6.036: Midterm, Spring 2022

Do not tear exam booklet apart!

- This is a closed book exam. One page (8 1/2 in. by 11 in. or A4) of notes, front and back, is permitted. Calculators are not permitted.
- The total exam time is 2 hours.
- 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 piece of paper.**

Name: ______________________ MIT Email: ______________________

Question	Points	Score
1	12	
2	25	
3	22	
4	15	
5	26	
Total:	100	

Name: ______________________________

1 How to do ML?

1. (12 points) Please answer each question with a phrase or short sentence.

   (a) If you are training a linear regression model and find that you are getting very low training error but much higher prediction error on a held-out data set from the same source, you should

   ____ .

   (b) If you have trained a classifier on half of your available data and want to estimate how well it will work in actual practice, you should

   ____ .

   (c) If you have two different neural-network architectures and you are trying to decide which one to use to perform well given your current data set, you should

   ____ .

   (d) If you are training a logistic regression classifier using gradient descent and find that the prediction accuracy stays around 50% but you look at the weights and they're really big, you should

   ____ .

   (e) If you are training a logistic regression classifier using gradient descent and find that the prediction accuracy goes up to 100% but you look at the weights and they're really big, you should

   ____ .

   (f) If you are training a neural network classifier and find that the loss decreases for a while, but then stops decreasing before it gets to an acceptable accuracy, you should

   ____ .

2 Media consumption

2. We are trying to help a new public relations company, NetFlacks, try to make predictions about the popularity of various TV shows.

   (a) (6 points) Assume that we are trying to predict the number of views of a particular TV show in the next month. What is the best way to encode the following inputs? Specify the number of features that would be used to represent each of these inputs, and how the given input value would be represented. It is fine to leave numerical expressions, if you need them, unevaluated.

   i. Genre: non-fiction, comedy, drama, science fiction (each show has a single genre)

   Number of features: __________

   How would we represent a comedy?

   ii. Famous actors: which members of a list of the 200 most famous actors ([*Aarnold Aardvark, ..., Ziggy Ztarduzt*]) appear in it

   Number of features: __________

   How would we represent a movie starring only Aarnold Aardvark and Ziggy Ztarduzt?

   iii. Production cost:

   Number of features: __________

   If our dataset had 5 shows with costs $1M, $2M, $3M, $4M, $5M, how would we represent the cost of the first of these shows?

(b) (5 points) After the show has been released there are a lot of ratings available. Assume that, for each show, you collect all the published reviews, and compute how many of those reviews gave the show 1 star, how many gave it 2 stars, etc., up to 5 stars. You end up with 5 features for each show: $(c_1, c_2, c_3, c_4, c_5)$, where c_i is an integer number of reviews with i stars. So, for example, the following shows might have the following features:

- *Big Blockbuster Bonanza* (100, 1000, 1M, 1000, 10)
- *Obscure Omniglot Omnibus* (0, 0, 1, 1, 10)

You want to do linear regression to predict the number of views of the show.

You consider using these features just as they are, but also consider some different encodings.

1. raw counts, $c_1, \ldots, c_5$ (5 features)
2. average $a = \frac{\sum_{i=1}^{5} i \cdot c_i}{\sum_{i=1}^{5} c_i}$ (1 feature)
   (Note that this was mis-defined in the original exam.)
3. average a and the sum, $s = \sum_{i=1}^{5} c_i$ (2 features)
4. the normalized counts c_i/s (5 features)
5. the normalized counts c_i/s and the sum s (6 features)

In order to pick a good encoding you want to be sure it can encode some plausible models of profitability, via linear regression. For each of the rules below, indicate **all** of the encodings that will allow it to be expressed via *linear regression*, or explain why none of them will do.

i. The number of reviews predicts the number of viewers in the next month.
○ 1. ○ 2. ○ 3. ○ 4. ○ 5.

ii. The percentage of reviews with more than 3 stars predicts the number of viewers next month.
○ 1. ○ 2. ○ 3. ○ 4. ○ 5.

