Exercise 1. (Train vs. test datasets)

You want to predict whether or not a product fails based on historical data on the amount of the different materials used to make the product. You have a set of 1,000 data points, where each data point contains the amount of the 5 different materials ($x^i \in \mathbb{R}^d$) and the information on failure or non-failure of the product, $y^i \in \{0, 1\}$. Here 0 denotes no failure and 1 denotes failure.

1. Determine what type of learning problem you are dealing with: i.e. supervised or unsupervised learning? If supervised learning, is it a regression or classification problem.
2. You randomly split the data such that 800 data points are used as the training set and 200 are used as the test set. Why shouldn't we use all the 1,000 available data points to train the model?
3. After training your model, you observe it performs well on the training data, but poorly on the test data. What is one possible explanation? What could you try to improve the performance on the test data?

Exercise 2. (Polynomial embedding)

You are given a data set $\{(x^i, y^i)\}_{i=1}^N$ with $x^i \in \mathbb{R}^3, y^i \in \mathbb{R}$. We aim to map the independent variables x^i using an appropriate feature vector $\Phi(x) = \{\Phi_i(x)\}_{i=1}^p$, $\Phi_i : \mathbb{R}^3 \to \mathbb{R}$. Then, we will use linear regression, $y = w^T\Phi(x)$. Your job is to determine an appropriate feature map Φ based on some expert knowledge as follows.

An expert on the data and associated application believes that a polynomial Φ will give a good model. Specifically, she believes that a good prediction model can be found as a degree two polynomial, with degree in each component x_l, $l = 1, 2, 3$, no more than one.

We describe the "degree". A polynomial of a vector $x \in \mathbb{R}^3$ is a linear combination of terms $x_1^p x_2^q x_3^r$, which are called monomials, where p, q, r are nonnegative integers, called the degree of the monomial in x_1, x_2, x_3, respectively. The degree of the monomial $x_1^p x_2^q x_3^r$ is $p+q+r$. The degree of a polynomial $\Phi(x)$ is the maximum of the degrees of all of its monomials, and the degree of $\Phi(x)$ in each x_l is the maximum of the degrees of its monomials in x_l. For example, the polynomial $p(x) = x_1x_2 + x_1^3 + x_1x_2^3 + x_3^2$ has degree four and its degree in x_2 is three.

Suggest an appropriate embedding $\Phi(x)$ based on the expert's advice. What is the dimension p? In other words, how many parameters do you need to be identifying?

Exercise 3. (Constant predictors)

In this problem we will investigate a "constant" predictor. Given $\{(x^i, y^i)\}_{i=1}^N$, the constant predictor gives a constant value, regardless of the independent variable x^i. We will see how to formulate this as a linear regression and what kind of predictor we get if we use the mean-square-error (MSE) loss function.

1. Consider the feature vector 1. Formulate the loss function for the problem.
2. Show that the optimal solution is given by $w_0^* = \frac{1}{N}\sum_{i=1}^N y^i$. In other words, we will get the mean of the labels as the constant predictor.

Reflection: The above predictor gives the mean of the dependent variables, y^i's. This predictor could be useful in some applications. The mean of the data is an example of the statistical information of the data. If we use other types of loss function, we can get other statistical information of the data. For example, using the mean absolute error, $L(w_0) = \frac{1}{N}\sum_{i=1}^{N} |w_0 - y^i|$ instead of the MSE error, we get the median of the data. Optional: for more about constant predictors, you can see this topic explored in the Stanford course on machine learning.

