CS 129
Problem Set 1

Instructors: Andrew Ng, Younes Bensouda Mourri

Due: 11:59pm Tuesday Jan 27th, 2026

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.

   YOUR SOLUTION HERE

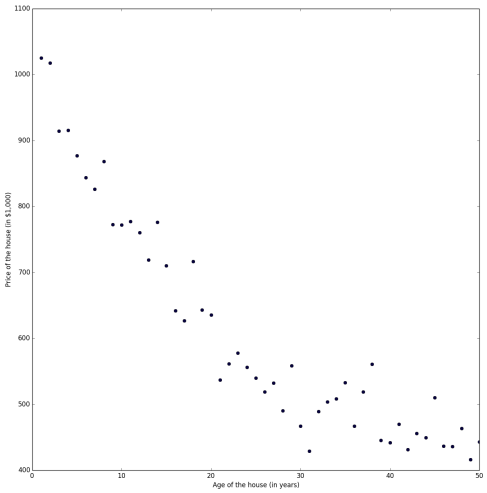
2. (5 points) We decide to fit a linear regression using gradient descent. As we have seen in lectures, the gradient descent algorithm depends on two parameters: α, the learning rate and t, the number of iterations. We tried three different sets of parameters (α, t). For each of those three sets, we 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 α/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 α 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.

   YOUR SOLUTION HERE

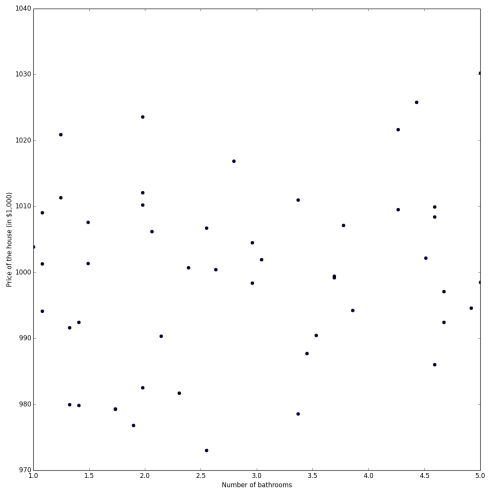
3. (5 points) We run a linear regression using only one feature: the number of rooms (Figure 3). You visit two houses: the first one has 3 rooms, the second one has 8 rooms. According to the model that is shown on Figure 3, what are the predicted prices of each house that you visited?
   YOUR SOLUTION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

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.

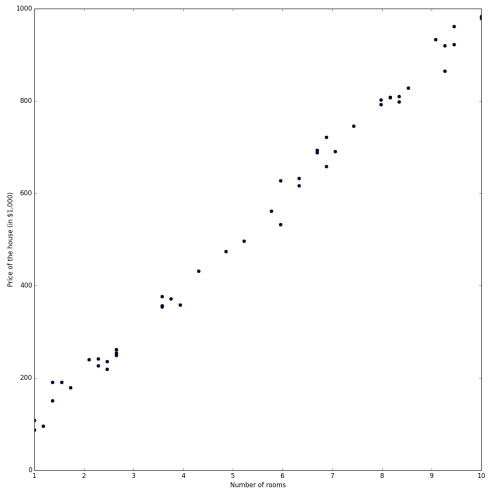
   YOUR SOLUTION HERE



(a)

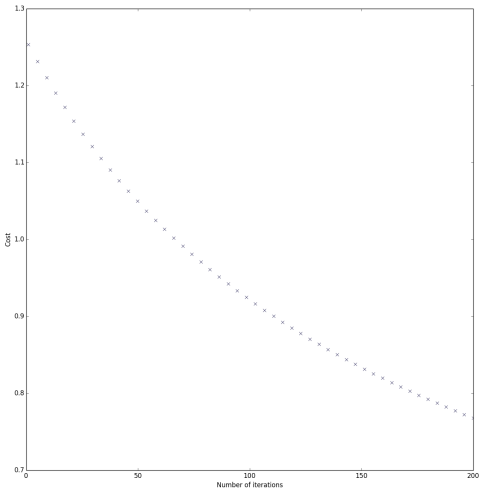


(b)