(c) (6 points) What is a good choice of *output* encodings if we are trying to make the following predictions with a neural network? Please also specify an appropriate loss function and output activation function (if one is needed).

i. Number of viewers in the next month
Output encoding, including number of dimensions

____

Loss function

____

Output activation

____

ii. Whether each member of the current 6.036 class (with 450 students) will watch it in the next month, predicting all values as outputs of a single neural network.
Output encoding, including number of dimensions

____

Loss function

____

Output activation

____

iii. Whether the exclusive rights to distribute it will be sold to network A, network B, or TwoYoob.
Output encoding, including number of dimensions

____

Loss function

____

Output activation

____

(d) (8 points) You are interested in the effect of the language of a show on its popularity. You consider two encodings of the language:

- A: One-hot encoding of one of 100 modern languages
- B: An integer index into a list of 100 modern languages

You do regression on a data set with two examples in which the language is the only feature. A show in language 2 has had 200 views and one in language 8 has had 800: $\{((2), 200), ((8), 800)\}$

Recall that in linear regression, we do not regularize θ_0.

For each of the following learning methods, what prediction would the resulting hypothesis make for language 10 on the list? Provide an approximate numeric output or indicate that it is under-specified.

i. Using strongly regularized linear regression on encoding A?

ii. Using strongly regularized linear regression on encoding B?

iii. Using unregularized linear regression on encoding A?

iv. Using unregularized linear regression on encoding B?

3 Relugression

3. Let's consider regression in one dimension, so our inputs $x^{(i)}$ and outputs $y^{(i)}$ are in $\mathbb{R}$.

(a) (4 points) Linny uses regular linear regression. Given the following dataset,

$$\mathcal{D} = \{((1), 1), ((2), 2), ((3), 4), ((3), 2)\}$$

what values of θ and θ_0 optimize the mean squared error of hypotheses of the form $h(x; \theta, \theta_0) = \theta x + \theta_0$?

$\theta =$ ____________ $\theta_0 =$ ____________

(b) (4 points) What property has to be true of your solution above in order for it to be at least a local optimum of the mean squared error objective function

$$J(\theta, \theta_0) = \frac{1}{4} \sum_{(x,y) \in \mathcal{D}} (\theta x + \theta_0 - y)^2 \ ?$$

Provide conditions on $\partial J / \partial \theta$ and $\partial J / \partial \theta_0$

(c) (4 points) Rolly read about neural networks and thinks linear regression would be improved by using a ReLU unit. Recall that ReLU is defined as

$$\text{ReLU}(z) = \begin{cases} 0 & \text{if } z < 0 \\ z & \text{otherwise} \end{cases} = \max(0, z)$$

Now we consider a new dataset:

$$\mathcal{D} = \{((0), 0), ((1), 0), ((2), 2), ((3), 4), ((4), 6)\}$$

What values of θ and θ_0 optimize the mean squared error of hypotheses of the form $h_R(x; \theta, \theta_0) = \text{ReLU}(\theta x + \theta_0)$?

$\theta =$ ____________ $\theta_0 =$ ____________

(d) (4 points) Rolly insists their hypothesis class is bigger than Linny's, in the sense that any data set that can be fit with 0 MSE using a hypothesis in Linny's class can also be fit with 0 MSE using a hypothesis in Rolly's class.

Is Rolly right?

○ Yes ○ No

Do one of the following:

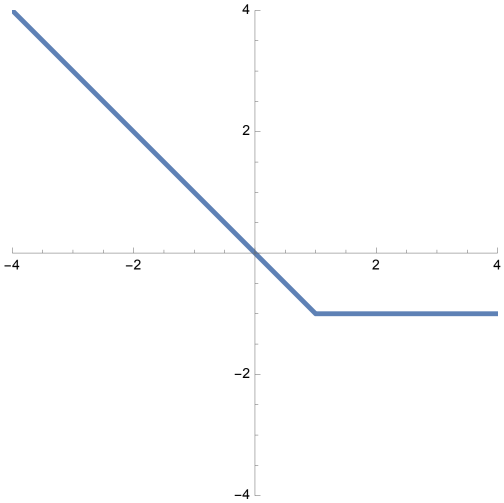
- Argue that Rolly is right by providing, for any finite data set $\mathcal{D}$ such that there is a linear hypothesis θ, θ_0 with 0 MSE, the parameters of a ReLU hypothesis that also has 0 MSE.

- Argue that Rolly is wrong by providing a small dataset that has a 0 MSE linear hypothesis but for which no 0 MSE ReLU hypothesis exists.

(e) (3 points) Lulu is interested in the following hypothesis class, with parameters a, b, c, and d:

$$h(x; a, b, c, d) = a + b\,\mathrm{ReLU}(cx + d)$$

For the plot below, provide values of a, b, c, and d that would generate it. Choose parameter values equal to -1, 0, or 1; or say that it's not possible given this hypothesis class.

