Problem 0: Problem Set Logistics

Please read carefully! This problem contains essential information regarding problem set logistics.

1. (5 points) Logistics

   • **Submission**: the class uses Gradescope to grade Problem Sets and Projects. Enroll using the following code: 9B4B7B. Submit on Gradescope. **DO NOT** send the teaching staff emails with your submissions, they will be discarded.

   • **Submission rules**: Here are some simple rules to ensure a fast and fair grading! Do not re-write the question in the submission: just write your answer. For True/False questions, only justify when necessary. When a justification is needed, respect the word limit. Finally, the **most important** part is to correctly assign pages of your submission on Gradescope. If you assign pages incorrectly or forget to assign pages, **points will be deducted**. **Note**: you can assign pages after the deadline as long as the .pdf is uploaded. Therefore, start by uploading your submission and then properly assign pages without stress.

   • **Regrade requests**: regrade requests **must be** submitted through Gradescope. **DO NOT** send the teaching staff emails with your regrade requests, they will be discarded. You will have **one week** after grades are released to submit regrade requests.

   • **Late day policy**: you will submit 3 Problem Sets, 1 Project Proposal, 1 Project Milestone, 1 Project Final Report and 1 Project Poster. The Project Final Report and the Project Poster are due on Gradescope on **December 12th, 2019 at 11:59 pm**. All the other submissions will be due on Gradescope at 11:59pm on Tuesdays. You have **three late days** over the quarter. Late days extend the deadline by one calendar day. Late days **cannot** be used on the Project Final Report/Project Poster. Elements submitted after the late day **will not be graded**. You can not use two late days on the same assignment.

   **Note**: deadlines on Gradescope are set-up automatically.

**Acknowledge and accept the aforementioned rules**

2. (5 points) 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**

**Problem 1: Linear Regression**

In this problem, we work with the boston housing data. The features represent the characteristics of a house (surface area, number of rooms, etc.) and the outcome variable is the price of the house. Therefore, the goal is to predict the price of a house given its features.

1. (5 points) We first explore the data to determine which features will be helpful in predicting the price. For 3 different features (age, number of bathrooms, number of rooms) we plot the price against the feature specified (Figure 1). Which feature will be the most useful to predict the price? Which feature will be the less useful to predict the price? In other words, rank the features by their predictive power. *Hint: Using only the plots, which model do you think would have the lowest error and why? No code required.*

2. (5 points) We decide to fit a linear regression using gradient descent. As we have seen in lectures, the gradient descent algorithm depends on two parameters: \( \alpha \), the learning rate and \( t \), the number of iterations. We tried three different sets of parameters \((\alpha, t)\). For each of those three sets, we plot the cost function against the number of iterations (Figure 2). Looking at those plots, you can tell that the learning parameters are not well chosen. For each figure, give one parameter change (i.e. increase/decrease \( \alpha/t \)) that would increase the performance (i.e. reduce the cost) as well as a one-two line(s) explanation. *Note: for each figure, you are only asked to change \( \alpha \) or \( t \) but not both. We could do both at the same time but for simplification purposes, we will focus on changing one parameter at a time.*

3. (5 points) We run a linear regression using only one feature: the number of rooms (Figure 3). You visit two houses: the first one has 3 rooms, the second one has 8 rooms. According to the model that is shown on Figure 3, what are the predicted prices of each house that you visited?

4. (5 points) After careful analysis, the relationship between the price and the age of the house does not seem to be linear. After performing some data transformation/augmentation, we are able to fit the following "line" (Figure 4). Explain how we were able to capture such a non-linear relationship.
Figure 1: Plot of the price against the age (a), the number of bathrooms (b), the number of rooms (c)
Figure 2: Plot of the cost function for three different sets of parameters
Figure 3: Price of the house against the number of rooms. The fitted line is represented in red.

Figure 4: Price of the house against the age of the house. The fitted line is represented in red.
Problem 2: Regularization

In class, we saw that the cost function for linear regression is:

$$\frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2$$

For this problem, we will try a different function, called the regularized cost function. Regularization is often useful in Machine Learning. Specifically, it allows models to generalize well (i.e. make better predictions to unseen data when you have many features). The new cost function has an extra penalty term and is defined as:

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

Note: You will see regularization in more detail in the upcoming weeks: this is just an introduction. The context given is enough for you to solve this problem.

1. (5 points) Derive the gradient of the regularized cost function.

2. (5 points) Now that you have your gradient, how would you update theta?

