CS 542 – Machine Learning

Midterm Exam

Spring 2020

Instructions:

1- Log onto the lecture zoom link
2- Share your video
3- Set an alarm for 1:35pm. You have 1 hour and 15 minutes to solve the exam.
4- Solve the exam using paper and pen
5- Print your name and BU ID clearly on the top right of the first page (the page that will have the solution to Problem 1)
6- Start a new page for each question
7- Stop solving at 1:35pm
8- Take photos of your solutions using your phone
9- Sort the photos based on question number
10- Convert the photos into a single pdf named ***[BUusername]_[first]_[last]_MT.pdf***
11- Submit your pdf file using this submission link
12- You will receive a confirmation message from us that we received your submission. **It is your responsibility to make sure the file contains solutions to all the problems you solved.**

Notes:

** There are five questions. The page has some common formulas*
** If you have any inquiries during the exam please message us in the zoom chat*
** Total points: 85*

Good luck ☺

Q1. [20 points] Short Questions

Answer the following questions in brief one or two sentence answers.

a) [4 points] Suppose you have a training dataset $\mathrm{D} = \{x_i, y_i\}$ where x_i are the inputs and y_i are the class labels, and you are designing a *discriminative* classifier. Which, if any, marginal, joint, or conditional probability distribution function(s) over these variables would you model, and why?

b) [4 points] Suppose you work for a startup that has just designed a new electric vehicle, and you are asked to build a machine learning system to detect engine failure. You have collected data from 1000 tests, only 5 of which led to the engine failing. What family of machine learning methods would you use, and why?

c) [4 points] What is the effect of the class prior on the decision boundary of Linear Discriminant Analysis?

d) [4 points] Alice decides to use Principal Component Analysis to implement image compression. Briefly explain how she should train the algorithm and how she should use it to compress a new image. What parameter controls the amount of compression?

e) [4 points] The following is the formulation for linearly separable SVMs. The formulation maximizes the margin and ensures data points are correctly classified.

$$\min \ \frac{1}{2}\|\mathbf{w}\|^2 \quad s.t.\ (\mathbf{w}^T\mathbf{x}_i + b)y_i \geq 1 \ \ \forall i$$

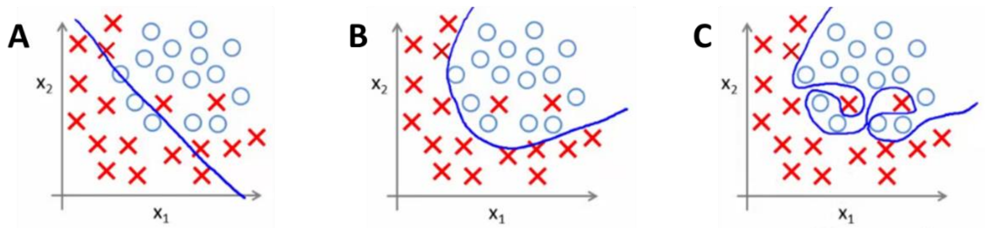
Write the corresponding mathematical formulation for a relaxed version that allows some violation of the constraints for a problem that is not strictly linearly separable.

Q2. [10 points] Regularization and Bias/Variance

Suppose you want to fit a Logistic Regression model to predict whether an email is spam ($y = 1$) or not spam ($y = 0$) based on the frequency of the words "buy" (feature x_1) and "click" (feature x_2). You decide to use polynomial basis functions to represent the input features and to apply regularization. You have fit three models by minimizing the regularized Logistic Regression cost function

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}\left[-y^{(i)}\log\left(h_\theta\left(x^{(i)}\right)\right) - \left(1-y^{(i)}\right)\log\left(1-h_\theta\left(x^{(i)}\right)\right)\right] + \frac{\lambda}{2m}\sum_{j=2}^{n}\theta_j^2$$

for $\lambda = 10^{-2}, 10^{0}, 10^{2}$. The following are sketches of the resulting decision boundaries.



a) [3 points] Which value of λ goes with each of the plots?

A: **B:** **C:**

b) [2 points] You try your models on a test set. Which of the three models will have the highest error due to **bias**?

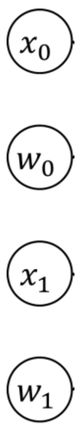
c) [2 points] Which model will have the highest error due to **variance**?

d) [3 points] You plot model complexity *M* (number of polynomials) versus the cost function, computed on the test data, and a similar curve for the training data. Explain how you can use these curves to detect when the model is overfitting and draw an example to illustrate.

Q3. [15 points] Backpropagation

Suppose we want to compute the gradients for the function $h(x) = q(w_0x_0 + w_1x_1)$, where $x = [x_0\ x_1]^T$ is the input vector, $w = [w_0\ w_1]^T$ is the parameter vector, and q is the *tanh* function: $\mathrm{q(u)} = \tanh(u) = (e^u - e^{-u})/(e^u + e^{-u})$. The *tanh* function is also plotted in the appendix.

a) [2 points] Complete the computational graph for this function below by adding two nodes $f_1(u, c) = c * u$ (multiplication), one node $f_2(u, c) = u + c$ (addition), and one node $f_3(u) = q(u)$. Label the nodes clearly with f_1, f_2 or f_3 and leave plenty of space between them.



b) [3 points] Write down the gradient $\frac{\partial f}{\partial u}$ for functions f_1, f_2, f_3.
Hint: note that $tanh'(u) = 1 - tanh(u)^2$.

