CS 229A — Spring 2019
Midterm
Instructors: Andrew Ng, Younes Bensouda Mourri
Friday 24th of May 2019

Instructions
The Stanford University Honor Code

A. The Honor Code is an undertaking of the students, individually and collectively:

a. That they will not give or receive aid in examinations; that they will not give or receive unpermitted aid in class work, in preparation of reports, or in any other work that is to be used by the instructor as the basis of grading;

b. That they will do their share and take an active part in seeing to it that others as well as themselves uphold the spirit and letter of the Honor Code.

B. The faculty on its part manifests its confidence in the honor of its students by refraining from proctoring examinations and from taking unusual and unreasonable precautions to prevent the forms of dishonesty mentioned above. The faculty will also avoid, as far as practicable, academic procedures that create temptations to violate the Honor Code.

C. While the faculty alone has the right and obligation to set academic requirements, the students and faculty will work together to establish optimal conditions for honorable academic work.

Acknowledge and accept the Honor Code:

Before delving into the questions, make sure to read all the problems. There are 7 problems and the midterm is 14 pages long. Problems decompose as follows:

- Problem 1: 20 points
- Problem 2: 15 points
- Problem 3: 20 points + 2 Bonus points
- Problem 4: 15 points + 1 Bonus point
- Problem 5: 10 points
- Problem 6: 20 points + 1 Bonus point
- Problem 7: 10 points

Name:

SUNET Id:
Assume that \( \theta \in \mathbb{R}^n \), \( a \in \mathbb{R}^n \), \( A \in \mathbb{R}^{n^2} \), and \( A \) is symmetric. Then,

\[
\frac{\partial a^T \theta}{\partial \theta} = a \quad (V1)
\]

\[
\frac{\partial \theta^T A \theta}{\partial \theta} = 2A\theta \quad (V2)
\]

Assume that \( x \in \mathbb{R}^n \), \( W \in \mathbb{R}^{d \times n} \), \( f : \mathbb{R} \rightarrow \mathbb{R} \) element-wise function, and \( \delta \in \mathbb{R}^d \). Then,

\[
\frac{\partial x}{\partial x} = 1_{n \times n} \text{ with } 1_{n \times n} \text{ the identity matrix} \quad (B1)
\]

\[
\frac{\partial Wx}{\partial x} = W \quad (B2)
\]

\[
\delta^T \frac{\partial Wx}{\partial W} = \delta x^T \quad (B3)
\]

\[
\frac{(f(x_1), \ldots, f(x_n))}{\partial x} = \operatorname{Diag}[f'(x_1), \ldots, f'(x_n)] \quad (B4)
\]

Assume that \( y \in \mathbb{R}^K \), a one-hot vector, and \( z \in \mathbb{R}^K \). If

\[
\begin{cases}
\hat{y} = \operatorname{Softmax}(z) \\
J = \operatorname{CE}(y, \hat{y})
\end{cases}
\]

with

\[
\text{Softmax}(z) = \left( \frac{1}{Z} e^{z_1}, \frac{1}{Z} e^{z_2}, \ldots, \frac{1}{Z} e^{z_d} \right) \quad (Z \text{ the normalization constant})
\]

and

\[
\text{CE}(y, \hat{y}) = -\sum_{k=1}^K y_k \log(\hat{y}_k)
\]

Then,

\[
\frac{\partial J}{\partial z} = (\hat{y} - y)^T \quad (B5)
\]

Cost function of linear regression (with regularization):

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

Cost function of logistic regression (with regularization):

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

Sigmoid function:

\[
\sigma(z) = \frac{1}{1 + e^{-z}}
\]
Problem 1: Linear Regression

In this question, $Y$ is the GPA of seniors (out of 5) and it is a continuous variable. We use the following features:

- $X_1$: number of hours studied per week
- $X_2$: GPA as a junior

1. (5 points) In Linear Regression, we define $Y = h_\theta(X)$ with $X = (1, X_1, X_2)$ (a row vector). Write down the explicit expression of $h_\theta(X)$ as a function of $\theta \in \mathbb{R}^3$.

2. (5 points) We fit a linear regression on the data mentioned above. We get the following values for $\theta$: $\theta_0 = -1.5$, $\theta_1 = \frac{1}{15}$, $\theta_2 = 0.75$. What is the expected GPA of a student who studied 45 hours a week and had a GPA of 3 as a junior? Give a numerical answer.

