DSGA-3001.007 Introduction to Machine Learning (Fall 2019)

PRACTICE MIDTERM (October 23)

The exam has 6 pages. Answer the questions in the spaces provided. If you run out of room for an answer, use page 6 at the end of the test.

Name: ______________________________

NYU NetID: ______________________________

Question	Points	Score
1	3	
2	1	
3	4	
4	4	
5	4	
Total:	16	

1. Let $\mathcal{X} = \{1, 2, 3\}$, let $\mathcal{Y} = \{1, 2, 3, 4, 5\}$, and let $\mathcal{A} = \mathcal{Y}$. Suppose the data generating distribution, P, has marginal $X \sim \text{Unif}\{1, 2, 3\}$ and conditional distribution $Y|X = x \sim \text{Unif}\{x, x+1, x+2\}$. Assume we are using the square loss $\ell(a, x) = (a - x)^2$. [Note: Unif denote the uniform distribution on the given set.]

   (a) (1 point) What is the target function?

   Solution: $f^*(x) = x + 1$.

   (b) (2 points) What is the risk of the target function?

   Solution:

$$E[(Y - f^*(X))^2] = E[E[(Y - (X+1))^2|X]] = E[2/3] = \frac{2}{3}.$$

2. (1 point) Which **one** of the following statements is **least plausible** (i.e., probably FALSE) about minibatches for gradient descent.

   ☐ Improved implementation or improved hardware can allow us to increase the minibatch size and simultaneously reduce convergence time (in seconds).

   ☐ In general, enlarging the minibatch size (chosen randomly, with replacement) lets us get a better estimate of the full training set gradient.

   ■ **In general, if we increase the size of our training set by a factor of 1000, then the best minibatch size (with respect to convergence time, in seconds) should also increase by a factor of 1000.**

3. Let $\mathcal{X} = \mathbb{R}^d$ and let $\mathcal{Y} = \mathcal{A} = \mathbb{R}$. Define the infinite collection of hypothesis spaces $\{\mathcal{F}_r \mid r \geq 0\}$ where

$$\mathcal{F}_r = \{f(x) = w^T x + b \mid w \in \mathbb{R}^d, b \in \mathbb{R}, \|w\|_2 \leq r\}.$$

Define the additional hypothesis space

$$\mathcal{F}_\infty = \{f(x) = w^T x + b \mid w \in \mathbb{R}^d, b \in \mathbb{R}\}.$$

Fix a training set $(x_1, y_1), \ldots, (x_n, y_n)$ where $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$. Throughout, assume we are using some arbitrary fixed loss function ℓ.

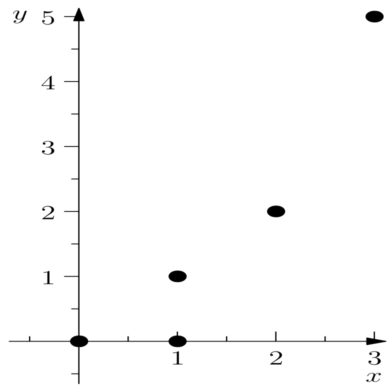
(a) (1 point) $\underline{\mathcal{F}_\infty}$ Among all hypothesis spaces $\mathcal{F}_r$ for $r \geq 0$, and $\mathcal{F}_\infty$, give a hypothesis space that has empirical risk minimizer with the smallest empirical risk.

(b) (1 point) $\underline{\mathcal{F}_\infty}$ Among all hypothesis spaces $\mathcal{F}_r$ for $r \geq 0$, and $\mathcal{F}_\infty$, give a hypothesis space that has the lowest approximation error.

(c) (1 point) __F__ **True or False**: Let f_∞ denote the empirical risk minimizer over $\mathcal{F}_\infty$, and let f_c denote the empirical risk minimizer over $\mathcal{F}_c$, where c was chosen by minimizing the loss on a validation set. Then we **always** have $R(f_c) \leq R(f_\infty)$.

