DS-GA-3001.007: Introduction to Machine Learning (Fall 2019)

Midterm Exam (October 23 1:00-2:40PM)

- You have 90 minutes to complete the exam.
- The exam is closed book, closed notes, closed computer, closed calculator, except one hand-written 8.5" $\times$ 11" reference sheet of your own creation.
- The exam has 6 pages. Mark your answers on the exam itself. We will not grade answers written on scratch paper.

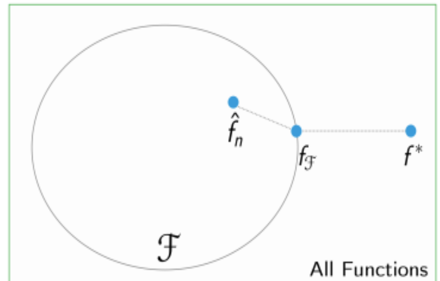
Name: ______________________________

NYU NetID: ______________________________

NYU Email: ______________________________
(as it appears on Gradescope)

Question	Points	Score
Decomposing Risk	11	
Model Selection	4	
Gradient Descent	4	
Regularization	4	
Scaling	4	
Perceptron	3	
Computing Risk	8	
Total:	38	

1. Fix a space of features $\mathcal{X}$ and labels $\mathcal{Y}$. Fix a loss function ℓ. Consider hypothesis space $\mathcal{F}$ of functions from $\mathcal{X}$ to $\mathcal{Y}$ and sample S drawn from $\mathcal{X} \times \mathcal{Y}$.
Take



- $f^* = \underset{f}{\operatorname{argmin}}\ \mathbb{E}\left[\ell(f(x), y)\right]$
- $f_{\mathcal{F}} = \underset{f \in \mathcal{F}}{\operatorname{argmin}}\ \mathbb{E}\left[\ell(f(x), y)\right]$
- $\hat{f} = \underset{f \in \mathcal{F}}{\operatorname{argmin}}\ \frac{1}{|S|} \sum_{i=1}^{|S|} \ell(f(x_i), y_i)$
  where $|S|$ is the number of sample in S.

(a) Recall that the approximation error is the difference of risks $R(f_{\mathcal{F}}) - R(f^*)$.

i. (1 point) Is the approximation error

■ **Positive** ☐ Negative ☐ Cannot be Determined

ii. (1 point) Is the approximation error random or non-random?

☐ Random ■ **Non-Random** ☐ Cannot be Determined

iii. (1 point) If we increase the size of $\mathcal{F}$, then is the approximation error

☐ Increased or Unchanged ■ **Decreased or Unchanged** ☐ Cannot be Determined

iv. (1 point) If we increase the size of S, then is the approximation error

☐ Changed ■ **Unchanged** ☐ Cannot be Determined

v. (1 point) Do we need to know the data generating distribution to compute approximation error

■ **True** ☐ False

(b) Recall that the estimation error is the difference of risks $R(\hat{f}) - R(f_{\mathcal{F}})$.

i. (1 point) Is the estimation error

■ **Positive** ☐ Negative ☐ Cannot be Determined

ii. (1 point) For fixed sample S is the estimation error random or non-random?

☐ Random ■ **Non-Random** ☐ Cannot be Determined

iii. (1 point) If we increase the size of $\mathcal{F}$, then do we ***expect*** the estimation error to

☐ Increase ☐ Decrease ☐ Unchanged ■ **Cannot be Determined**

iv. (1 point) If we increase the size of S, then do we ***expect*** the estimation error to

☐ Increase ☐ Decrease ☐ Unchanged ■ **Cannot be Determined**

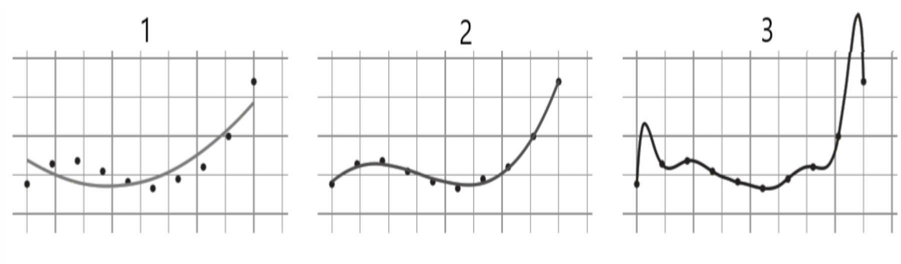
v. (1 point) Do we need to know the data generating distribution to compute approximation error

■ **True** ☐ False

(c) (1 point) Suppose we choose an algorithm to determine the empirical risk minimizer. Does our choice impact

