CS689: Machine Learning

Fall 2021 - Midterm Exam

Name: _____

Instructions: Write your name on the exam sheet. The duration of this exam is **two hours**. No electronic devices may be used during the exam. You may consult either your paper notes and/or a print copy of texts during the exam. Sharing of notes/texts during the exam is strictly prohibited. Show your work for all derivation questions. Provide answers that are as detailed as possible for explanation questions. Attempt all problems. Partial credit may be given for incorrect or incomplete answers. If you need extra space for answers, write on the back of the **preceding** page. If you have questions at any time, please raise your hand.

Problem	Topic	Page	Points	Score
1	Loss Functions	1	10	
2	Gradient Descent	3	10	
3	Regularized Risk Minimization	5	10	
4	Basis Expansion	7	10	
5	Bias	9	10	
6	Regularization Dynamics	11	10	
7	Sub-Differentials	13	10	
8	Lagrangian Functions	15	10	
9	Lagrange Duals	17	10	
10	Modeling	19	10	
Total:			100	

1. (*10 points*) **Loss Functions**. Consider the function $f(z)$ given below when answering the following questions.

$$f(z) = \begin{cases} 1 - 2z & z < 0 \\ (z-1)^2 & 0 \leq z < 1 \\ 0 & 1 \leq z \end{cases}$$

a. (*5 pts*) What is the derivative of $f(z)$? Show your work.

b. (*5 pts*) Let the function $h(z) = [z \leq 0]$ be equal to 1 when z is less than or equal to 0 and let it be 0 otherwise. Explain the significance of the fact that $f(z) \geq h(z)$ when considering using $l(y, g_\theta(\mathbf{x})) = f(y \cdot g_\theta(\mathbf{x}))$ as a classification loss function.

2. (*10 points*) **Gradient Descent**. Consider the learning problem given below for the linear regression model $f_\theta(\mathbf{x}) = \mathbf{x}\mathbf{w} + b$ with $\theta = [\mathbf{w}, b]$ when answering the following questions.

$$\hat{\theta} = \arg\min_{\theta} \sum_{n=1}^{N} (y_n - f_\theta(\mathbf{x}))^2$$

a. (*5 pts*) Give pseudo code for a gradient descent algorithm for finding $\hat{\theta}$, including the computation of the gradient.

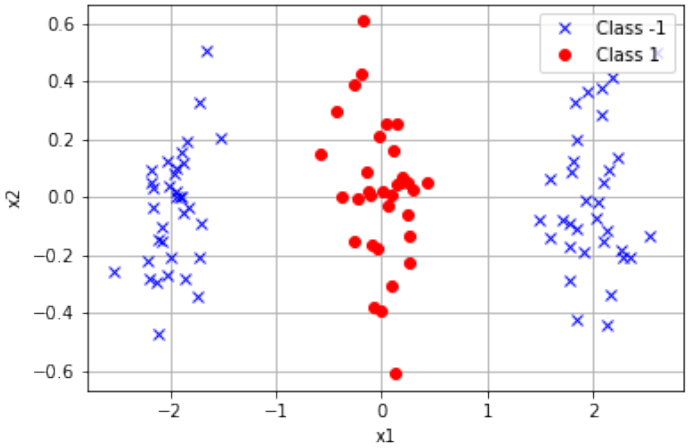
b. (*5 pts*) Give two reasons why we might prefer to use a gradient descent algorithm to solve the OLS learning problem instead of the closed-form estimator $\hat{\theta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$.

3. (*10 points*) **Regularized Risk Minimization**. Consider the linear regression model $f_\theta(\mathbf{x}) = \mathbf{x}\mathbf{w} + b$ with $\theta = [\mathbf{w}, b]$.

a. (*5 pts*) Give the regularized risk minimization problem when fitting the model $f_\theta(\mathbf{x})$ using the absolute loss and squared two norm regularizer. Explain your answer.

b. (*5 pts*) Explain why we should not use regular gradient descent to solve this regularized risk minimization problem.

