Problem 0: Problem Sets Logistics

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

1. (10 points) Logistics

- **Submission**: the class uses Gradescope to grade Problem Sets and Projects. Enroll using code: M7D3RE. 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 rewrite the question in the submission: just write your answer! For True/False questions without justification (the question will always mention whether a justification is required) **do not** write justifications (if you do so, the question will not be graded!). When a justification is needed, respect the instruction given in terms of length (it will always be specified): if your justification is much longer than the instruction given, it will not be graded. Finally, the **most important** part is to correctly assign pages of your submission to the questions of the Problem Set on Gradescope. If you assign pages incorrectly or forget to assign pages, **you will not receive any points**. In other words, what is graded is what you have submitted and assigned correctly! **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 the grades are released to submit regrade requests.

- **Late day policy**: you will submit 4 Problem Sets, 1 Project Proposal, 1 Project Milestone, 1 Project Final Report and 1 Project Poster on Gradescope. All these submissions will be due at 11:59pm Tuesdays. You have two late periods over the quarter. Late periods extend the deadline to the following Friday (at 11:59pm). Late periods **cannot** be used on the Project Final Report/Project Poster. Elements submitted after the late deadline will not be graded.

  **Note**: deadlines on Gradescope are set-up automatically. Therefore, there is no room for leniency with deadlines. Do not play with fire and always keep some safety time to submit on time.

**Acknowledge and accept the aforementioned rules**

Acknowledged and accepted

2. (10 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

Acknowledged and accepted

Problem 1: Linear Regression

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

   Linear regression measures correlation. Therefore, features are ranked from the most correlated to the
   less correlated as follows:

   number of rooms > age of the house > number of bathrooms

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.

   We recommend the following:
   • Figure 3.a: increase the number of iterations (increase learning rate is also valid)
Figure 1: Plot of the price against the age (a), the number of bathrooms (b), the number of rooms (c)
• Figure 3.b: increase the learning rate (increase the number of iterations is also valid)
• Figure 3.c: decrease the learning rate

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?

• 3 rooms ⇒ $300,000
• 8 rooms ⇒ $800,000

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.

The solution consists of creating non-linear relationships. Indeed, let $y$ be the price of the house, $x_1 = x$, the age of the house be our first feature and $x_2 = x^2$ the age of house squared be our second feature. Then fitting a linear regression on $y$ with these two features yields:

$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 = \theta_0 + \theta_1 x + \theta_2 x^2$$

The above relationship is linear in the features but the relationship modeled between $x$ and $y$ is not.
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. Regularization is often relevant in Machine Learning. Specifically, it allows models to generalize (i.e. predict on new unseen data with performance similar to seen data). The new cost function is equal to the unregularized cost function plus a penalty term penalizing high weights:

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

Note: You will see regularization more in details in the upcoming weeks: this is just an introduction!

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

\[
\forall j \neq 0, \quad \frac{\partial \text{Cost}}{\partial \theta_j} = \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})x_j^{(i)} + \frac{\lambda}{m} \theta_j
\]

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

\[
\forall j \neq 0, \quad \theta_j := \theta_j - \frac{\alpha}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})x_j^{(i)} - \frac{\alpha \lambda}{m} \theta_j = \text{Unregularized Update} - \alpha \frac{\lambda}{m} \theta_j
\]

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

The penalty term shrinks the weights in the update rule. Indeed if \( \theta_j > 0 \) then the new update rule shrinks its value. Same for \( \theta_j < 0 \).
Problem 3: Logistic Regression

In this problem, we work with medical data. The features represent the characteristics of a tumor (size, darkness, depth, etc.) and the outcome variable is the type of a variable: 1 if the tumor is malignant, 0 otherwise. Therefore, 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, 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 of \( P(Y = 0|X) \)?

\[
P(Y = 0|X) = 1 - P(Y = 1|X) = \frac{e^{-X\theta}}{1 + e^{-X\theta}}
\]

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?

*Assign 1 if \( P(Y = 1|X) > \frac{1}{2} \)*
4. (5 points) Another decision rule to assign class 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)

Using this rule, we assign 1 if \( P(Y = 1|X) > P(Y = 0|X) \).

\[
P(Y = 1|X) > P(Y = 0|X) \iff P(Y = 1|X) > 1 - P(Y = 1|X) \iff P(Y = 1|X) > \frac{1}{2}
\]
Problem 4: Confusion matrix - ROC

When doing classification, the most natural metric is the accuracy: how many samples are correctly labelled? However, despite its great simplicity, accuracy has some limitations. First, 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 without doing anything useful. Let’s take the example of a virus with a prevalence lower than 1%. Assume you are building a model to test whether someone is contaminated. Then predicting always “not-contaminated” will give you a classifier with accuracy higher than 99%. Yet, this is not satisfying! Second, misclassifications are not always equal: indeed in the case of a virus test, it is much worse to tell someone with the virus that they are not contaminated (because that person will not take action) than to tell someone without the virus that they are contaminated (a more comprehensive test will show they are not). Therefore, we need to build more descriptive metrics taking into account misclassifications per class. This is the confusion matrix. It is defined as follows:

<table>
<thead>
<tr>
<th></th>
<th>Predicted 1</th>
<th>Predicted 0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual 1</td>
<td>TP</td>
<td>FN</td>
</tr>
<tr>
<td>Actual 0</td>
<td>FP</td>
<td>TN</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. (5 points) The True Positive Rate (TPR, also called sensitivity or recall) is defined by the percentage of actual positive datapoints correctly labelled as positive. The False Positive Rate (FPR) is equal to 1 - TNR with TNR the True Negative Rate (defined in similar way to TPR but with actual negative datapoints instead). Define TPR and FPR using TP, TN, FP, FN.

\[
TPR = \frac{TP}{TP + FN}, FPR = 1 - TNR = 1 - \frac{TN}{TN + FP} = \frac{FP}{TN + FP}
\]

2. (5 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.

\[
\begin{array}{c|c|c}
  c = 0 & c = 1 \\
  TPR & 1 & 0 \\
  FPR & 1 & 0 \\
\end{array}
\]

3. (5 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?

For a random classifier, on average:

\[
TPR = FPR = \frac{1}{2}
\]
4. (a) (5 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. One example is given below (Figure 6). What is the shape of the ROC for a perfect classifier?

For the perfect classifier, the TPR is 1 and the FPR is 0 for all values of $c \not\in \{0, 1\}$. ROC Curve is plotted (Figure 7).

(b) (5 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.

The goal of a classifier is to be as close as possible to the perfect classifier. In the case of the ROC, the ROC of the perfect classifier is the left corner and its AUC is 1. Therefore, the closer the ROC of a classifier is to the left corner, the higher its AUC is. Hence the AUC is a good proxy for model performance.
Figure 7: ROC curve of our classifier vs. a random classifier and the perfect classifier