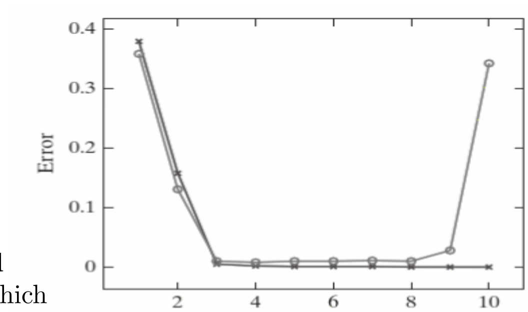
☐ Approximation Error ☐ Estimation Error ■ **Neither**

2. (a) (1 point) Based on the sample, how would you ***expect*** to label the following fits:



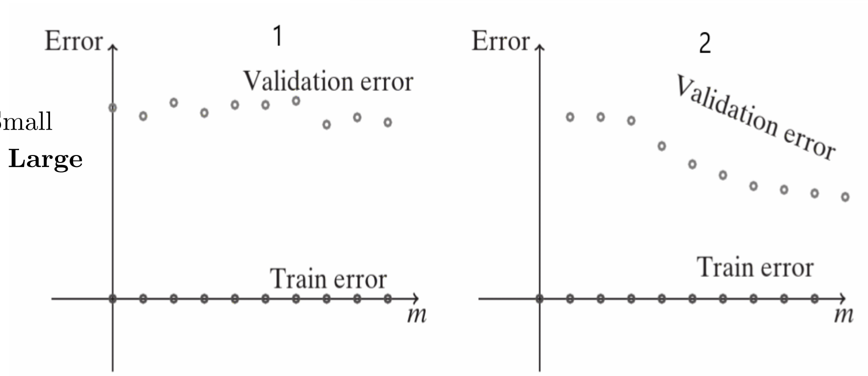
☐ 1. Overfitting 2. Neither 3. Underfitting
☐ 1. Neither 2. Overfitting 3. Underfitting
☐ 1. Underfitting 2. Overfitting 3. Neither
■ **1. Underfitting 2. Neither 3. Overfitting**

(b) (1 point) Suppose we are fitting data with polynomials. Denote the degree of the polynomial by d. The following plots error on the training set and the validation set. How would you label the plot



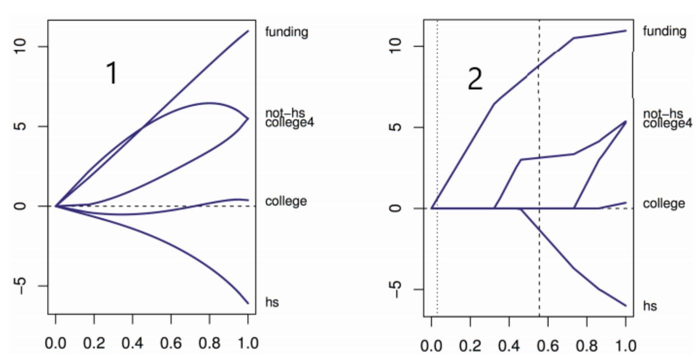
■ **X Training O Validation**
☐ X Validation O Training

(c) (1 point) Suppose we are fitting a small dataset and a large dataset. Take m to be number of epochs. Which of the following would likely correspond to the small dataset and large dataset?



☐ 1. Large 2. Small
■ **1. Small 2. Large**

(d) (1 point) Label the regularization paths as Ridge Regression or Lasso Regression



■ **1. Ridge 2. Lasso**
☐ 1. Lasso 2. Ridge

3. (a) Suppose we're studying linear regression. We want to minimize a sum of squares $f(w)$ with (stochastic) gradient descent. Assume we have weights w_t with update rules

- $w_{t+1} \leftarrow w_t - \eta v_{\text{GD}}$ for gradient descent
- $w_{t+1} \leftarrow w_t - \eta v_{\text{SGD}}$ for stochastic gradient descent

Recall that a vector v is a descent direction at w_t when $f(w_t + \eta v) \leq f(w_t)$ for η small.

i. (1 point) **T** **True or False:** $-v_{\text{GD}}$ is necessarily a descent direction.

ii. (1 point) **F** **True or False:** $-v_{\text{SGD}}$ is necessarily a descent direction.

(b) (2 points) In mini-batch gradient descent, we can randomly choose from the sample with replacement or without replacement. The following snippet of code implements the update rule for mini-batch gradient descent. Note that the design matrix X_b has m rows.