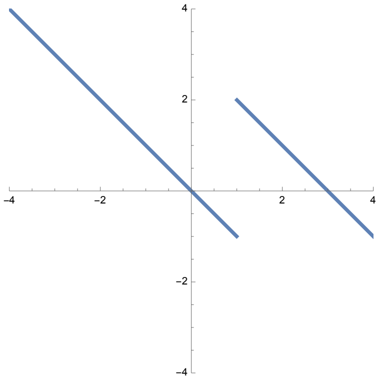


$a =$ ____________ $b =$ ____________ $c =$ ____________ $d =$ ____________

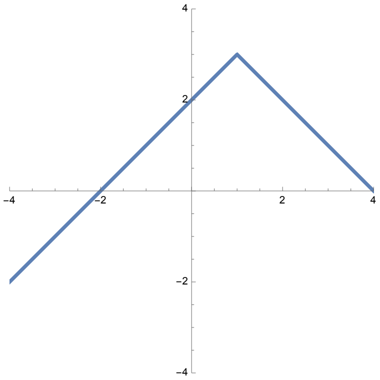
(f) (3 points) Renee wants to be even fancier, and considers hypothesis class with additional parameters e, f, and g:

$$h(x; a, b, c, d, e, f, g) = a + b\,\mathrm{ReLU}(cx + d) + e\,\mathrm{ReLU}(fx + g)$$

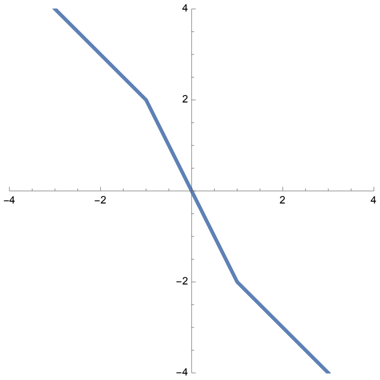
For each plot below, indicate whether it can be expressed using Renee's class.



○ Yes ○ No



○ Yes ○ No

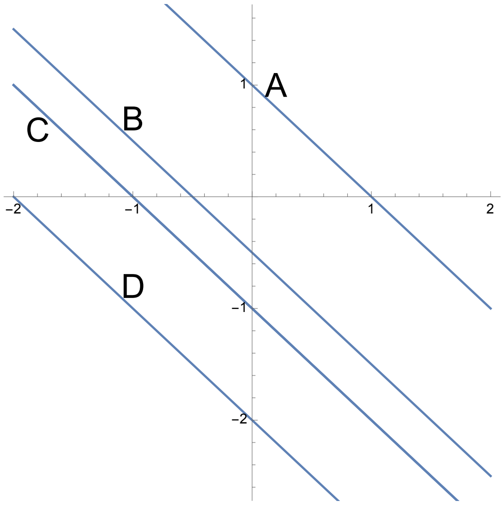


○ Yes ○ No

4 Class struggle

4. We are interested in doing linear classification of points in a 2-dimensional space.

   (a) (5 points) We started with a classifier, described by $\theta = [1, 1]$, $\theta_0 = 1$. All our friends had theories about how to transform it. They drew their candidate separators on a napkin at a restaurant, but forgot to label the or indicate which side was positive vs negative, so we just have drawings of the separators but not the normals.



For each of the transformations below, indicate which separator in the diagram it corresponds to.

i. Multiply θ by 2

○ A ○ B ○ C ○ D ○ none

ii. Multiply θ and θ_0 by 2

○ A ○ B ○ C ○ D ○ none

iii. Multiply θ by -1

○ A ○ B ○ C ○ D ○ none

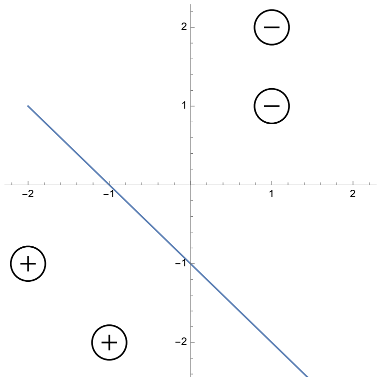
iv. Multiply θ and θ_0 by -1

○ A ○ B ○ C ○ D ○ none

v. Add 1 to θ_0

○ A ○ B ○ C ○ D ○ none

(b) (4 points) Consider the following data set, shown with the separator corresponding to the original linear classifier with parameters $\theta = [1, 1]$, $\theta_0 = 1$.



Remembering to think about which side of the separator is assigned the positive class, select the transformation below that causes the largest decrease in negative log likelihood (NLL)?

