regularization.**

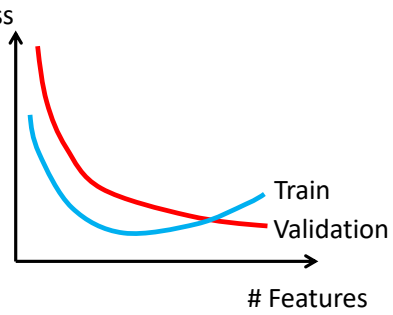
(a) (3 points) Suppose you are trying to choose a good subset of features for a least-squares linear regression model. Let Algorithm A be forward stepwise selection, where we start with zero features and at each step we add the new feature that most decreases validation error, stopping only when validation error starts increasing. Let Algorithm B be similar, but at each step we include the new feature that most decreases training error (measured by the usual cost function, mean squared error), stopping only when training error starts increasing. What is the relationship between the number of features that the two algorithms will end up selecting?

(b) (2 points) You are selecting a subset of features for a machine learning program. Which of the following will you likely observe in terms of training and validation loss?

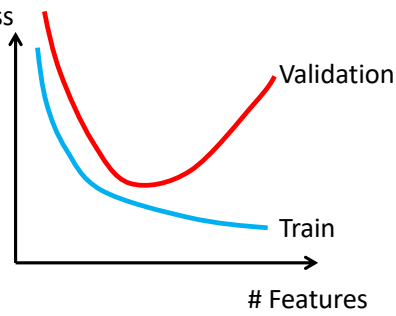
A)



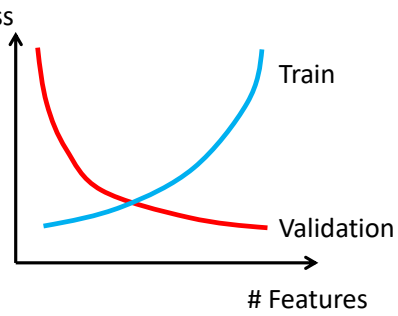
B)



C)



D)



(c) (2 points) Select all true statements below.

1. Ridge regression has an analytical solution.
2. Lasso regression has an analytical solution.
3. Lasso regression tends to produce sparse solution.
4. Both L1 and L2 regularizers encourage weights to be close to zero.

(d) (3 points) Suppose that you are building a binary classification for logistic regression. Recall that logistic regression loss is $L = -y_i \log(p_i) - (1 - y_i) \log(1 - p_i)$, where $p_i = \frac{1}{1+\exp(-\mathbf{w}^\top \mathbf{x}_i + b)}$.
Suppose that the dataset is linearly separable, explain why you may need to apply L2 regularization.

4. **Support Vector Machines.** Recall the (soft-margin) SVM primal problem:

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}||w||^2 + \frac{c}{n}\sum_{i=1}^{n}\xi_i \\ \text{subject to} \quad & -\xi_i \leq 0 \quad \text{for } i = 1, \ldots, n \\ & \left(1 - y_i\left[w^T x_i + b\right]\right) - \xi_i \leq 0 \quad \text{for } i = 1, \ldots, n \end{aligned}$$

and its dual problem:

$$\begin{aligned} \text{maximize} \quad & \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j x_j^T x_i \\ \text{subject to} \quad & \sum_{i=1}^{n}\alpha_i y_i = 0 \\ & \alpha_i \in \left[0, \frac{c}{n}\right] \quad \text{for } i = 1, \ldots, n \end{aligned}$$

(a) (1 point) The primal objective is

A. convex

B. concave

C. neither convex nor concave

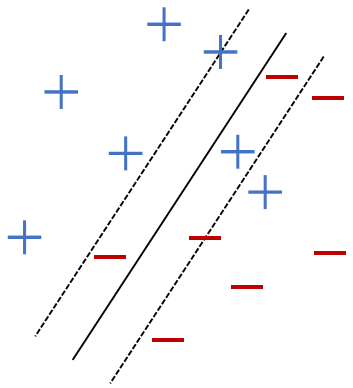
(b) (1 point) The dual objective is

A. convex

B. concave

C. neither convex nor concave

(c) Recall that given the dual solution α_i^*'s, the primal solution is given by $w^* = \sum_{i=1}^{n}\alpha_i^* y_i x_i$, and the support vectors are defined to be x_i's where $\alpha_i > 0$. The figure below shows a toy dataset and the SVM decision boundary (the solid line) with the corresponding margin (indicated by the two dotted lines).



i. (2 points) Draw a triangle around all points that have the slack variable likely to be zero ($\xi_i = 0$) (no partial credits).

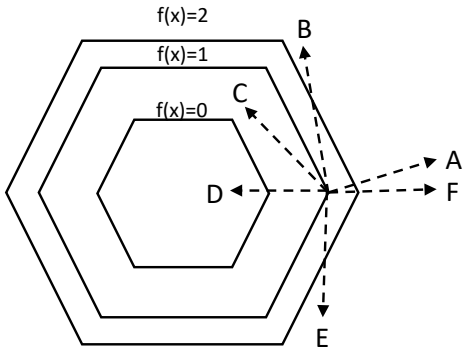
ii. (2 points) Draw a circle around all points that are likely support vectors (no partial credits).

(d) Assuming the data is not linearly separable, increasing c is likely to result in (circle the right answer)

i. (2 points) smaller / larger geometric margin

ii. (2 points) fewer / more support vectors

(e) (3 points) Circle all vectors in the following figure that is in the subdifferential.



5. **Kernels.**

(a) (3 points) The Gaussian kernel is a popular choice for kernel regression and is defined as:

$$K(x_i, x_j) = \exp\left(-\frac{(x_i - x_j)^2}{2h^2}\right)$$

where x_i and x_j are two input values, and h is the kernel bandwidth or smoothing parameter.

Discuss the bias-variance tradeoff in kernel regression. How does the choice of the kernel bandwidth affect the tradeoff? Provide an example of a situation where increasing the kernel bandwidth might improve the performance of the kernel regression model.

(b) (2 points) Why is it important to know that a solution is in the "span of the data"?

(c) (4 points) How does the representer theorem affect the tradeoff between model complexity and generalization performance? Provide an example of a situation where the representer theorem can be used to extract useful information from a linear model.

(d) (2 points) Why we should use a feature extractor?

(e) (4 points) What is the difference between linear and nonlinear models, and how do kernels help in modeling nonlinear relationships?

Congratulations! You have reached the end of the exam.