6.036: Midterm, Spring 2018

Solutions

- This is a closed book exam. Calculators not permitted.
- The problems are not necessarily in any order of difficulty.
- Record all your answers in the places provided. If you run out of room for an answer, continue on a blank page and mark it clearly.
- If a question seems vague or under-specified to you, make an assumption, write it down, and solve the problem given your assumption.
- **Write your name on every page.**
- Come to the front to ask questions.

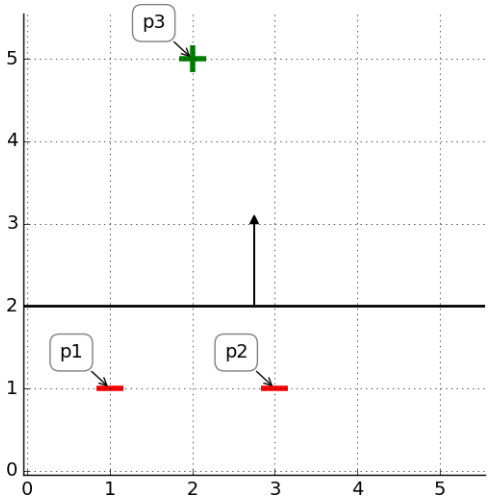
Name: _____ Athena ID: _____

Question	Points	Score
1	20	
2	10	
3	18	
4	28	
5	14	
6	10	
Total:	100	

Name: _____

Margin

1. (20 points) Here are some points and a hypothesis.



(a) Give parameters θ and θ_0 for the separator, such that $\|\theta\| = 1$.

i. θ: $[0, 1]^T$

ii. θ_0: -2

(b) What is the margin of each point with respect to that separator?

i. Point 1: **1**

ii. Point 2: **1**

iii. Point 3: **3**

(c) What would be the next values of θ and θ_0 after one step of batch gradient descent where the objective is

$$J(\theta, \theta_0) = \frac{1}{3}\left(\sum_{i=1}^{3} L_{hinge}(y^{(i)}(\theta^T x^{(i)} + \theta_0))\right) + \lambda\|\theta\|^2,$$

with $\lambda = 0$ and step size $= 1$?

i. θ: $[0, 1]^T$

ii. θ_0: -2

Name: _____

(d) What is the margin of the *whole data set* with respect to the new separator? Your answer should be a single number (or fraction).

1

(e) What would be the next values of θ and θ_0 after one step of batch gradient descent where the objective is

$$J(\theta, \theta_0) = \frac{1}{3}\left(\sum_{i=1}^{3} L_{hinge}(y^{(i)}(\theta^T x^{(i)} + \theta_0))\right) + \lambda\|\theta\|^2,$$

with $\lambda = .01$ and step size $= 1$?

i. θ: $[0, .98]^T$

ii. θ_0: -2

(f) What is the *margin of the whole data set* with respect to the new separator? Your answer should be a single number (or fraction).

1.04

Now consider the situation when $\theta = [0, 1]^T, \theta_0 = -1/2$.

(g) What would be the next values of θ and θ_0 after one step of batch gradient descent where the objective is

$$J(\theta, \theta_0) = \frac{1}{3}\left(\sum_{i=1}^{3} L_{hinge}(y^{(i)}(\theta^T x^{(i)} + \theta_0))\right) + \lambda\|\theta\|^2,$$

with $\lambda = 0$ and step size $= 1$?

i. θ: $[-4/3, 1/3]^T$

ii. θ_0: $-7/6$

Name: _____

Sources of error

2. (10 points) Recall that *structural* error arises when the hypothesis class cannot represent a hypothesis that performs well on the test data and *estimation* error arises when the parameters of a hypotheses cannot be estimated well based on the training data.

   Following is a collection of potential cures for a situation in which your learning algorithm generates a hypothesis with high test error.

   For each one, indicate whether it **can reduce** structural error, estimation error, neither, or both.

   (a) Penalize $\|\theta\|^2$ during training
   ◯ structural error √ **estimation error** ◯ both ◯ neither

   (b) Penalize $\|\theta\|^2$ during testing
   ◯ structural error ◯ estimation error ◯ both √ **neither**

   (c) Increase the amount of training data
   ◯ structural error √ **estimation error** ◯ both ◯ neither

   (d) Increase the order of a fixed polynomial basis
   √ **structural error** ◯ estimation error ◯ both ◯ neither

   (e) Decrease the order of a fixed polynomial basis
   ◯ structural error √ **estimation error** ◯ both ◯ neither

   (f) Add more layers with linear activation functions to your neural network
   ◯ structural error ◯ estimation error ◯ both √ **neither**

   (g) Add more layers with non-linear activation functions to your neural network
   √ **structural error** ◯ estimation error ◯ both ◯ neither

   (h) Stop training before training error reaches 0
   ◯ structural error √ **estimation error** ◯ both ◯ neither

   For each of the following situations, indicate whether the **poor performance is due to** high structural error, high estimation error, neither, or both.