Exercise 4. (Logistic regression)

Consider a binary classification problem with data $\{x^i, y^i\}_{i=1}^N$, $x^i \in \mathbb{R}^d$, $y^i \in \{0, 1\}$. Let our predictor be 1 if $z^i > 0$, where $z = w^T x + b$ and $z^i \in \mathbb{R}$ corresponds to using x^i above, and 0 otherwise. Suppose you are using classification for a fault diagnosis scenario, and $x^i \in \mathbb{R}^d$ is some attributes of the process, whereas the two classes $y^i = 0$ and $y^i = 1$ correspond to *no fault* and *fault*, respectively. Suppose we are more disturbed by a *false negative* prediction than a false positive because we want to ensure we do not miss any faults. As such we slightly modify the loss function for training by introducing a constant $c \in \mathbb{R}$:

$$L(w, b) = \frac{1}{N}\sum_{i=1}^{N} c \times y^i \log(1 + e^{-z_i}) + (1 - y^i)\log(1 + e^{z_i}), \tag{0.1}$$

1. Should we choose $c > 1$ or $c < 1$? Justify your answer.
2. Derive the gradient of the loss and write the gradient descent procedure to find the parameters (b, w) of the logistic regression.
3. Explain an approach to verify the convexity of $L(w, b)$.

Exercise 5 (Logistic function)

In this exercise, we will analyze the behaviour of the logistic function. To this end, we consider a binary classification problem with one-dimensional input data $x \in \mathbb{R}$ and class prediction $P(Y = 1 \mid x) = \sigma(wx + b)$, where $\sigma(z) = \frac{1}{1+e^{-z}}$ is the logistic or sigmoid function.

1. (Plotting) When given a plot for $y = f(x)$, one can obtain the plot for $y = f(wx + b)$ by shifting the origin to $x = \frac{-b}{w}$ and compressing the axis by a factor of $|w|$. In case $w < 0$, we flip the plot around $x = -\frac{b}{w}$. Verify this by plotting $\sigma(wx + b)$ for $w = -2$ and $b = 4$.
2. For the above choice of w and b, what happens with $P(Y = 1 \mid x)$ as x increases? Where do we have $P(Y = 1 \mid x) = 0.5$?
3. Fix $w = 1$ and consider a data point $x = 1$. How does changing the value of b from $b = 0$ to $b = -2$ affect our prediction of $P(Y = 1 \mid x)$?
4. Now fix $b = 0$. Plot (qualitatively) $P(Y = 1 \mid x) = \sigma(wx)$ for $w \in \{0.5, 1, 2\}$. What happens with $P(Y = 1 \mid x)$ as $w \to 0$ and $w \to \infty$?
5. **Bonus:** Consider the training data set

   $$\mathcal{D} = \{x_i, y_i\}_{i=1}^8 = \{(20, 1), (5, 0), (25, 1), (10, 0), (7.5, 1), (22.5, 0), (17.5, 1), (12.5, 0)\}.$$

   Plot the training data on the real-line using circles, '$\circ$', for $y = 0$, and crosses, '$\times$', for $y = 1$. Are there w and b such that for all i we have $y_i = 1$ if and only if $wx_i + b > 0$? What can you say about the benefits of using a probabilistic prediction model for the above classification problem?

Bonus Exercise (Logistic loss using Maximum Likelihood Estimation)
In this exercise, we will derive the logistic loss function using *Maximum Likelihood Estimation*. Maximum Likelihood Estimation is a statistical method used to estimate the unknown parameters of a probability distribution based on observed data. This exercise serves as a complement to help understand the logistic loss from a different perspective.

1. Consider a sequence of N independent coin tosses, with the outcome for the i^{th} toss denoted by Y_i, where $Y_i = 1$ corresponds to the outcome *heads* and $Y_i = 0$ to *tails*. Suppose the probability of *heads* is given by some parameter $p \in [0, 1]$, i.e. $P(Y^i = 1) = p$.

   (a) Verify that the probability for outcome $Y_i = y_i \in \{0, 1\}$ can be written as

   $$P(Y_i = y_i) = p^{y_i}(1-p)^{1-y_i}, \quad y_i \in \{0, 1\}.$$

   Hint: Evaluate for $P(Y_i = 1)$ and $P(Y_i = 0)$.

   (b) Verify that the joint probability distribution of $Y := (Y_1, \ldots, Y_N)$ can be written as

   $$P(\boldsymbol{Y} = \boldsymbol{y}) = \prod_{i=1}^{N} P(Y_i = y_i) = \prod_{i=1}^{N} p^{y_i}(1-p)^{1-y_i}.$$

   The above expression is called the Likelihood function for the sequence of outcomes. It is often denoted as $\mathcal{L}(p \mid \boldsymbol{y}) = P(\boldsymbol{Y} = \boldsymbol{y})$.