3. (5 points) Using the update rule you mentioned above, explain how this new penalty term affects the weights?
Problem 3: Logistic Regression

In this problem, we use medical data. The features represent the characteristics of a tumor (size, darkness, depth, etc.) and the outcome variable is 1 if the tumor is malignant and 0 if it is benign. The goal is to predict the type of tumor given its features.

1. (5 points) We plot the type of tumor against the tumor size (Figure 5). Give one reason why linear regression will work poorly on this problem. Note: there are several but one is enough.

![Figure 5: Type of tumor against the tumor size](image)

2. (5 points) Logistic regression does not predict the outcome variable. It predicts the probability that the outcome variable belongs to class 1 given the data. It is defined as:

\[
P(Y = 1 | X) = \frac{1}{1 + e^{-X\theta}}
\]

What is the expression for \( P(Y = 0 | X) \)?

3. (5 points) When the probability is computed, we need to assign a class to each tumor. Given the probability, how do we decide if a tumor is benign or malignant?

4. (5 points) Another way to classify a tumor is to choose the class with the biggest probability. For example, if \( P(Y = 0 | X) > P(Y = 1 | X) \), we assign 0. Show that this decision rule is equivalent to the previous one (i.e. it always leads to the same assignment)
Problem 4: Normal Equation for Regularized Linear Regression

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

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

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) Give a vectorized expression of $J(\theta)$. In other words, write $J(\theta)$ as a function of $X$, $Y$, $\theta = (\theta_0, \theta_1, \ldots, \theta_n)^T$, $\lambda$, $m$ without summations. Note: justification for vectorized expression is not required.

2. (5 points) Derive the gradient of $J$ with respect to $\theta$, i.e. compute $\nabla_{\theta} J(\theta)$. Mention explicitly the formulas you use.

3. (5 points) Derive the normal equations for this cost function. In other words, find $\theta$ that minimizes the cost function $J(\theta)$ as a function of $X$, $Y$, $\lambda$, and $m$. State any assumption you make in order to derive the solution.
Problem 5: Confusion matrix - ROC

When doing classification, the most natural way to evaluate your model is through its accuracy. Accuracy is defined as the number of correct predictions over the total number of predictions. However, this is not always a good evaluation metric. If classes are imbalanced (i.e. one class has many more examples than the other classes), it is really easy to get a good score by just predicting the most popular class all the time. For example, assume you are building a model to test whether someone is contaminated with a virus or not. The virus has a prevalence < 1%. Predicting always 'not-contaminated' will give you a classifier with an accuracy that is higher than 99%! Yet, this is not satisfying, because your model is useless. Furthermore, miss-classifications are not always equal: it is worse to tell someone who is contaminated that they are not contaminated (because that person will not get treated and die...) than the other way around. Therefore, we need to build more descriptive metrics taking into account miss-classifications. This is the confusion matrix and is defined as:

<table>
<thead>
<tr>
<th>Predicted 1</th>
<th>Predicted 0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual 1</td>
<td>TP</td>
</tr>
<tr>
<td>Actual 0</td>
<td>FP</td>
</tr>
</tbody>
</table>

- TP (True Positives): number of datapoints correctly labelled 1
- TN (True Negatives): number of datapoints correctly labelled 0
- FP (False Positives): number of datapoints incorrectly labelled 1
- FN (False Negatives): number of datapoints incorrectly labelled 0

We will talk more about this later in the Coursera videos, but for the purpose of this problem, you have all the context needed to solve it.

1. (4 points) The True Positive Rate (TPR, also called sensitivity or recall) is defined by the percentage of actual positive data points correctly labelled. The False Positive Rate (FPR) is equal to 1 - TNR. (where TNR, True Negative Rate, is defined like the TPR but with actual negative data points instead). Define TPR and FPR using TP, TN, FP, FN.

2. (4 points) One way to assign classes given the probability is to choose a threshold $c$ such that if $P(Y = 1 | X) > c$, we assign 1. The higher the $c$, the more conservative you are in assigning class 1. What is the value of the TPR for $c = 0$, $c = 1$? Answer the same question for the FPR.

3. (4 points) What is on average the TPR for a random classifier (i.e. a classifier assigning 1 with probability 1/2)? How about the FPR?

4. (4 points) More generally, we can compute the value of (FPR, TPR) for every value of $c$. We then plot those points and that gives us the ROC curve. An example is given below in Figure 6. What is the shape of the ROC for a perfect classifier?

5. (4 points) Using the ROC curve, we can compute the AUC or Area Under the Curve of the ROC. For example, the AUC of the random classifier is $\frac{1}{2}$. Explain in two-three lines why the AUC is a good metric.
Figure 6: ROC curve of our classifier vs. a random classifier