6.036: Midterm, Spring 2019

Do not tear exam booklet apart!

- This is a closed book exam. One page (8 1/2 in. by 11 in.) of notes, front and back, is permitted. Calculators are 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.
- If you absolutely *have* to ask a question, come to the front.
- **Write your name on every page.**

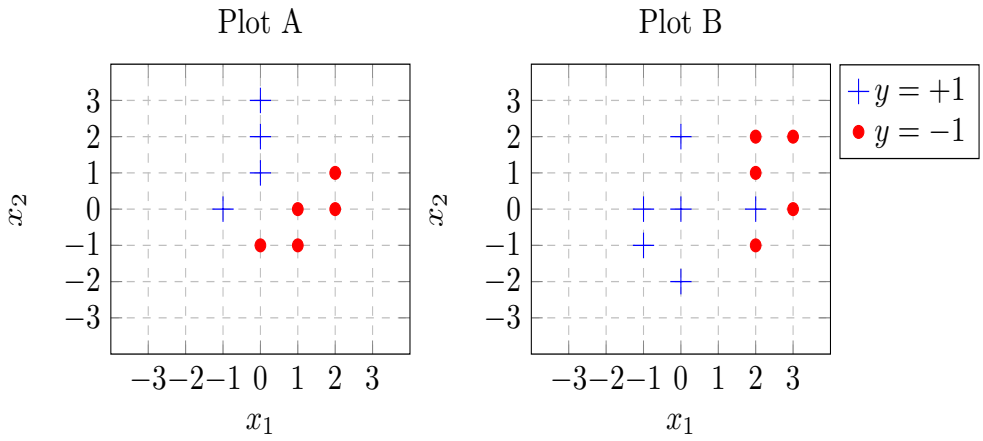
Name: _____ Athena ID: _____

Question	Points	Score
1	14	
2	17	
3	15	
4	10	
5	14	
6	15	
7	15	
Total:	100	

Name: _____

Linear Classifiers

1. (14 points) In the plots below, we give you 2D points with $+1$ and -1 labels.



Answer the following questions for both plot A and plot B:

(a) Using a linear separator $h(p; \theta, \theta_0) = sign(\theta^\top p + \theta_0)$, what is the minimum possible number of misclassified points?

Plot A:

Plot B:

(b) What are the values of $\theta \in \mathbb{R}^2$ and $\theta_0 \in \mathbb{R}$ that define your separator?

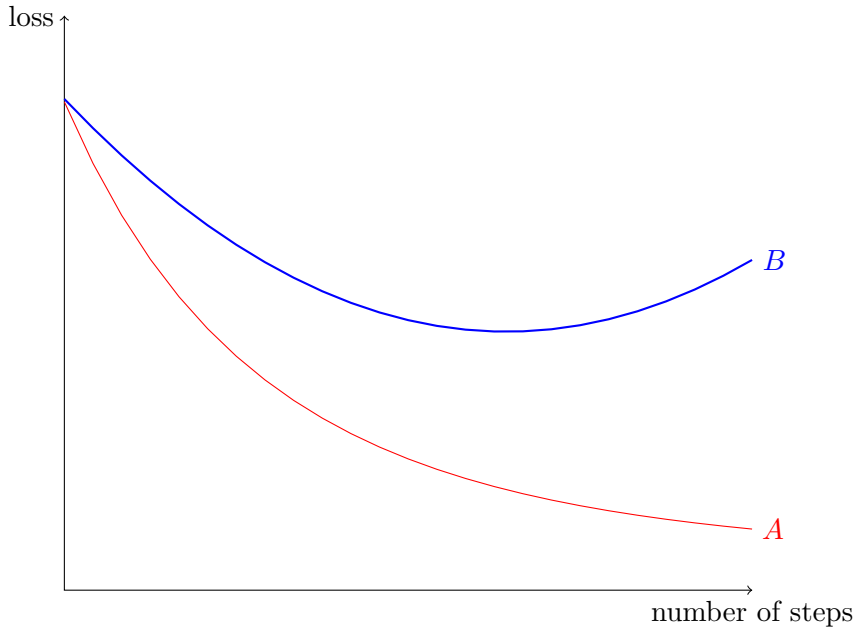
Plot A:

Plot B:

(c) For a given point p, what does $\frac{\theta^\top p + \theta_0}{\|\theta\|}$ intuitively represent?