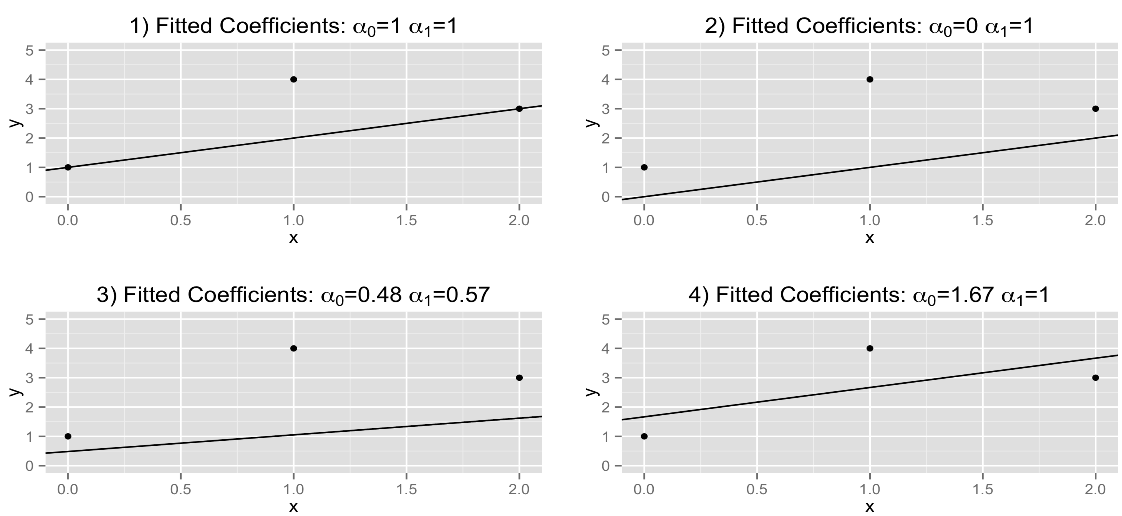
```
for epoch in range(n_iterations):
    shuffled_indices = np.random.permutation(m)
    X_b_shuffled = X_b[shuffled_indices]
    y_shuffled = y[shuffled_indices]
    for i in range(0, m, minibatch_size):
        t += 1
        xi = X_b_shuffled[i:i+minibatch_size]
        yi = y_shuffled[i:i+minibatch_size]
        gradients = 2/minibatch_size * xi.T.dot(xi.dot(theta) - yi)
        eta = learning_schedule(t)
        theta = theta - eta * gradients
        theta_path_mgd.append(theta)
```

Is the code with replacement or without replacement: **Without replacement**

4. (a) We have a dataset $= \{(0,1), (1,4), (2,3)\}$ that we fit by minimizing an objective function of the form:

$$J(\alpha_0, \alpha_1) = \sum_{i=1}^{3} (\alpha_0 + \alpha_1 x_i - y_i)^2 + \lambda_1 (\alpha_0 + \alpha_1) + \lambda_2(\alpha_0^2 + \alpha_1^2),$$

and the corresponding fitted function is given by $f(x) = \alpha_0 + \alpha_1 x$. We tried four different settings of λ_1 and λ_2, and the results are shown below.



For each of the following parameter settings, give the number of the plot that shows the resulting fit.

i. (1 point) __**1**__ $\lambda_1 = 0$ and $\lambda_2 = 2$.

ii. (1 point) __**4**__ $\lambda_1 = 0$ and $\lambda_2 = 0$.

iii. (1 point) __**3**__ $\lambda_1 = 0$ and $\lambda_2 = 10$.

iv. (1 point) __**2**__ $\lambda_1 = 5$ and $\lambda_2 = 0$.

5. Suppose we have $\mathcal{X} = \{-1.5, -0.5, 0.5, 1.5\} \times \{-0.001, 0.001\}$ and $\mathcal{Y} = \{-1, 1\}$. Assume the data generating distribution gives

- y has equal probability of being $-1, 1$.
- x_1 has equal probability of being $\{-1.5, -0.5, 0.5, 1.5\}$. x_1 is related to y through $y = x_1 + 0.5z$ where $z = \pm 1$ with equal probability.
- x_2 has equal probability of being $\{-0.001, 0.001\}$. x_2 is related to y through $y = 1000x_2$.

Suppose we have Ridge Regression with two features, one label and m samples

$$J(w) = \frac{1}{m}\sum_{i=1}^{m}\left(w_1 x_1^{(i)} + w_2 x_2^{(i)} - y_i\right)^2 + \lambda(w_1^2 + w_2^2),$$

We're trying to decide between weights $w_{\text{accurate}} = [0, 1000]$ and $w_{\text{simple}} = [1, 0]$.

(a) (1 point) What is the value of $J(w_{\text{accurate}})$?