   (i) Neural network has very low training error but high testing error.
   ◯ structural error √ **estimation error** ◯ both ◯ neither

   (j) Neural network training error is persistently high, as is test error.
   √ **structural error** ◯ estimation error ◯ both ◯ neither

Name: _____

Formulation

3. (18 points) We want to design a neural network to solve each of these problems. For each one, indicate a good choice of:

   - representation of **target output values** y
   - activation function on the output layer
   - loss function

   Note: Write a mathematical expression for the loss function, not just the type of loss in words. You can assume "*guess*" and "*actual*" are scalars or vectors depending on context; use subscripts on these variables to index the output if it is a vector.

   (a) Predict when a train will arrive based on the day, time, and weather, in minutes relative to the scheduled arrival time. If your prediction is *after* the train actually arrives, it has loss 100. If before, then the loss is the number of minutes early you predict.

   i. Representation of target output value:
   ◯ integer ✓ **real number** ◯ one-hot vector ◯ vector of values in $\{0, 1\}$

   ii. Output activation function: **linear**

   iii. Loss function (provide full equation):
   $Loss(guess, actual) =$

   Solution:
   $$\begin{cases} 100 & \text{if } guess > actual \\ actual - guess & \text{otherwise} \end{cases}$$

   (b) Predict which items—out of 10,000 possible items sold by Things 'R' Us—a shopper will purchase during one shopping trip, based on their previous shopping history. You care only about whether or not an item is bought (*not* the quantity purchased), and any given customer can order multiple items.

   i. Representation of target output value:
   ◯ integer ◯ real number ◯ one-hot vector ✓ **vector of values in** $\{0, 1\}$

   ii. Output activation function: **sigmoid**

   iii. Loss function (provide full equation):
   $Loss(guess, actual) =$

   Solution:
   $$-\sum_{i=1}^{10000} \left(actual_i \log guess_i + (1 - actual_i) \log(1 - guess_i)\right)$$

(c) Predict the nationality of a person (out of 100 possible values) based on their walking speed and clothing.

i. Representation of target output value:
◯ integer ◯ real number √ **one-hot vector** ◯ vector of values in $\{0, 1\}$

ii. Output activation function: **softmax**

iii. Loss function (provide full equation):
$Loss(guess, actual) =$

Solution:

$$-\sum_{i=1}^{100} actual_i \log guess_i$$

Name: _____

Radial basis features

4. (28 points) We will consider a systematic way of creating a new feature space called *radial basis functions*. To define the new features, we need a set of example points $E = (E_1, \ldots, E_k)$ where each example $E_i \in \mathbb{R}^d$ where d is the dimension of the original input space. Our feature transformation is:

$$\phi(x) = \begin{bmatrix} \exp(-\beta\|E_1 - x\|^2) \\ \exp(-\beta\|E_2 - x\|^2) \\ \cdots \\ \exp(-\beta\|E_k - x\|^2) \end{bmatrix}$$

for some $\beta > 0$.

(a) What is the dimension of $\phi(x)$? __k **or** $k \times 1$__

(b) Consider the following concrete example:
$d = 1$, $E = [[1], [2]]$, $\beta = 1$
Original data set $X = [[0], [1], [2]]$, $Y = [[+1], [-1], [+1]]$

i. Is the data set X, Y linearly separable in the original space?
◯ Yes √ **No**

If yes, provide parameters that describe a separator, otherwise write 'None.'

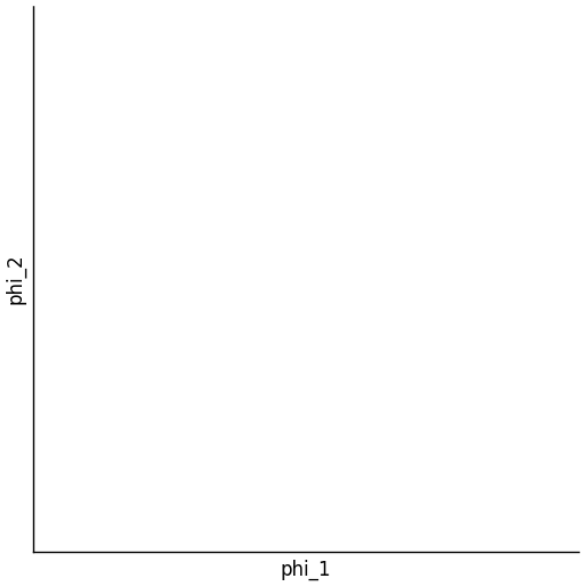
ii. θ: __**None**__

iii. θ_0: __**None**__

Name: _____

(c) On the axes below, plot the points $\phi([0]), \phi([1]), \phi([2])$. Label them clearly. This table may be useful:

v	$\exp(-v)$
0	1.0
1	0.37
2	0.13
4	0.02
8	0.0003



(d) Is the data set $\phi(X), Y$ linearly separable?
√ **Yes** ◯ No

(e) If so, provide parameters that describe a separator.

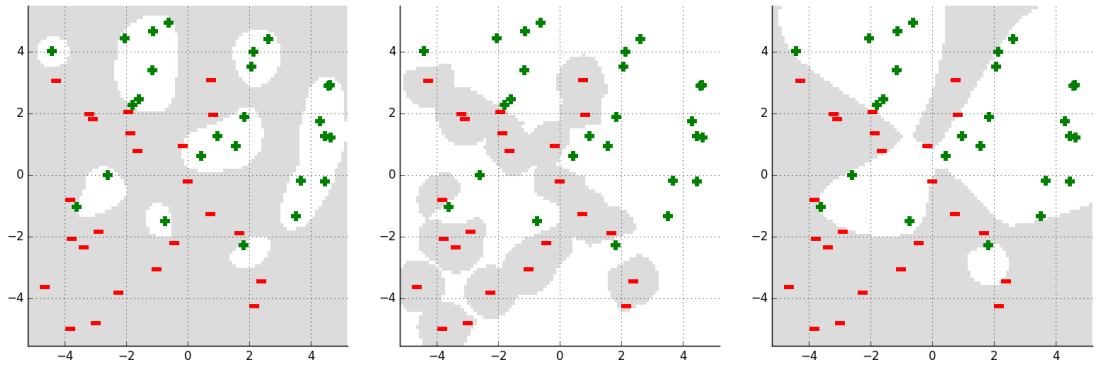
i. θ: $[-1, 0]$

ii. θ_0: 0.5

One common strategy is to use the x values in the training set as E. We will do that in the following questions.

(f) Now, in a 2D domain, the following plots show separators obtained for values of $\beta \in \{0.1, 1.0, 1000\}$ using the perceptron algorithm. The shaded area corresponds to regions of space that will be classified as negative.

Match the plot to the value.



Beta: **1** Beta: **1000** Beta: **0.1**

(g) In the data set above, the perceptron algorithm made 344, 42, and 162930 mistakes on three of the runs, but we forgot which β values they corresponded to. Match the number of mistakes to the $\beta \in \{0.1, 1.0, 1000\}$.

i. Mistakes: 42 Beta: **1000**

ii. Mistakes: 344 Beta: **1.0**

iii. Mistakes: 162930 Beta: **0.1**

(h) In the limit as β approaches ∞, what familiar form does this feature representation take?

one hot encoding!

Now consider the case where we fix the number, k, of example points, but allow the coordinates of the points to be adjusted by the learning algorithm. We can think of this as a kind of neural network, parameterized by $E_1, \ldots, E_k$ as well as a k-dimensional weight vector W and offset W_0, so that the output of the network is

$$\hat{y} = W^T \phi(x) + W_0$$

(i) If our loss function on a single data point is $Loss(\hat{y}, y) = (\hat{y} - y)^2$, what is $\nabla_W Loss(\hat{y}, y)$?

Solution:

$$\begin{aligned}
\nabla_W Loss(\hat{y}, y) &= \nabla_W (\hat{y} - y)^2 \\
&= 2(\hat{y} - y) \nabla_W \hat{y} \\
&= 2(\hat{y} - y)\phi(x)
\end{aligned}$$

(j) What is $\nabla_{E_1} Loss(\hat{y}, y)$? (Remember that $\frac{d}{dt}\exp(t) = \exp(t)$.)