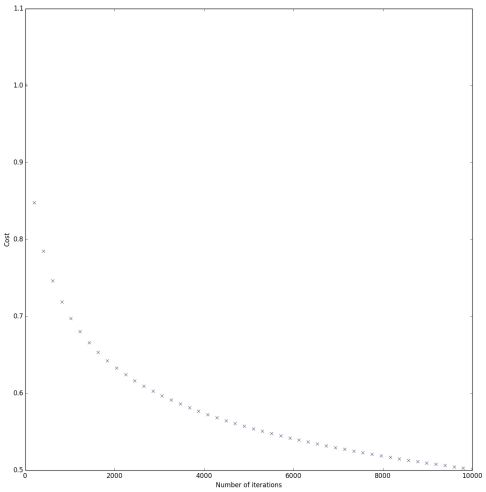


(c)

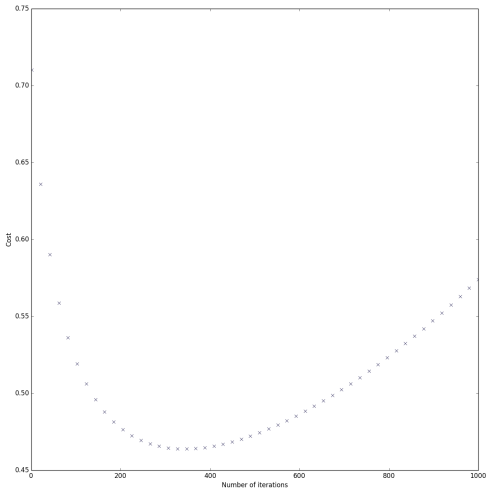
Figure 1: Plot of the price against the age (a), the number of bathrooms (b), the number of rooms (c)



(a)



(b)



(c)

Figure 2: Plot of the cost function for three different sets of parameters

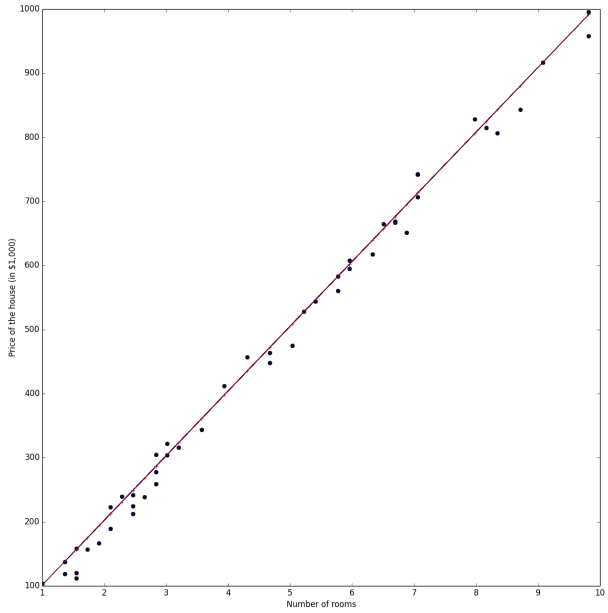


Figure 3: Price of the house against the number of rooms. The fitted line is represented in red