Name: _____

Consider the following plot from the previous classification task. The two curves show the train and test error vs. the number of steps in the optimization algorithm.



(d) Assign the appropriate labels:

Test error (select one): ◯ A ◯ B

Train error (select one): ◯ A ◯ B

(e) Which of the following options can improve the final performance of the trained classifier on the test data set? Note: augmentation of a data set refers to taking the existing data set and adding many points which are slightly perturbed versions of the original points. Select all that apply.

◯ A. Augment the training data set and retrain the classifier.

◯ B. Augment the test data set and retrain the classifier.

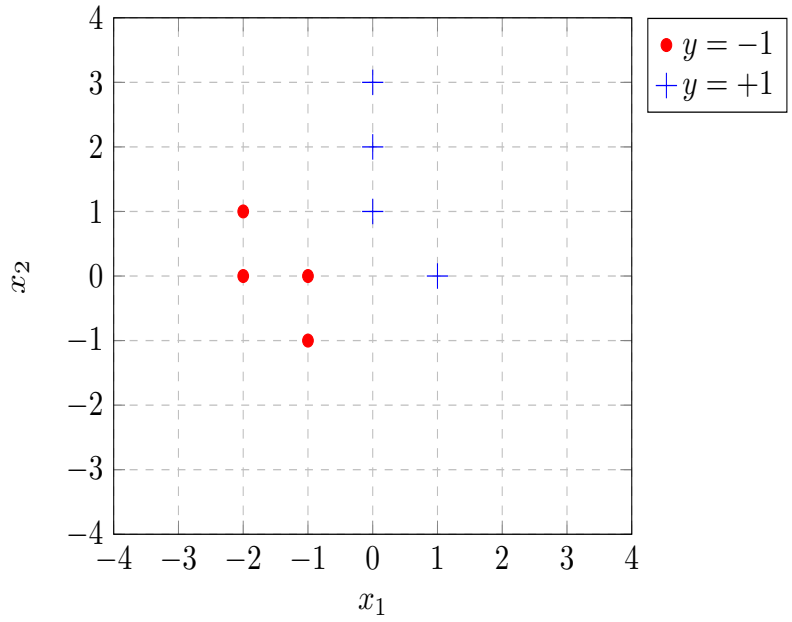
◯ C. Terminate the training process earlier.

◯ D. Add a penalty on the magnitude of the parameter values and retrain the classifier.

Name: _____

Perceptron Algorithm

2. (17 points) Alice plans to apply the Perceptron algorithm to solve a classification problem on the data set shown in the figure below. For this problem, we will only consider linear separators that pass through the origin.



(a) What is the theoretical upper bound on the number of steps for the Perceptron algorithm to find the linear separator on this data set?

(b) Alice suggests a feature transformation of the form: $\phi((x_1, x_2)) = (\alpha x_1, x_2)$. Is there a value of α that would reduce the upper bound on the number of mistakes made by the Perceptron algorithm? If so, provide one such value. If not, explain why not.

(c) Will scaling both coordinates uniformly, i.e., $\phi((x_1, x_2)) = (\beta x_1, \beta x_2)$, decrease the bound on the number of mistakes? Explain.

(d) When the point is classified incorrectly, the algorithm updates θ. Is the point guaranteed to be classified correctly after the update is made? Explain.

(e) A separator is trained using Perceptron with the points $i = 1, ..., N$ in the data set. Write the expression for the general final separator in terms of $\{x^{(i)}\}$, their labels $\{y^{(i)}\}$, and the number of mistakes $\{n_i\}$ that Perceptron made on each of the points.

