10-301/601 Machine Learning
Spring 2024
Exam 2
03/28/2024
Time Limit: 120 minutes

Name: Zoe Rudnick
Andrew ID: zrudnick
Room: DH A302
Seat: G01
Exam Number: 339

Instructions:

- Verify your name and Andrew ID above.
- This exam contains 18 pages (including this cover page).
  The total number of points is 48.
- Clearly mark your answers in the allocated space If you have made a mistake, cross out the invalid parts of your solution, and circle the ones which should be graded.
- Look over the exam first to make sure that none of the 18 pages are missing.
- No electronic devices may be used during the exam.
- Please write all answers *darkly* in pencil or in pen.
- You have 120 minutes to complete the exam. Good luck!

Instructions for Specific Problem Types

For "Select One" questions, please fill in the appropriate bubble completely:

Select One: Who taught this course?

- ● Matt Gormley
- ○ Marie Curie
- ○ Noam Chomsky

If you need to change your answer, you may cross out the previous answer and bubble in the new answer:

Select One: Who taught this course?

- ● Henry Chai
- ○ Marie Curie
- ⊗ Noam Chomsky

For "Select all that apply" questions, please fill in all appropriate squares completely:

Select all that apply: Which are instructors for this course?

- ■ Matt Gormley
- ■ Henry Chai
- ■ Hoda Heidari
- □ I don't know

Again, if you need to change your answer, you may cross out the previous answer(s) and bubble in the new answer(s):

Select all that apply: Which are the instructors for this course?

- ■ Matt Gormley
- ■ Henry Chai
- ■ Hoda Heidari
- ⊗ I don't know

For questions where you must fill in a blank, please make sure your final answer is fully included in the given space. You may cross out answers or parts of answers, but the final answer must still be within the given space.

Fill in the blank: What is the course number?

10-601 10-~~6~~301

1 Optimization, Regularization, and Modeling (8 points)

1. Given a dataset with n points $(\mathbf{x}^{(i)}, y^{(i)})$ where $\mathbf{x}^{(i)}$ is the feature vector of the ith point and $y^{(i)}$ is the corresponding true output, a linear regression model predicts the output as $\hat{y}^{(i)} = \mathbf{w}^T\mathbf{x}^{(i)} + b$, where $\mathbf{w} \in \mathbb{R}^M$ is the weight vector and $b \in \mathbb{R}$ is the bias. The L2 regularized Mean Squared Error (MSE) objective function is defined as:

$$J_{\text{MSE}}(\mathbf{w}, b) = \left(\frac{1}{n}\sum_{i=1}^{n}(\hat{y}^{(i)} - y^{(i)})^2\right)$$
$$J_{\text{reg}}(\mathbf{w}, b) = J_{\text{MSE}}(\mathbf{w}, b) + \lambda\|\mathbf{w}\|^2$$

where λ is the regularization parameter that controls the amount of regularization, and $\|\mathbf{w}\|^2$ is the L2 norm of the weight vector $\mathbf{w}$.

(a) (2 points) **Short Answer:** Derive $\frac{\partial J_{\text{reg}}}{\partial \mathbf{w}}$, i.e. the partial derivatives of $J_{\text{reg}}(\mathbf{w}, b)$ with respect to $\mathbf{w}$. *Hint:* Note that $\frac{\partial J_{\text{MSE}}}{\partial \mathbf{w}} = \frac{2}{n}\sum_{i=1}^{n}(\hat{y}^{(i)} - y^{(i)})\mathbf{x}^{(i)}$

(b) (2 points) **Short Answer:** Derive $\frac{\partial J_{\text{reg}}}{\partial b}$, i.e. the partial derivatives of $J_{\text{reg}}(\mathbf{w}, b)$ with respect to b.

(c) (2 points) **Short answer:** Why do we generally avoid regularizing the bias term b in linear regression?

(d) (2 points) **Select all that apply:** Our objective function, L2 regularized MSE, is strictly convex. Which of the following is true?