(d) (1 point) __T__ **True or False**: Let f_∞ and f_c be as defined previously. Suppose, mistakenly, we reused the training set as the validation set when choosing c. Then we **always** have $\hat{R}(f_c) = \hat{R}(f_\infty)$ (where $\hat{R}$ still refers to the empirical risk on the training set).

4. Let $\mathcal{X} = [0,1]$ and $\mathcal{Y} = \mathcal{A} = \mathbb{R}$. Suppose you receive the (x, y) data points $(0,0)$, $(1,0)$, $(1,1)$, $(2,2)$, $(3,5)$. Throughout assume we are using the $0-1$ loss function $\ell(a, y) = \mathbf{1}(a \neq y)$.



(a) (1 point) Suppose we restrict to the hypothesis space $\mathcal{F}_1$ of constant functions. What is the empirical risk minimizer $\hat{f}(x)$?

Solution: $\hat{f}(x) = 0$

(b) (1 point) Suppose we restrict to the hypothesis space $\mathcal{F}_1$ of constant functions. What is $\hat{R}(\hat{f})$, the empirical risk of $\hat{f}$, where $\hat{f}$ is the empirical risk minimizer?

Solution:

$$\frac{3}{5}$$

(c) (2 points) Suppose we restrict to the hypothesis space $\mathcal{F}_2$ of increasing functions. What is the empirical risk of the associated empirical risk minimizer?

Solution:

$$\frac{1}{5}$$

5. Consider the following version of the elastic-net objective:

$$J(w) = \frac{1}{n}\|Xw - y\|_2^2 + \lambda_1\|w\|_1 + \lambda_2\|w\|_2^2.$$

Here we have a training set $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^d \times \mathbb{R}$, $X \in \mathbb{R}^{n\times d}$ has x_i^T as its ith row, and $y \in \mathbb{R}^n$ has y_i as its ith coordinate. We fit our data 3 times with the following configurations:

1. Configuration A) $(\lambda_1, \lambda_2) = (0, 0)$
2. Configuration B) $(\lambda_1, \lambda_2) = (5, 0)$
3. Configuration C) $(\lambda_1, \lambda_2) = (0, 5)$

Answer the following questions based on the above information.

(a) For each of the following, state **one** of the configurations that is **most likely** being described. Below w^* represents a minimizer of J.

i. (1 point) __**B**__ w^* has several entries that are 0.

ii. (1 point) __**A**__ The decision function corresponding to w^* has the lowest training error out of all of the configurations.

(b) (2 points) Suppose each data point x has 2 features (x_1, x_2), and that we are using Configuration C. We applied feature normalization which resulted in new scaled features

$$\tilde{x}^T = (\tilde{x}_1, \tilde{x}_2) = (2x_1, x_2/3).$$

This gives the new objective

$$J_s(\tilde{w}) = \frac{1}{n}\|\tilde{X}\tilde{w} - y\|_2^2 + 5\|\tilde{w}\|_2^2$$

which when minimized gives decision function

$$f_{\tilde{w}}(\tilde{x}) = \tilde{w}^T\tilde{x} = 2\tilde{w}_1 x_1 + \tilde{w}_2 x_2/3.$$

Which **one** of the following unscaled objectives, when minimized, will yield the same decision function? Below we use the unscaled decision function

$$f_w(x) = w_1 x_1 + w_2 x_2$$

and want $f_w(x) = f_{\tilde{w}}(\tilde{x})$.

☐ $J(w) = \frac{1}{n}\|Xw - y\|_2^2 + 5w_1^2 + 5w_2^2$

■ $J(w) = \frac{1}{n}\|Xw - y\|_2^2 + 5w_1^2/4 + 45w_2^2$

☐ $J(w) = \frac{1}{n}\|Xw - y\|_2^2 + 20w_1^2 + 5w_2^2/9$

.