(f) Is this algorithm guaranteed to find the classifier with maximum margin?
◯ yes ◯ no

Name: _____

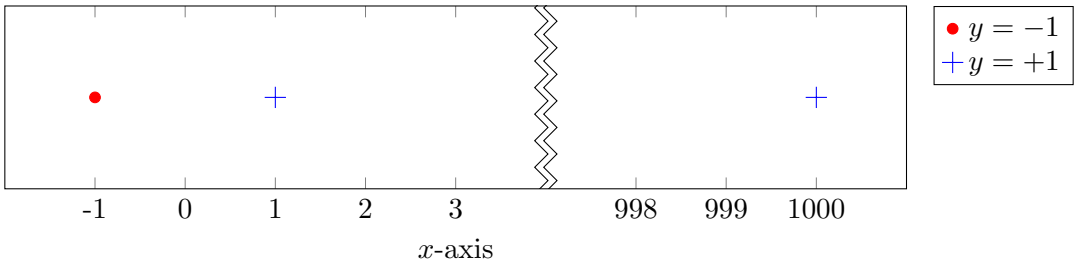
Margin Maximization

3. (15 points) In this problem, we will consider using linear regression for classification problems, and its relationship to margin maximization.

   Student Chris has a data set of samples $\{(x_i, y_i)\}$ with $x_i \in \mathbb{R}$ and classes $y_i \in \{-1, +1\}$. Chris mixed up classification and regression, and ended up computing a linear regression $\hat{y} = \theta^\top x + \theta_0$ instead of a classifier. Having made this mistake, Chris figures out that it is possible to convert the regression into a classifier by taking its sign, i.e. $\hat{y} = \text{sign}(\theta^\top x + \theta_0)$ where:

$$\text{sign}(z) = \begin{cases} 1 & \text{when} \quad z > 0 \\ -1 & \text{when} \quad z \leq 0 \end{cases}$$

   Consider the data set $x = [[-1], [1], [1000]]^\top$ with labels $y = [-1, 1, 1]^\top$.



   (a) What is its maximum margin separator: θ, θ_0 ?

   (b) Chris uses a loss function $L(g, a) = (g - a)^2$ that takes g (guess) and a (actual) as parameters. Specify Chris's objective function for the linear regression problem using θ, θ_0, x_i, y_i:

With this data set, it turns out that the linear regression solution is approximately

$$\hat{y} = 0.001x - 0.000999$$

.

(c) What then is the decision boundary defined by Chris's classifier?

(d) Does the classifier correctly classify all of the points in the training set?

◯ Yes.
◯ No.

(e) Can you add another point to the data set so that the data set is still linearly separable, but so that using linear regression to train it would result in a classifier that mis-classifies one or more points? If yes, specify such a data point. If no, explain why not.

(f) Would you expect this classifier to generalize well? Explain.

(g) Would you expect a maximum margin classifier to generalize better than Chris's? Explain.

Name: _____

Model Evaluation

4. (10 points) Lisa trains models for classification problems. She is provided with different image data sets (e.g., trains, people, cars, cats, dogs) by Snapbook. Each data set has both positive and negative examples. In fact, Snapbook provides Lisa only a fraction of each data set, the remainder is left for internal Snapbook testing. Lisa trains a separate model on each data set. She measures model training accuracy, and she estimates test accuracy using cross-validation. For each model, Snapbook measures the accuracy of the model on the data that was held out (not provided to Lisa). These experiments yield the following results:

	training accuracy	cross-validation accuracy	held-out tests accuracy
data set 1	52%	54%	51%
data set 2	97%	71%	70%
data set 3	93%	92%	55%
data set 4	91%	91%	89%
data set 5	50%	53%	70%

For which data set(s):

(a) Lisa's model is overfitting (check all that apply):
◯ data set 1 ◯ data set 2 ◯ data set 3 ◯ data set 4 ◯ data set 5

(b) It is likely that more training data drawn from the same distribution would improve the quality of the held-out accuracy (check all that apply):
◯ data set 1 ◯ data set 2 ◯ data set 3 ◯ data set 4 ◯ data set 5

(c) Lisa's hypothesis class might not be expressive enough (check all that apply):
◯ data set 1 ◯ data set 2 ◯ data set 3 ◯ data set 4 ◯ data set 5

(d) Held-out data set is not likely from the same distribution as Lisa's (check all that apply):
◯ data set 1 ◯ data set 2 ◯ data set 3 ◯ data set 4 ◯ data set 5