- ☐ Solving for the optimal parameters in closed form will always be more computationally efficient than solving for them with gradient descent.
- ☐ The objective has a unique global minimum and gradient descent converges to that minimum, if it converges.
- ☐ Stochastic gradient descent and gradient descent will typically converge to different local minima of the function.
- ☐ Nonce of the above

2 Logistic Regression (7 points)

1. (2 points) A medical dataset contains information on patients and whether they are high risk (Yes $\equiv$ 1 or No $\equiv$ 0). You decide to use binary logistic regression to predict risk based on two features: Weight (kg) and Height (cm). The training dataset consists of 20 patients, not shown. The test dataset contains the following information for 4 patients:

Patient	Weight	Height	High Risk?
1	80	170	1
2	60	180	0
3	70	120	0
4	100	150	1

Select all that apply: After training, the weights are $w_0 = 0$ (intercept term), $w_1 =$ 0.2 (Weight term), w_2 = -0.1 (Height term). For which of the test patients does the model make an *incorrect* prediction?

- ☐ $i = 1$
- ☐ $i = 2$
- ☐ $i = 3$
- ☐ $i = 4$
- ☐ None of the above

2. (1 point) **Fill in the blank:** *The decision boundary for binary logistic regression is* _____. **Select one.**

- ◯ linear
- ◯ quadratic
- ◯ exponential
- ◯ None of the above

3. Suppose we have a dataset $\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^{N}$ where each data point $(\mathbf{x}^{(i)}, y^{(i)})$ is sampled i.i.d. from a probability distribution $p^*(\mathbf{x}, y)$.

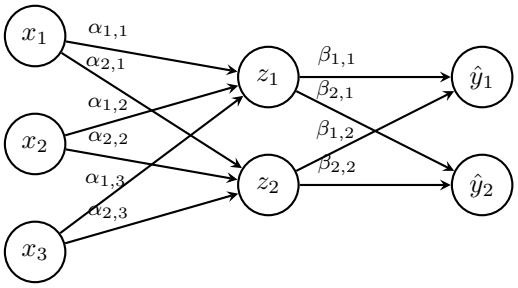
   (a) (2 points) **Select all that apply:** Which of the following objective functions could be maximized or minimized to obtain a value of $\boldsymbol{\theta}$ for a binary logistic regression model?

   - ☐ $\prod_{i=1}^{N} p_{\boldsymbol{\theta}}(y^{(i)} \mid \mathbf{x}^{(i)})$
   - ☐ $\log \prod_{i=1}^{N} p_{\boldsymbol{\theta}}(y^{(i)} \mid \mathbf{x}^{(i)})$
   - ☐ $-\log \prod_{i=1}^{N} p_{\boldsymbol{\theta}}(y^{(i)} \mid \mathbf{x}^{(i)})$
   - ☐ $\sum_{i=1}^{N} \log p_{\boldsymbol{\theta}}(y^{(i)\mid} \mid \mathbf{x}^{(i)})$
   - ☐ $-\sum_{i=1}^{N} \log p_{\boldsymbol{\theta}}(y^{(i)} \mid \mathbf{x}^{(i)})$
   - ☐ None of the above

   (b) (2 points) **Short Answer:** Neural the Narwhal has a new dataset $\mathcal{D}'$ sampled from a different probability distribution with a single parameter $\theta \in \mathbb{R}$. Instead of finding just the value of θ that maximizes the likelihood of their dataset, Neural decides to find the values of both θ and of N' (the size of $\mathcal{D}'$) that maximizes its likelihood. Will Neural's plan help them make predictions on unseen data? Briefly justify your answer in 1-2 sentences.

3 Neural Networks & Backpropagation (17 points)

1. Consider the neural network with 1 hidden layer shown below for a binary classification problem, where $\mathbf{x} \in \mathbb{R}^3$ is the input feature vector and $\mathbf{y} \in \mathbb{R}^2$ is a one-hot vector representing the correct class. Note: this network does not contain bias terms.



$$a_1 = \alpha_{1,1}x_1 + \alpha_{1,2}x_2 + \alpha_{1,3}x_3$$
$$a_2 = \alpha_{2,1}x_1 + \alpha_{2,2}x_2 + \alpha_{2,3}x_3$$
$$z_j = \max(0, a_i),\ \forall j \in \{1, 2\}$$
$$b_1 = \beta_{1,1}z_1 + \beta_{1,2}z_2$$
$$b_2 = \beta_{2,1}z_1 + \beta_{2,2}z_2$$
$$\hat{y}_k = \exp(b_k)/(\exp(b_1) + \exp(b_2)),\ \forall k \in \{1, 2\}$$
$$\ell = -\sum_{k=1}^{2} y_k \log(\hat{y}_k)$$

(a) (2 points) **Numerical answer:** Given $\mathbf{x} = [1, 2, 0]^T$, $\alpha_{j,i} = 1\ \forall j, i$, $\beta_{k,j} = 1\ \forall k, j$. Compute b_2. (You should ignore these numerical values for all subsequent questions.)