☐ 1000λ ☐ 1000 ■ $1000^2\lambda$ ☐ 1000^2

(b) (1 point) Assuming that m is large, we can calculate the empirical risk from the risk. Under this assumption, what is the value $J(w_{\text{simple}})$?

☐ $0.5 + \lambda$ ■ $0.25 + \lambda$ ☐ $1 + \lambda$ ☐ $0.75 + \lambda$

(c) (1 point) Using your answers above, determine λ^* such that we would choose w_{simple} for any $\lambda > \lambda^*$.

Solution: $\frac{0.25}{1000^2 - 1}$

(d) (1 point) For most values of λ, we would choose w_{simple}. How could we transform the features to avoid choosing the less accurate weights?

Solution: Scaling features – for example, min-max scaler

6. (3 points) We can add ℓ^2 regularization to the Perceptron algorithm. Remember the Perceptron loss is $\ell(\hat{y}, y) = \max\{0, -\hat{y}y\}$. Adding $\lambda\|w\|^2$ to the empirical risk, we obtain

$$J(w) = \lambda\|w\|^2 + \frac{1}{n}\sum_{i=1}^{n}\max\left\{0, -y_i(w^T \cdot x_i)\right\}.$$

Suppose we want to minimize J through stochastic gradient descent. Take the learning rate to be 1. With weights w_t, we select (x_t, y_t) from the training set. What is the update rule determined by $\lambda\|w\|^2 + \max\left\{0, -y_t(w^T \cdot x_t)\right\}$

☐

$$w_{t+1} \leftarrow \begin{cases} 2\lambda w_t & \text{if } y_t(w_t^T \cdot x_t) < 0 \\ 2\lambda w_t - y_t x_t & \text{if } y(w^T \cdot x) \geq 0 \end{cases}$$

■

$$w_{t+1} \leftarrow \begin{cases} (1-2\lambda)w_t & \textbf{if } y_t(w_t^T \cdot x_t) \geq 0 \\ (1-2\lambda)w_t + y_t x_t & \textbf{if } y(w^T \cdot x) < 0 \end{cases}$$

□

$$w_{t+1} \leftarrow \begin{cases} 2\lambda w_t & \text{if } y_t(w_t^T \cdot x_t) \geq 0 \\ 2\lambda w_t - y_t x_t & \text{if } y(w^T \cdot x) < 0 \end{cases}$$

□

$$w_{t+1} \leftarrow \begin{cases} (1-2\lambda)w_t & \text{if } y_t(w_t^T \cdot x_t) < 0 \\ (1-2\lambda)w_t + y_t x_t & \text{if } y(w^T \cdot x) \geq 0 \end{cases}$$

7. Consider feature space $\mathcal{X} = \{1,2,3,4\}$ and label space $\mathcal{Y} = \{1,2,3,4\}$. Suppose the data generating distribution gives

- Equal probability $\frac{1}{4}$ to features $\{1,2,3,4\}$, that is, $X \sim \text{Unif}\{1,2,3,4\}$
- Equal probabilities $\frac{1}{x}$ to labels $\{1,\ldots,x\}$ conditional on feature x, that is, $Y \mid X \sim \text{Unif}\{1,\ldots,x\}$

(a) Assume we are using the square loss: $\ell(\hat{y}, y) = (\hat{y} - y)^2$.

i. (1 point) Fix x. Take a derivative to determine the constant c such that $\mathbb{E}\left[(Y-c)^2|X=x\right]$ is minimized.

Solution: $c = E[Y|X]$

ii. (1 point) What is the target function, that is, for fixed x how should we choose $f^*(x)$ to minimize the expected square loss.

Solution: $f^*(x) = (x+1)/2$.

iii. (2 points) What is the expected square loss of the target function?

Solution:

$$E[(Y - f^*(X))^2] = E[E[(Y-(X+1)/2)^2|X]] = \frac{26}{48}.$$

(b) Assume we are using the 0-1 loss:

$$\ell(\hat{y}, y) = \begin{cases} 0 & \text{if } \hat{y} = y \\ 1 & \text{if } \hat{y} \neq y \end{cases}.$$

i. (1 point) Fix x. What value of y is most probable? Is it unique?

Solution: Values $1,\ldots,x$ meaning not unique

ii. (1 point) What is the target function, that is, for fixed x how should we choose $f^*(x)$ to minimize the expected 0-1 loss.

Solution: $f^*(x)$ any number $1,\ldots,x$.

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

Solution: Take $f^*(x) = 1$ for all x. We have

$$E[\ell(Y,1)] = E[E[\ell(Y,1)|X]] = \frac{6}{16}.$$

END OF EXAM – PRESENT YOUR NYU ID AT SUBMISSION