○ Multiply θ and θ_0 by 2
○ Multiply θ and θ_0 by 1/2
○ Multiply θ and θ_0 by -1
○ Multiply θ and θ_0 by -2

(c) (6 points) Now consider a problem with input dimension $d = 10$, with separator

$$\theta = [1, 1, 1, 1, 1, 1, 1, 1, 1, 1], \quad \theta_0 = 0$$

We are wondering about the following data points:

- $x^{(1)} = [0, 0, 0, 0, 0, 0, 0, 0, 0, 0]$
- $x^{(2)} = [1, 1, 1, 1, 1, 1, 1, 1, 1, 1]$
- $x^{(3)} = [-1, -1, -1, -1, -1, -1, -1, -1, -1, -1]$
- $x^{(4)} = [1, -1, 1, -1, 1, -1, 1, -1, 1, -1]$
- $x^{(5)} = [1, 0, 0, 0, 0, 0, 0, 0, 0, 0]$

i. Which point or points are assigned to class 1 (positive) with highest confidence?

○ $x^{(1)}$ ○ $x^{(2)}$ ○ $x^{(3)}$ ○ $x^{(4)}$ ○ $x^{(5)}$

ii. Which point or points are assigned to class 0 (negative) with highest confidence?

○ $x^{(1)}$ ○ $x^{(2)}$ ○ $x^{(3)}$ ○ $x^{(4)}$ ○ $x^{(5)}$

iii. Which point or points is this classifier maximally uncertain about?

○ $x^{(1)}$ ○ $x^{(2)}$ ○ $x^{(3)}$ ○ $x^{(4)}$ ○ $x^{(5)}$

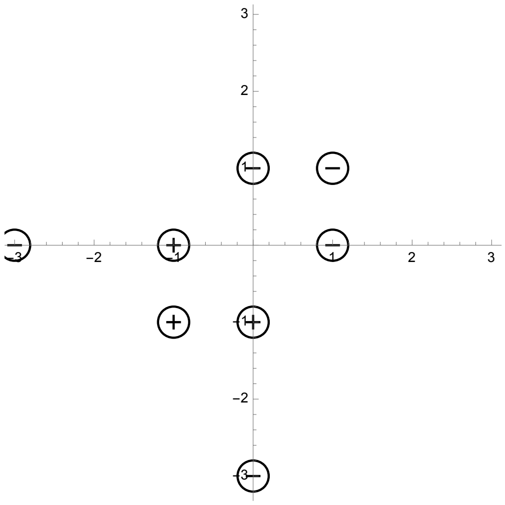
5 Circulation

5. Instead of linear classifiers, we're going to think about circular classifiers (CCs). In d dimensions a CC is parameterized by a point $\theta \in \mathbb{R}^d$ and a radius θ_0. We will classify a point x as positive if $\|x - \theta\|_2 \leq \theta_0$ where

$$\|x - \theta\|_2 = \sqrt{\sum_{j=1}^{d} (x_j - \theta_j)^2}\ .$$

(a) (2 points) Let's start in 1D! Describe the set of points that would be classified as positive by a CC with parameters $\theta = (2), \theta_0 = 1$.

(b) (3 points) Here is a data set in two dimensions. Provide θ and θ_0 for a CC that correctly classifies all the points.



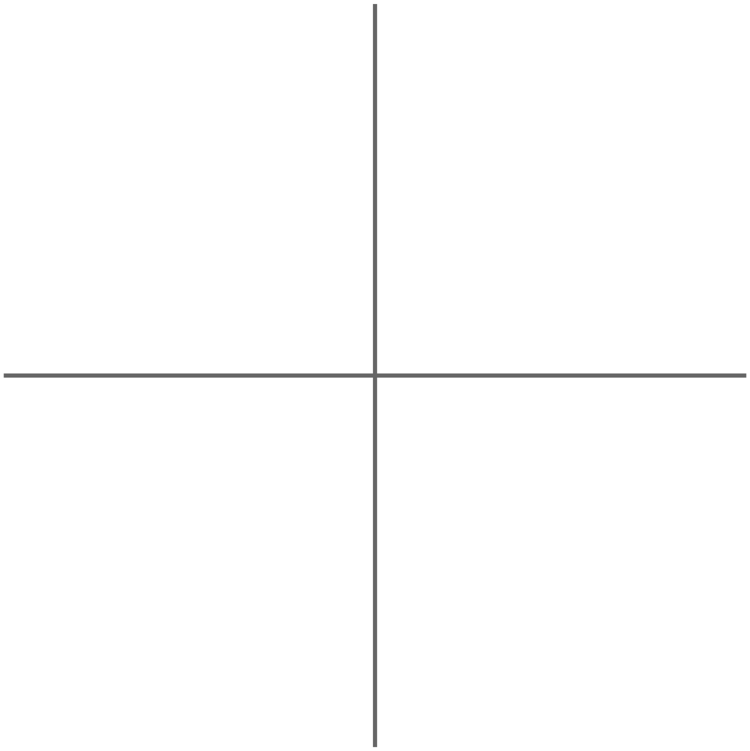
$\theta =$ ______________ $\theta_0 =$ ______________