(b) (2 points) **Math:** What is the chain of partial derivatives needed by symbolic differentiation to calculate the derivative $\frac{\partial \ell}{\partial \alpha_{j,i}}$ for $i \in \{1, 2, 3\}$ and $j \in \{1, 2\}$?

Your answer should be in the form: $\frac{\partial \ell}{\partial \alpha_{j,i}} = \frac{\partial ?}{\partial ?}\frac{\partial ?}{\partial ?}\ldots$ Make sure each partial derivative $\frac{\partial ?}{\partial ?}$ in your answer cannot be decomposed further into simpler partial derivatives. You may intersperse summations between the $\frac{\partial ?}{\partial ?}$ terms.

(c) (3 points) **Math:** What is the sequence of partial derivatives *stored* by the backpropagation algorithm before it computes *any* of the derivatives $\frac{\partial \ell}{\partial \alpha_{j,i}}$ for $i \in \{1, 2, 3\}$ and $j \in \{1, 2\}$?

Your answer should be in the form of a list: $[\frac{\partial ?}{\partial ?}, \frac{\partial ?}{\partial ?}, \ldots, \frac{\partial ?}{\partial ?}]$, such that each item is stored by backpropagation before all items that appear after it in the list. Make sure each partial derivative $\frac{\partial ?}{\partial ?}$ in your answer cannot be decomposed further into simpler partial derivatives.

(d) (2 points) **Math:** Write an expression for how backpropagaion computes $\frac{\partial \ell}{\partial \alpha_{j,i}}$ for $i \in \{1, 2, 3\}$ and $j \in \{1, 2\}$, after the algorithm has stored all the partial derivatives in your list from the previous question.

Your answer should be in the form: $\frac{\partial \ell}{\partial \alpha_{j,i}} = \frac{\partial ?}{\partial ?} \frac{\partial ?}{\partial ?} \ldots$

(e) (3 points) Complete the stochastic gradient descent implementation below in order to update $\alpha_{j,i}$ and $\beta_{k,j}$. (You may use $\frac{\partial \ell}{\partial \alpha_{j,i}}$ and/or $\frac{\partial \ell}{\partial \beta_{k,j}}$, if needed.)

Initialize weights $\alpha_{j,i}$ and $\beta_{k,j}$ randomly
Choose learning rate η
for each epoch **do**
  for each training example $(\mathbf{x}, \mathbf{y})$ **do**
    // Compute gradients

    // Update weights

(f) (2 points) Yay! You just finished training your network using stochastic gradient descent. But, Neural the Narwhal tests it out, tells you that it doesn't classify well enough yet, and suggests you add 3 more neurons to the hidden layer. He also suggests prepending a bias term to your input, $\mathbf{x}$.

With these updates to your network architecture, what are the new dimensions of the weights matrix, $\boldsymbol{\alpha}$? Express your answer as $\boldsymbol{\alpha} \in \mathbb{R}^{r \times c}$, where r is the number of rows and c is the number of columns.

(g) (1 point) **True or False:** If we switch all nonlinear functions in the given neural network to the identity function, then the resulting neural network will behave identically to a pair of linear regression models.

◯ True

◯ False

2. (2 points) **Short answer:** Describe three differences between a neural network diagram and a computation graph diagram.

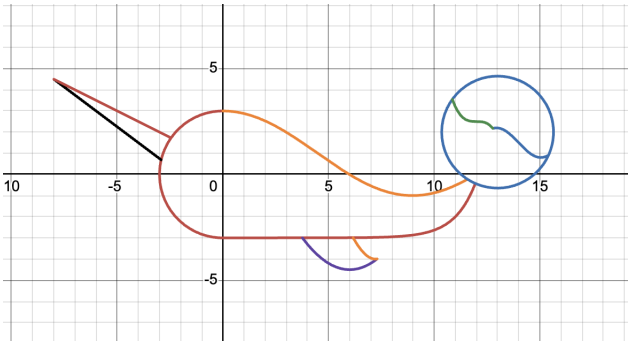
1.

2.

3.

4 Learning Theory (6 points)

1. Desmos is a powerful browser-based graphing software. It is so powerful that it can even create this graph of a narwhal; we will call this the special narwhal equation.