4. (*10 points*) **Basis Expansion**. Consider the data set shown below where $\mathbf{x} \in \mathbb{R}^2$. Give a minimal basis expansion (a basis expansion with the fewest components) that will perfectly classify these data when using a linear classifier in the basis expanded space. Explain your answer.



5. (*10 points*) **Bias**. Consider a regression problem where $\mathbf{X}$ is the matrix of feature vectors and $\mathbf{Y}$ is the vector of targets. Recall that the OLS estimator for the linear regression model given these data is $\hat{\theta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ (assume $\mathbf{X}^T\mathbf{X}$ is invertible).

Suppose we re-scale $\mathbf{Y}$ forming new targets $\mathbf{Z} = s\mathbf{Y}$ for $s > 0$. Consider the estimator $\hat{\phi} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Z}$. Under the assumption that $E_{p(\mathbf{Y}|\mathbf{X})}[\mathbf{Y}] = \mathbf{X}\theta_*$, prove that $\hat{\phi}/s$ is an unbiased estimator for θ_*.

6. (*10 points*) **Regularization Dynamics**. Recall the the linear classifier is given by $f_\theta(\mathbf{x}) = \text{sign}(g_\theta(\mathbf{x}))$, $g_\theta(\mathbf{x}) = \mathbf{x}\mathbf{w} + b$, $\theta = [\mathbf{w}, b]$. Consider the regularized logistic regression learning problem for the linear classifier shown below when answering the following questions.

$$\hat{\theta} = \arg\min_{\theta} \sum_{n=1}^{N} \log(1 + \exp(-y_n g_\theta(\mathbf{x}_n)) + \lambda \|\mathbf{w}\|_2^2$$

a. (*5 pts*) Suppose we fit the model with $\lambda = 1$ and find that the train error is much lower than the test error. Should we increase or decrease λ to try to improve the test error? Explain your answer.

b. (*5 pts*) Suppose that we run a grid search using a validation set split from the training set. However, at the optimal value of λ the error on the test set is still much higher than the error on the training set. Assuming that all implementations are correct, explain how this might happen.

7. (*10 points*) **Sub-Differentials**. Consider the function $f(z) = \max(0, 1 - z)$. Derive the sub-differential of $f(z)$. Show your work and/or explain your answer.

8. (*10 points*) **Lagrangian Functions**. Recall that the linear classifier is given by $f_\theta(\mathbf{x}) = \text{sign}(g_\theta(\mathbf{x}))$, $g_\theta(\mathbf{x}) = \mathbf{x}\mathbf{w} + b$, $\theta = [\mathbf{w}, b]$. Consider the SVC learning problem for the linear classifier shown below.

$$\hat{\theta} = \arg\min_\theta \; C \sum_{n=1}^{N} \max(0, 1 - y_n g_\theta(\mathbf{x}_n)) + \|\mathbf{w}\|_2^2$$

Suppose we need to solve this problem with the additional simplex constraints $\forall\, d \; w_d \geq 0$ and $\sum_{d=1}^{D} w_d = 1$. Give a Lagrangian function for this problem and specify any constraints on the Lagrange multipliers. Show your work and/or explain your answer.

9. (*10 points*) **Lagrange Duals**. Consider the Lagrangian function given below. x and y are the primal parameters and λ is the Lagrange multiplier. Derive the Lagrange dual function and provide the Lagrange dual optimization problem. Show your work and explain your answer.

$$\mathcal{L}(x, y, \lambda) = x^2 + y^2 - \lambda(x + y - 2)$$

10. (*10 points*) **Modeling**. In some real-world supervised learning problems, the targets are semi-continuous. This means that for some fraction p of the data cases, the target values y_n are exactly 0. For the remainder of the cases, the target values y_n are strictly positive real numbers. Which cases have targets equal to 0 and which do not typically depends on the values of the feature variables $\mathbf{x}_n$. The target values not exactly 0 are also dependent on the values of the features. Describe how we could use the models we have seen to date to provide a prediction function for this problem. Briefly explain how you would learn the parameters of the prediction function.