(c) (2 points) Circe wants to optimize CC's using gradient descent. Her first thought is to use 0-1 loss, but she decides against it. Why would 0-1 loss be a poor choice?

(d) (3 points) Circe decides to emulate logistic regression and use hypothesis class

$$h_c(x; \theta, \theta_0) = \sigma(10\theta_0 - 10\|x - \theta\|_2)$$

On the axes below, sketch $\sigma(10 - 10\|x - 2\|_2)$. Clearly label the X and Y axes.



(e) (6 points) Let's try to optimize negative log likelihood for this hypothesis class. Just using your intuition, for each case, indicate whether a small gradient step on the specified parameter would increase or decrease its value in order to improve the objective value.

i. Let $\theta = 5$, $\theta_0 = 2$, $x = 4$, $y = 0$.

θ should ○ increase ○ decrease ○ stay the same

θ_0 should ○ increase ○ decrease ○ stay the same

ii. Let $\theta = 5$, $\theta_0 = 2$, $x = 0$, $y = 1$.

θ should ○ increase ○ decrease ○ stay the same

θ_0 should ○ increase ○ decrease ○ stay the same

iii. Let $\theta = 5$, $\theta_0 = 2$, $x = 5$, $y = 1$.

θ should ○ increase ○ decrease ○ stay the same

θ_0 should ○ increase ○ decrease ○ stay the same

(f) (5 points) Now, let's do it more formally and for a general d-dimensional input $x \in \mathbb{R}^d$. Let g be the output of the hypothesis,

$$g = \sigma(10\theta_0 - 10\|x - \theta\|_2)$$

and recall that

$$\frac{\partial \mathcal{L}_{nll}(g, y)}{\partial \theta} = \frac{\partial \mathcal{L}_{nll}(\sigma(z), y)}{\partial \theta} = (\sigma(z) - y)\frac{\partial z}{\partial \theta} \quad .$$

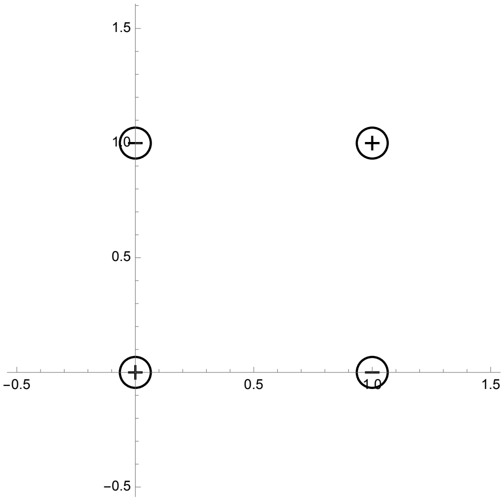
Note also that

$$\frac{\partial}{\partial \theta}\|x - \theta\|_2 = -\frac{x - \theta}{\|x - \theta\|_2} \quad .$$

Given current values θ, θ_0 and a single data point (x, y), write an expression for a gradient-descent update to θ in terms of θ, θ_0, x, and g and step-size η.

Name: ______________________________

(g) (5 points) Consider a very simple neural network, with two CC units in the hidden layer, and a single output with a sigmoid. Our goal is to use this network to correctly classify the familiar negative XOR data set:



Indicate values for the weights in this network that will result in a correct classification of all four points.

$$h(x) = \sigma(w_0^2 + w_1^2 h_c(x; (w_{11}^1, w_{21}^1), w_{01}^1) + w_2^2 h_c(x; (w_{12}^1, w_{22}^1), w_{02}^1))$$

$w_{01}^1 =$ ______________ $w_{11}^1 =$ ______________ $w_{21}^1 =$ ______________

$w_{02}^1 =$ ______________ $w_{12}^1 =$ ______________ $w_{22}^1 =$ ______________

$w_0^2 =$ ______________ $w_1^2 =$ ______________ $w_2^2 =$ ______________

Work space