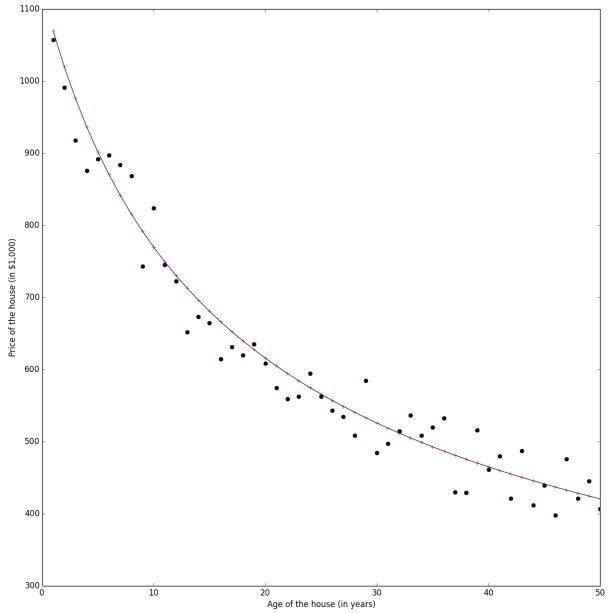


Figure 4: Price of the house against the age of the house. The fitted line is represented in red

Problem 2: Regularization

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

$$\frac{1}{2m}\sum_{i=1}^{m}(f_{\vec{w},b}(x^{(i)}) - y^{(i)})^2$$

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:

$$\frac{1}{2m}\sum_{i=1}^{m}(f_{\vec{w},b}(x^{(i)}) - y^{(i)})^2 + \frac{\lambda}{2m}\sum_{j=1}^{n} w_j^2$$

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 with respect to w_j.

   YOUR SOLUTION HERE

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

   YOUR SOLUTION HERE

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

   YOUR SOLUTION HERE

4. (2 points) How does changing the value of λ influence the magnitude of the learned parameters w?

   YOUR SOLUTION HERE

Problem 3: Normal Equation for Regularized Linear Regression

The cost for a linear regression with L_2 regularization is:

$$J(\vec{w}) = \frac{1}{2m}\sum_{i=1}^{m}\left(f_{\vec{w}}(x^{(i)}) - y^{(i)}\right)^2 + \frac{\lambda}{2m}\sum_{j=1}^{n} w_j^2$$

Note: 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.
To make the problem easier you can disreagard the bias (ie assume $b = 0$)

1. (5 points) Give a vectorized expression of $J(\vec{w})$. In other words, write $J(\vec{w})$ as a function of X, Y, $\vec{w} = (w_1, w_2, \ldots, w_n)$, λ, m without summations. Where $\vec{w}$ is a one by n matrix, X is an m by n matrix and Y is an m by one matrix. Feel free to transpose any of the matrices as you wish.

   Note: justification for vectorized expression is not required.

   YOUR SOLUTION HERE

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

   YOUR SOLUTION HERE

3. (5 points) Derive the normal equations for this cost function. In other words, find $\vec{w}$ that minimizes the cost function $J(\vec{w})$ as a function of X, Y, λ, and m. State any assumption you make in order to derive the solution.

   YOUR SOLUTION HERE

Problem 4: 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 tumor: 1 if the tumor is malignant, 0 otherwise. Therefore, the goal is to predict the type of tumor given its features.

1. (4 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.

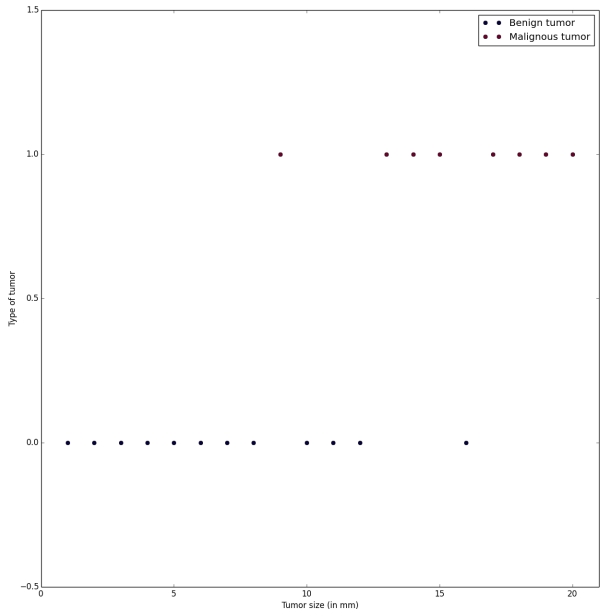


Figure 5: Type of tumor against the tumor size

YOUR SOLUTION HERE

2. (4 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:

$$f_{\vec{w},b}(\vec{x}) = \frac{1}{1 + e^{-(\vec{w}\cdot\vec{x}+b)}} = \mathbb{P}(Y = 1|\vec{x})$$

What is the expression for $\mathbb{P}(Y = 0|\vec{x})$?