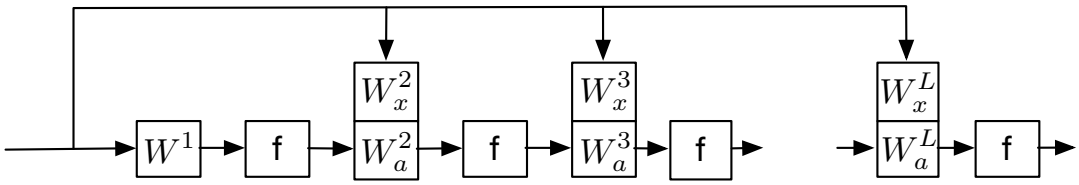
Solution:

$$\begin{aligned}
\nabla_{E_1} Loss(\hat{y}, y) &= \nabla_{E_1} (\hat{y} - y)^2 \\
&= 2(\hat{y} - y) \nabla_{E_1} \hat{y} \\
&= 2(\hat{y} - y) \nabla_{E_1} W^T \phi(x) \\
&= 2(\hat{y} - y) W_1 \nabla_{E_1} \phi_1(x) \\
&= 2(\hat{y} - y) W_1 \nabla_{E_1} \exp(-\beta \|E_1 - x\|^2) \\
&= -2\beta(\hat{y} - y) W_1 \exp(-\beta \|E_1 - x\|^2) \nabla_{E_1} \|E_1 - x\|^2 \\
&= -4\beta(\hat{y} - y) W_1 \exp(-\beta \|E_1 - x\|^2)(E_1 - x)
\end{aligned}$$

Name: _____

Shortcut connections

5. (14 points) In some neural-network models, it has proved useful to pass the input value, unchanged, into each of the subsequent layers, as shown in the figure below.



The forward pass is governed by the following equations:

$$
\begin{aligned}
a^0 &= x \\
a^1 &= f({W^1}^T a^0 + W_0^1) \\
z^l &= {W_a^l}^T a^{l-1} + {W_x^l}^T x + W_0^l \\
a^l &= f(z^l)
\end{aligned}
$$

Note that the second line above defines how to compute the output of the *first* layer, and the third and fourth lines define how to compute the output of all subsequent layers for $l = 2 \ldots L$.

(a) Let $m^l = n^{l-1}$ be the number of inputs entering layer l and n^l be the number of outputs. So, m^0 is the dimension of the input vector.

What is the dimension of W_a^l for $l > 1$? $\underline{m^l \times n^l \textbf{ or } n^{l-1}} \times n^l$

(b) What is the dimension of W_x^l for $l > 1$? $\underline{\quad m^0 \times n^l \quad}$

(c) Now we will think of the backward pass of back-propagation for the "linear" modules in this network. Given $\partial \text{Loss}/\partial z^l$, a^{l-1}, W_a^l and x

i. Write an expression for $\partial \text{Loss}/\partial a^{l-1}$.

Solution: $W_a^l \cdot \partial \text{Loss}/\partial z^l$

ii. Write an expression for $\partial \text{Loss}/\partial W_a^l$.

Solution: $a^{l-1}(\partial \text{Loss}/\partial z^l)^T$

iii. Write an expression for $\partial \text{Loss}/\partial W_x^l$.

Solution: $x(\partial \text{Loss}/\partial z^l)^T$

Name: _____

Passive-aggressive algorithm

6. (10 points) The perceptron algorithm, through the origin, iterates through its data set, considering each point $(x^{(i)}, y^{(i)})$, where $x^{(i)} \in R^d$ and $y^{(i)} \in \{+1, -1\}$, and making changes to its parameters θ based on that point. If the point is classified correctly, it makes no change. If the point is not classified correctly, the algorithm performs the update:

$$\theta = \theta + y^{(i)} x^{(i)}$$

After this update, the data point $x^{(i)}, y^{(i)}$ may still not be classified correctly.

(a) In two dimensions (when $d = 2$), provide values for θ, $x^{(i)}$ and $y^{(i)}$ for which this is the case (there are *many* possible answers—any one will do).

i. θ : **0, 100**

ii. $x^{(i)}$: **(0, -1)**

iii. $y^{(i)}$: **+1**

(b) The *passive-aggressive* algorithm is a variation of the perceptron algorithm which performs an update for any point satisfying $y^{(i)}\theta^T x^{(i)} < 1$. The update has the form

$$\theta = \theta + cy^{(i)}x^{(i)}$$

where c may be a function of $x^{(i)}$, $y^{(i)}$, and θ. We are interested in finding a minimal value of $c > 0$ for which the data point will satisfy

$$y^{(i)}\theta_{new}^T x^{(i)} \geq 1 \quad .$$

after the update. It turns out that the solution has the form:

$$c = \alpha(1 - y^{(i)}\theta^T x^{(i)})$$

Give an expression for α that represents the smallest magnitude update that will cause $y^{(i)}\theta_{new}^T x^{(i)} \geq 1$ to be true. You may use θ, $x^{(i)}$, and/or $y^{(i)}$ in your expression.

Solution:

$$\frac{1}{\|x^{(i)}\|^2}$$

Name: _____

Work space