$\frac{\partial f_1}{\partial u} =$

$\frac{\partial f_2}{\partial u} =$

$\frac{\partial f_3}{\partial u} =$

c) [4 points] Perform a forward pass for $x = [5 \quad 5]^T$ and $w = [1 \quad -1]^T$, **writing values on top of the arrows** in your computational graph. What is the output of the forward pass, *i.e.* $h(x)$?

d) [6 points] Perform a backward pass for the example in (c), **writing values below the arrows** in the graph. What are the gradients of h with respect to x_0, x_1, w_0, w_1?

$\frac{\partial h}{\partial x_0} =$ $\qquad$ $\frac{\partial h}{\partial x_1} =$

$\frac{\partial h}{\partial w_0} =$ $\qquad$ $\frac{\partial h}{\partial w_1} =$

Q4. [20 points] Neural Networks and Deep Learning

Answer the following questions in brief one or two sentence answers.

a) [4 points] How would you design a deep learning architecture to be used for hand gesture recognition. Assume the input to your system is short video clips, and the desired output is a predicted class of hand gesture.

b) [3 points] Name the regularization technique specifically designed for Deep Neural networks and briefly describe how it achieves such regularization.

c) [3 points] Is it recommended to do feature engineering first and then apply deep learning? Contrast deep learning with other machine learning algorithms in terms of feature engineering.

d) [3 points] Suppose we have a neural network with ReLU activation function. Let's say, we replace ReLu activations by linear activations. Would this new neural network be able to approximate a non-linear function? And why?

e) [3 points] Name an example of a data augmentation approach and explain how it helps improve model generalization.

f) [2 points] What is the main difference between a fully-connected and a convolutional network?

g) [2 points] List two ways to downsize feature maps in convolutional neural networks.

Q5. [8 points] Maximum A Posteriori (MAP) Solution

To deal with bias, we can use a Bayesian model and obtain a posterior solution for the parameter. Suppose we are estimating the probability of seeing 'heads' ($\mathrm{x} = 1$) or 'tails' ($\mathrm{x} = 0$) after tossing a coin, with μ being the probability of seeing 'heads'. The probability distribution of a single binary variable $x \in \{0,1\}$ that takes value 1 with probability μ is given by the *Bernoulli* distribution

$$Bern(x|\mu) = \mu^x (1-\mu)^{1-x}$$

Suppose we have a dataset of **independent** coin flips $D = \{x^{(1)}, x^{(2)}, \dots, x^{(m)}\}$ and we would like to estimate μ by maximizing the posterior probability. Recall that we can write down the data likelihood as

$$p(x^{(i)}|\mu) = \mu^{x^{(i)}} (1-\mu)^{1-x^{(i)}}$$

Consider the following prior on μ, which believes that the coin is either fair, or slightly biased towards 'tails':

$$p(\mu) = \begin{cases} 0.5 \ if \ \mu = 0.5 \\ 0.5 \ if \ \mu = 0.4 \\ 0 \ otherwise \end{cases}$$

a) [3 points] Write down the formulation for the likelihood $p(D|\mu)$.

b) [5 points] Write down the posterior estimate for μ under this prior as a function of the likelihood and the prior. *(Hint: use the argmax function).*

Q6. [12 points] Loss functions for Classification

For this question, suppose that we have a classification problem with inputs $x \in R^n$ and corresponding outputs $y \in \{-1, +1\}$. We would like to learn a classifier that computes a linear function of the input $f(x) = w^T x$, and predicts $y = +1$ if $f(x) \geq 0$, or $y = -1$ otherwise.

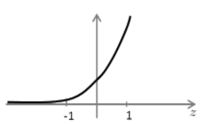
a) [3 points]

(i) What can be said about the correctness of the classifier's prediction if the expression $yf(x) > 0$ is true?

(ii) What if $yf(x) = 0$?

(iii) What if $yf(x) < 0$?

b) [4 points] Suppose we use a loss function $L(z) = L(yf(x))$, where $yf(x)$ is defined as above, to compute the loss for a given training pair of input and output $\{x, y\}$.
What effect does the loss function shown below have on the resulting classifier?



Will this loss function produce a reasonable solution? Explain why or why not.

c) [5 points] Suppose you decide to use the loss $L(z) = exp(-2z - 1)$. Write down the gradient descent algorithm for minimizing this loss on a set of training examples $\{x_i, y_i\}, i = 1, \ldots, m$, using a squared L-2 norm regularizer $R(w) = ||w||^2$ on the parameter vector.

(i) What is the objective function?

(ii) Should it be minimized or maximized?

(iii) What is the corresponding gradient descent update step for w?

Appendix: Common Formulas

Matrix Derivatives

For vectors x, y and matrix A,

$y = Ax$, then $\frac{\partial y}{\partial x} = A$

If $z = x^T A x$, then $\frac{\partial z}{\partial x} = x^T(A + A^T)$. For the special case of a symmetric matrix A, $\frac{\partial z}{\partial x} = 2x^T A$.

Chain Rule

If $\mathrm{z} = \mathrm{f(y)}$ and $\mathrm{y} = \mathrm{g(x)}$, then $\frac{dz}{dx} = \frac{dz}{dy}\frac{dy}{dx} = f'\big(g(x)\big) * g'(x)$

Hyperbolic Tangent Function (tanh)