3. (5 points) After releasing the model on Carta as a GPA simulator for students, you have gathered feedback from students. The main feedback you received is that students are extremely unhappy when the model predicts a higher GPA than their actual GPA whereas students do not really care if the prediction is below their actual GPA. How can you incorporate this feedback as you build a new model? Specifically, describe exactly how you would still use linear regression and the same features. Also explain why your new model works.

4. (5 points) After a careful analysis you realize that the relationship between $X_1$ and $Y$ is not linear but logarithmic. Can you still model this relationship with a linear regression?
Problem 2: Regularized Logistic Regression

The cost for a logistic regression with $L_2$ regularization is:

$$J(\theta) = -\frac{1}{m} \sum_{i=1}^{m} \left( y^{(i)} \log \left( h_\theta(x^{(i)}) \right) + \left(1 - y^{(i)}\right) \log \left(1 - h_\theta(x^{(i)})\right) \right) + \frac{\lambda}{2m} \sum_{j=0}^{n} \theta^2_j$$

Note 1: Here, we regularize the intercept $\theta_0$. We adopt this convention for calculus simplicity, even though we saw in class that intercept should not be regularized.

Note 2: Use the following convention: matrix $X$ is made of row vectors $x^{(i)}$ such that row $i$ of $X$ is $x^{(i)}$. $Y$ is a vector whose $i$-th component is $y^{(i)}$. Your matrix $X$ has $m$ training examples. As seen in lecture for linear regression, the convention is that $x_0^{(i)} = 1$ for all $i$.

1. (5 points) For a vector $x^{(i)}$, what is $h_\theta(x^{(i)})$? Remember this is logistic regression.

2. (5 points) Compute the gradient of the cost function, i.e., compute $\frac{\partial J}{\partial \theta_j}$ $\forall j$.
   Note: you can use the formula of the gradient of the non-regularized cost without justification.
3. (5 points) Give a vectorized expression of $\frac{\partial J}{\partial \theta}$, i.e. give an expression of the gradient that does not involve summations.
Problem 3: Neural Networks

Consider a neural network whose feedforward equations for a single input vector \( x \in \mathbb{R}^n \) are defined as follows (\( W_i \) denotes a weight matrix, \( b_i \) denotes a bias vector):

\[
\begin{align*}
  z_1 &= W_1 x + b_1 \\
  a_1 &= \sigma(z_1) \\
  z_2 &= W_2 a_1 + b_2 \\
  \hat{y} &= \text{Softmax}(z_2) \\
  J &= \text{CE}(y, \hat{y}) 
\end{align*}
\]

Note 1: \( \sigma \) denotes the sigmoid function:

\[
\sigma(z) = \frac{1}{1 + e^{-z}}
\]

Note 2: For \( z \in \mathbb{R}^d \), the softmax function is defined as:

\[
\text{Softmax}(z) = \left( \frac{1}{Z} e^{z_1}, \frac{1}{Z} e^{z_2}, \ldots, \frac{1}{Z} e^{z_d} \right) \quad \text{with } Z \text{ the normalization constant}
\]

Note 3: Those conventions are the same as the ones used in the section of backpropagation. Here, since the bias is represented by the \( b_i \) vector, the input vector is \( x = (x_1, \ldots, x_n)^T \) where \( (x_i) \) represents a feature: it does not include a 1 as a first coordinate.

Note 4: In this case, \( x \) is a column vector. So in this problem, each column of \( X \) represents a training example.

1. (5 points) \( a_1 \) has shape \((d_1, 1)\). This problem is a \( K \)-class classification problem. For a single training example input \( x \in \mathbb{R}^n \), give the shapes of the following elements:

- \( W_1 \)
- \( b_1 \)
- \( z_1 \)
- \( W_2 \)
- \( b_2 \)
- \( z_2 \)

2. (3 points) How many trainable parameters does this neural network have?

3. (12 points) Derive the following gradients:

- \( \frac{\partial J}{\partial W_1} \)
- \( \frac{\partial J}{\partial x} \)
- \( \frac{\partial J}{\partial b_1} \)
4. (2 points) **Bonus**: If you were to implement and train this model, you would have to tune non-trainable parameters. Give one example of such a parameter.
Problem 4: MLE or Least-Squares?