(a) (2 points) **Numerical Answer:** We would like to learn the special narwhal equation, but we only know that for a given point, whether it is or is not on the line calculated by the narwhal equation. However, we don't really know where to start, so we try brute-forcing the narwhal equation. We try functions where:

- The function is a piecewise function with 6 parts.
- Each piecewise equation is a polynomial with 5 degrees of freedom.
- Each piecewise equation has integer coefficients in the range $[-5, 5)$
- Each piecewise equation has an integer upper and lower bound of the interval on which the piece is defined. The lower bound is any integer in the range $[-10, 0)$ and the upper bound is any integer in the range $[0, 10)$.

How many points do we need to achieve error ϵ with probability $(1 - \delta)$? *Show your work.* Your final answer can include a log.

(b) (1 point) **Short Answer:** After discussing with a friend at Desmos, you realize that the VC-dimension of the set of all functions which can be plotted in Desmos is ∞ (i.e. this set can shatter datasets of arbitrarily large, finite size). Explain the theoretical implications of this for train error and test error.

2. The inequality below characterizes the relationship between the true error $R(h)$ and the empirical error $\hat{R}(h)$ for any hypothesis $h \in \mathcal{H}$ with VC-dimension $VC(\mathcal{H})$.

$$R(h) \leq \hat{R}(h) + \mathcal{O}\left(\sqrt{\frac{1}{N}\left(VC(\mathcal{H}) + \log(\frac{1}{\delta})\right)}\right)$$

(a) (1 point) **True or False:** Increasing $VC(\mathcal{H})$ always decreases the true error $R(h)$.

- ◯ True
- ◯ False

(b) (2 points) **Short answer:** $VC(\mathcal{H})$ can be seen as a measure of model complexity. Explain how adding regularization to the model also decreases its VC-dimension. What is the benefit of doing so for model selection?

5 Societal Impacts of ML (10 points)

1. (1 point) **Select one:** Suppose we have a model that has a 0.5 error rate on each of two distinct groups in the dataset. Which of the following fairness metrics will **always** be satisfied?

   ◯ False Negative Rate (FNR) parity

   ◯ False Positive Rate (FPR) parity

   ◯ Error parity

   ◯ A, B, and C

   ◯ All of the above

   ◯ None of the above

2. Markov is trying to build a model on predicting whether Cognitive Machine University (CMU') will admit a student. They are given the following dataset, split into two groups (Red/Blue), with entirely binary features.

   - **GPA above 3.7:** 1 indicates that the student has a GPA above 3.7 reported on their application, 0 otherwise
   - **Legacy:** takes on the value 1 if the student comes from a legacy family, 0 otherwise
   - **Athelete:** 1 if student is recruited for sports, 0 otherwise
   - **Admitted?:** shows the true label, 1 meaning the student was actually admitted, 0 otherwise

Protected Attribute	GPA above 3.7	Legacy	Athelete	Admitted?
Red	1	0	1	1
Red	0	1	0	0
Red	1	0	0	0
Red	0	1	1	0
Red	1	0	0	1
Blue	0	0	1	0
Blue	1	1	0	1
Blue	0	0	0	1
Blue	1	0	1	0
Blue	0	1	0	0

   Markov uses a model where we predict 1 if the "majority" of the features for a student is 1 (i.e. at least 2 of the features has value 1 for that student), and 0 otherwise.

(a) (1 point) **Numerical answer:** What is the negative predictive value on the Red group?

(b) (1 point) **Numerical answer:** What is the negative predictive value on the Blue group?

(c) (1 point) **True or False:** Does Markov's majority vote classifier achieve negative predictive value (NPV) parity?

- ◯ True
- ◯ False

(d) (1 point) **Numerical answer:** What is the false positive rate on the Red group?

(e) (1 point) **Numerical answer:** What is the false positive rate on the Blue group?

(f) (1 point) **True or False:** Does Markov's majority vote classifier achieve false positive rate (FPR) parity?

- ◯ True
- ◯ False

(g) (2 points) **Short answer:** Incorporating your responses from the previous parts, briefly discuss the fairness of the model. Can we say that the model is a fair between the Red and Blue groups?

(h) (1 point) **Short answer:** What are some societal impacts of this model. In particular, what consequences might false negatives and false positive predictions have on the applicants?

Do not remove this page! Use this page for scratch work.

Do not remove this page! Use this page for scratch work.

Do not remove this page! Use this page for scratch work.

Do not remove this page! Use this page for scratch work.