Name: _____

Learning as Optimization

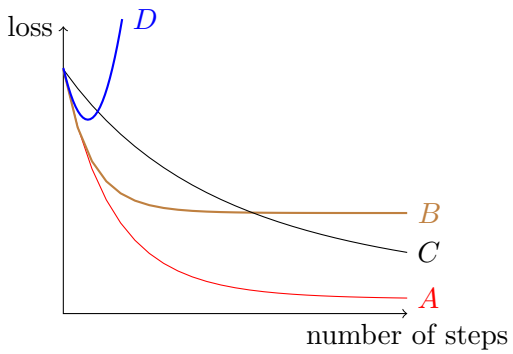
5. (14 points) Ben develops a new hypothesis class: $h(x; w_1, w_2) = w_1 x_1 + w_1 x_1^2 + w_2 x_2 + w_2 x_2^2$, where $x = (x_1, x_2)$. He plans to use it for a regression problem on the data set $S_n = \{(x^{(1)}, y^{(1)}), ..., (x^{(n)}, y^{(n)})\}$.

   (a) Ben will use batch gradient descent to compute model parameters w_1, w_2. His loss function is mean squared error (MSE). Derive an update rule for w_1 given the learning rate η.

   (b) Describe the shape of the MSE as a function of w_1 and w_2. How many minima will it have? Assume that the data set S_n is fixed.

Name: _____

(c) Ben tries different settings of the learning rate η. Depending on the setting he obtains different behavior of the gradient descent algorithm. Match each plot (A,B,C,D) to the best fitting description (assume MSE loss).



Learning rate too low (select one):
◯ A ◯ B ◯ C ◯ D

Learning rate about right (select one):
◯ A ◯ B ◯ C ◯ D

Learning rate too high (select one):
◯ A ◯ B ◯ C ◯ D

Learning rate much too high (select one):
◯ A ◯ B ◯ C ◯ D

(d) Alyssa suggests using a mean absolute error, instead, defined by:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|y^{(i)} - h(x^{(i)}, w_1, w_2)\right|$$

What could be an advantage of this approach?

Name: _____

Neural Networks

6. (15 points) Mira's father is an archaeologist who appraises Chinese antiques. Since his daughter recently took 6.036, he asked her a favor: to build a classifier to predict from which dynasty each antique artifact originates. Specifically, each antique artifact was built by one of the four dynasties: `Tang` (A.D. 618-907), `Song` (A.D. 960-1276), `Ming` (A.D. 1368–1644), `Qing` (A.D. 1636-1912). Mira decides to build a classifier using a neural network and train it using negative log likelihood (NLL) loss. Recall that the negative log likelihood loss for a single example is defined as:

$$L_{NLL}(\hat{y}, y) = -\sum_{i=1}^{n_y} y_i \log \hat{y}_i$$

where $\hat{y} = (\hat{y}_1, \ldots, \hat{y}_{n_y})$ denotes the predicted probability distribution over the classes and $y = (y_1, \ldots, y_{n_y})$ is the ground truth, a one hot vector that has zero at each index except at the correct class: $y = (1, 0, 0, 0)$ for `Tang`, $y = (0, 1, 0, 0)$ for `Song`, $y = (0, 0, 1, 0)$ for `Ming`, $y = (0, 0, 0, 1)$ for `Qing`.

(a) Assume that Mira is given an antique that belongs to `Ming` dynasty ($y = (0, 0, 1, 0)$). Which of these predictions has the smallest NLL loss?

◯ A. $\hat{y} = (0.25, 0.20, 0.30, 0.25)$

◯ B. $\hat{y} = (0.01, 0.01, 0.44, 0.54)$

◯ C. $\hat{y} = (0.25, 0.25, 0.25, 0.25)$

◯ D. $\hat{y} = (0.97, 0.01, 0.01, 0.01)$

(b) Apart from the NLL loss, Mira is also thinking about trying out other loss functions. In particular, she is thinking about using the accuracy:

$$L_{accuracy}(\hat{y}, y) = \begin{cases} -1 & \arg\max(y) = \arg\max(\hat{y}) \\ 0 & otherwise \end{cases}$$

or the squared loss function:

$$L_{squared}(\hat{y}, y) = (1 - \hat{y}y^{\top})^2$$

as her new loss functions. Which of the these loss functions can be minimized by the stochastic gradient descent (SGD) algorithm (mark all that apply)?