In this question, we are trying to find a relationship between features \( X \) and continuous variables \( Y \). In the Linear Regression model, we usually make the following assumptions:

- \( y^{(i)} = x^{(i)} \theta + \epsilon^{(i)} \) (Here, \( x^{(i)} \) is a row-vector, \( \theta \) a column vector and \( y^{(i)} \) the associated outcome)
- \( \epsilon^{(i)} \) is independent of \( X \)
- \( \mathbb{E}[\epsilon^{(i)}] = 0, \text{Var}[\epsilon^{(i)}] = \sigma^2 \)
- Random variables \( \epsilon^{(i)} \) are i.i.d

In this question, we will make a stronger assumption about \( \epsilon \). We will assume that \( \epsilon \) is a normal random variable. Therefore, \( \epsilon \sim \mathcal{N}(0, \sigma^2) \).

1. (5 points) With this stronger assumption, we can prove that \( y^{(i)} \sim \mathcal{N}(x^{(i)} \theta, \sigma^2) \). Therefore, the likelihood of seeing one sample \((x^{(i)}, y^{(i)})\) is:

\[
L_i(\theta) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(y^{(i)} - x^{(i)} \theta)^2}{2\sigma^2}}
\]

*Note*: you do not have to prove that.

Assuming independent samples, compute the likelihood, \( L(\theta) \), of seeing the dataset, i.e the global likelihood over all samples:

\[
\left\{ \left(x^{(1)}, y^{(1)}\right), \ldots, \left(x^{(m)}, y^{(m)}\right) \right\}
\]

2. (5 points) Prove that the log-likelihood is given by:

\[
\log L(\theta) = -\frac{1}{2\sigma^2} \sum_{i=1}^{m} \left( y^{(i)} - x^{(i)} \theta \right)^2 - \frac{m}{2} \log(2\pi) - m \log(\sigma)
\]
3. (5 points) Show that maximizing the log-likelihood is equivalent to minimizing the sum of least squares.

4. (1 point) **Bonus**: you just proved that the MLE method yields the same result as the least-squares method. However, MLE has one drawback compared to the least-squares method! What is it?
Problem 5: True/False

Circle True/False. No justification needed. For each question, we will grade as follows:

- Correct: +2
- Not answered: +1
- Incorrect: 0

1. True/False It is recommended to use PCA to prevent overfitting because we can reduce the number of dimensions.

2. True/False Using regularization will be at least as effective as PCA when addressing overfitting.

3. True/False In general, decreasing λ (regularization parameter in regularized logistic regression) when overfitting reduces bias.

4. True/False In general, adding features when underfitting reduces bias.

5. True/False Recall is defined as $\frac{TP}{TP + FN}$ (with TP: True Positives, FN: False Negatives)
Problem 6: K-means algorithm

Suppose we have the following points in one dimension:

\[ x_1 = -7, x_2 = -5, x_3 = -4, x_4 = 3, x_5 = 4, x_6 = 6, x_7 = 7 \]

Run the 3-means clustering until convergence with the following initialization:

\[ \mu_1 = -8.5, \mu_2 = -3, \mu_3 = 8 \]

Note: in the case of a tie, assign the point to the class with a lower number (i.e. if one point is tied between class 1 and class 2, assign it to class 1).

1. (4 points) Draw a number line to help you visualize what is happening.

2. (4 points) How many iterations did you perform? Note: do not double count! Therefore if iterations \( n \) and \( n + 1 \) give the same result, the algorithm converges in \( n \) iterations.

3. (4 points) What is the final assignment?

4. (4 points) What are the final centroids?

5. (4 points) Remember that the loss in the \( K \)-means algorithm is given by:

\[
\text{Cost} = \frac{1}{m} \sum_{i=1}^{m} ||x_i - \mu_{z_i}||^2 \text{ with } z_i \text{ the cluster of point } i
\]

Compute the final loss.

6. (1 point) Bonus: as you can see, the final assignment is not optimal. There is a better assignment of the centroids. What is it?
Problem 7: Support Vector Machines

Figure 1: Separating hyperplane representation

In this question point A is denoted by $x^{(i)}$. The separating hyperplane (line) is defined as the set of all $x$ verifying $w^T x + b = 0$ with $w$ the normal vector and $b$ the intercept term. The training examples are $x^{(i)}$ and their corresponding labels are $y^{(i)}$.

1. (4 points) We define point $B = x^{(i)} - \gamma^{(i)} \frac{w}{||w||}$. Find the value of $\gamma^{(i)}$ in terms of $w, b, x^{(i)}$.

2. (4 points) Assume we want to derive the maximal margin SVM (assuming all points are linearly separable). What is the objective function? What are the optimization constraints? Write your answer in terms of $y^{(i)}, x^{(i)}, w, b, \gamma$.

- Objective:

- Constraints:
3. (2 points) Why does solving this optimization problem make sense? *Hint:* no mathematical explanation is required, use your intuition!