YOUR SOLUTION HERE

3. (4 points) The cost function for logistic regression, used to tune the model weights over a dataset $\{(\vec{x}^{(i)}, y^{(i)})\}_{i=1}^{m}$, is defined as

$$J(W, b) = -\frac{1}{m}\sum_{i=1}^{m}\left[y^{(i)}\log\hat{p}^{(i)} + (1 - y^{(i)})\log(1 - \hat{p}^{(i)})\right],$$

where $\hat{p}^{(i)} = \mathbb{P}(Y = 1 \mid \vec{x}^{(i)})$.

Consider a single data point with $y = 1$. What happens to its contribution to the cost as $\hat{p} \to 1$? What happens as $\hat{p} \to 0$? What does this cost penalize more: confident but incorrect predictions or uncertain predictions?

YOUR SOLUTION HERE

4. (4 points) Suppose we now wish to classify tumors whose type is unknown. After computing the predicted probability, we must assign a class label to each tumor. Since there are only two possible classes, one common decision rule is to assign label 1 if $\mathbb{P}(Y = 1|\vec{x}) > \frac{1}{2}$, and label 0 otherwise. An alternative decision rule is to assign the class with the larger predicted probability. For example, if $\mathbb{P}(Y = 0|\vec{x}) > \mathbb{P}(Y = 1|\vec{x})$, we assign label 0. Show mathematically that these two decision rules are equivalent (i.e. they always lead to the same class assignment).

YOUR SOLUTION HERE

5. (4 points) What ethical problem might arise from using a fixed decision threshold of $\frac{1}{2}$ to classify tumors? How could you address this issue? Hint: Consider the effects of increasing or decreasing the threshold.

YOUR SOLUTION HERE

Problem 5: 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:

	Predicted 1	Predicted 0
Actual 1	TP	FN
Actual 0	FP	TN

- 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

1. (4 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.

   YOUR SOLUTION HERE

2. (4 points) One way to assign classes given the probability is to choose a threshold c such that if $\mathbb{P}(Y = 1|X) > c$, 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.

   YOUR SOLUTION HERE

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?

   YOUR SOLUTION HERE

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. One example is given below (Figure 6). What is the shape of the ROC for a perfect classifier? Either be very detailed in your description or draw it on the figure.

   YOUR SOLUTION HERE

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.

   YOUR SOLUTION HERE

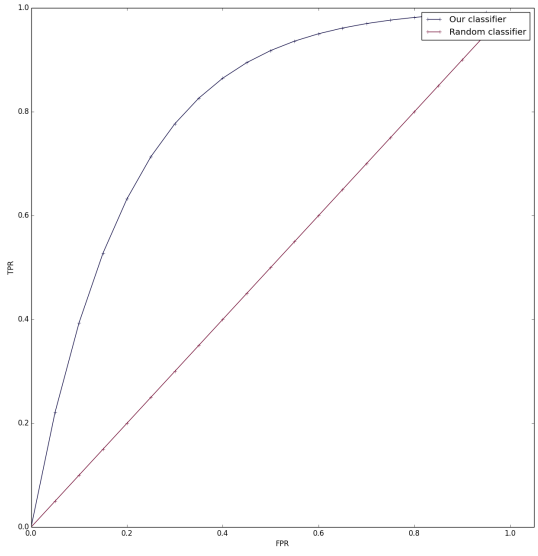


Figure 6: ROC curve of our classifier vs. a random classifier