   (c) Suppose that the true parameter p is unknown. You only have access to the observations y_i. The idea of maximum likelihood estimation is to choose the value of $\hat{p}$ which maximizes the probability of the observations $\boldsymbol{Y} = \boldsymbol{y}$.

      i. Argue that the value of $\hat{p}$ which maximizes the likelihood $\mathcal{L}(p \mid \boldsymbol{y}) = P(\boldsymbol{Y} = \boldsymbol{y})$ is also minimizing the negative log-likelihood, i.e. show that

      $$\text{argmax}_p \mathcal{L}(p \mid \boldsymbol{y}) = \text{argmin}_p - \log \mathcal{L}(p \mid \boldsymbol{y}).$$

      Hint: Use the monotonicity of the $\log(\cdot)$ function.

      ii. Show that the negative log-likelihood can be expressed as

      $$-\log \mathcal{L}(p \mid \boldsymbol{y}) = \sum_{i=1}^{N} -(y_i \log(p) + (1-y_i)\log(1-p)).$$

2. Now we move towards formulating the logistic regression problem in terms of minimizing log-likelihood of the dataset. We will do so using the following example. Suppose that we want to estimate the relationship between the amount of study hours x a student spends for a course and his/her passing outcome for the course, denoted by $Y \in \{0, 1\}$. We propose that the probability of passing $P(Y = 1 \mid x)$ is modelled as $P(Y = 1 \mid x) = \sigma(wx + b)$, where $\sigma(z) = \frac{1}{1+e^{-z}}$ and $w \in \mathbb{R}$.

   (a) Define $\hat{p}(z) := (\frac{1}{1+e^{-z}}, \frac{e^{-z}}{1+e^{-z}})$ for any $z \in \mathbb{R}$. Show that $\hat{p}$ is a probability distribution on $\{1, 0\}$ for every $z \in \mathbb{R}$. and also $\hat{p}(1) + \hat{p}(0) = 1 \ \forall z$. Hence $\hat{p}(z)$ is a probability distribution $\forall z \in \mathbb{R}$.

   (b) Now suppose that we are given a dataset $\{x_i, y_i\}_{i=1}^N$ for a class of N students where x_i is the amount of hours student i has studied and $Y_i \in \{0, 1\}$ is his/her outcome for the course. Assume that the relationship between passing probability and the amount of study hours for student i is given by $P(Y_i = 1) = \sigma(wx_i + b)$ for all students, with w, b being constant for all students.

i. Using the concepts you learned in the first part, show that the probability of outcome y_i for a student i in terms of his/her passing probability $p_i = P(Y_i = 1)$ is given by

$$P(Y_i = y_i) = p_i^{y_i}(1 - p_i)^{1-y_i}.$$

ii. Write the above expression for $P(Y_i = 1 \mid x_i) = \sigma(wx_i + b)$.

iii. Now, verify that the joint probability for outcomes $Y_1 = y_1, \cdots, Y_N = y_N$ when given the observations $\{x_i\}_{i=1}^N$ can be written as

$$P(\boldsymbol{Y} = \boldsymbol{y} \mid \boldsymbol{x}) = \prod_{i=1}^{N} (\sigma(wx_i + b))^{y_i}(1 - \sigma(wx_i + b))^{1-y_i}.$$

iv. Suppose you do not know the parameters w, b and you would like to estimate them from the dataset. You decide to use maximum likelihood estimation to do this. Show that the expression for negative log-likelihood can be written as

$$-\log \mathcal{L}(w, b \mid \boldsymbol{y}, \boldsymbol{x}) = \sum_{i=1}^{N} \left\{ y_i \log \left(\frac{1}{\sigma(wx_i + b)} \right) + (1 - y_i) \log \left(\frac{1}{1 - \sigma(wx_i + b)} \right) \right\}$$

What happens when you minimize the negative log-likelihood over the parameters w, b? How does the above compare to the logistic loss and logistic regression that you have seen in class?