◯ A. NLL-loss, L_{NLL}

◯ B. Accuracy, $L_{accuracy}$

◯ C. Squared loss, $L_{squared}$

(c) After trying out different model architectures, Mira finds that softmax classifier works well. When she uses softmax, the last layer of her network computes pre-activations $z = (z_1, ..., z_{n_y})$ which may be arbitrarily large or small (Note: a pre-activation is the linearly weighted sum that is an input to the activation function). Softmax function then computes $\hat{y} = (\hat{y}_1, ..., \hat{y}_{n_y})$ by normalizing z such that the sum of the $\hat{y}_i$ is 1:

$$\hat{y}_i = \frac{e^{z_i}}{\sum_{j=1}^{n_y} e^{z_j}}.$$

Mira finds that for some settings of the pre-activation values, the basic softmax function works poorly. She finds out that subtracting the maximum value $\max z_i$ from all pre-activations z_i produces more reliable results. Explain why.

(d) Say that Mira wanted to solve a slightly different problem: given an artifact, Mira would like to figure out what the probability is that the artifact is "typical" of each of the four time periods. E.g. there could be an antique crafted in a style which was popular both during the `Tang` and `Ming` eras, but not at all in the other two eras, in which case ideally we would output $y = (1, 0, 1, 0)$. Choose a different structure (activation function and number of nodes) for the last layer. Specify a loss function that would work better for this multi-class labeling task.

Activation function and output nodes:

Loss function:

Name: _____

Initialization is Important

7. (15 points) In this problem we will try to understand why proper initialization of weights in a network is important.

   Kim constructs a fully connected deep neural network with L=4 layers using negative log-likelihood (NLL) loss and ReLU activation functions for all hidden layers, and a softmax for the output layer. The ReLU activation function is implemented as $\text{ReLU}(z) = max(0, z)$, with $\partial\text{ReLU}(z)/\partial z = 1$ if $z > 0$, and 0 otherwise. Kim uses random initialization for all of the layers except for layer 2, where he uses zero initialization (*i.e.,* the layer weights are $W^2 = 0$ and $W_0^2 = 0$).

   (a) Before training, he is curious about the output of his network as initialized. What will Kim observe on the output when he provides different input examples, $x^{(i)}$?

   (b) The network will be trained with stochastic gradient descent (SGD). Specify an update rule for W^1 (layer 1 weights) in terms of $\frac{\partial L}{\partial W^1}$ and step size η. Similarly, specify an update rule for W^2 in terms of $\frac{\partial L}{\partial W^2}$ and step size η.

   Kim (correctly) derives the gradient of the loss function with respect to weights W^1 in terms of the activation functions A^l, weights W^l, pre-activations Z^l, and partials $\frac{\partial L}{\partial A^4}$, $\frac{\partial A^l}{\partial Z^l}$, for l = 1,...,4:

$$\frac{\partial L}{\partial W^1} = \frac{\partial Z^1}{\partial W^1} \cdot \frac{\partial L}{\partial Z^1}, \tag{2}$$

   where

$$\frac{\partial L}{\partial Z^1} = \frac{\partial A^1}{\partial Z^1} \cdot W^2 \cdot \frac{\partial A^2}{\partial Z^2} \cdot W^3 \cdot \frac{\partial A^3}{\partial Z^3} \cdot W^4 \cdot \frac{\partial A^4}{\partial Z^4} \cdot \frac{\partial L}{\partial A^4}. \tag{3}$$

(c) After one iteration of gradient descent, will the new weights W^1 be different than the initial weights W^1? Explain why.

(d) After that first iteration of gradient descent, will the new weights W^2 be different than the initial weights $W^2 = 0$? Explain why.

(e) Kim finds that his network with initialization of $W^2 = 0, W_0^2 = 0$ and his ReLU activation as defined above for $L = 2$ performs poorly. He switches to a sigmoid activation function for the hidden nodes in layer $L = 2$ but still uses zero initialization as previously for $L = 2$. Kim finds that the network trains and performs much better. Explain why.

Name: _____

Work space