Problem 1. (Linear regression) 3 points
You are given a dataset containing 150 data points. Each point represents weight (tonnes), motor power (horsepower), torque (Newtonmeter) of a car, along with its CO2 emission (kg/km). A linear regression was used to predict CO2 emissions as a function of the features.

1. Specify the dimensions of data set X, and regression parameters w, b.

   Dimensions: $X \in \ldots\ldots\ldots\ldots\ldots$, $w \in \ldots\ldots\ldots\ldots\ldots$, $b \in \ldots\ldots\ldots\ldots\ldots$

   What is the prediction for $x^{test} \in \mathbb{R}^3$? $y^{test} = \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$

   Solution: Let $X \in \mathbb{R}^{150\times 3}$ be the feature matrix with features car weight, cylinder capacity, motor power, and torque, let $w \in \mathbb{R}^3$ be the weights vector, and let $b \in \mathbb{R}$ be the bias or intercept. The linear regression model is written as follows:

   $$y^{test} = w^T x^{test} + b.$$

2. After performing linear regression, you notice that your trained model suffers from overfitting. Which method below will most likely **not** help to avoid overfitting? Circle the correct answer.

   (a) Add regularization in the loss function. (b) Remove a feature in the linear regression.
   (c) Conduct a polynomial feature expansion.

   Solution: Conducting a polynomial feature expansion will not help to avoid overfitting. On the contrary, it adds flexibility to the model.

Problem 2. (Logistic regression) 3 points

1. Consider a binary classification with $x \in \mathbb{R}^2$. We aim to use logistic regression to learn a classifier. Suppose all $x \in \mathbb{R}^2$ such that $z = w^T x + b > 0$ will be labeled as class 1, whereas $\sigma(z)$ gives probability of belonging to class 1. For $x^i = [-\frac{1}{2}, \frac{9}{2}]^T$, $w = [2, -1]^T$, $b = 4$ write $\sigma(z^i)$. You may leave your answer in terms of the exponential function. *Solution:*

   $$z = w^T x + b = -\frac{3}{2}$$
   $$\sigma(z) = \frac{1}{1 + e^{1.5}}$$

2. Suppose our training dataset consists of N data points. Recall that the logistic loss is written as $L(w, b) = -\frac{1}{N}\sum_{i=1}^{N}\left(y^i \log(\hat{y}^i) + (1 - y^i)\log(1 - \hat{y}^i)\right)$, where $\hat{y}^i = \sigma(w^T x^i + b)$. Write the gradient descent updates with $w_t = [2, -1]^T$. You don't need to calculate the gradient.

   $$w_{t+1} = \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

   Solution:

   $$w_{t+1} = w_t - \eta \nabla_w L(w_t, b_t) = [2, -1]^T - \eta \nabla_w L([2, -1]^T, b_t)$$

3. Suppose the true label of x^i is $y^i = 0$ and $\sigma(w^T x^i + b) = 0.1824$. Circle the term contributing to the logistic loss corresponding **only** to this point:

   (a) $-\frac{1}{N}\log\big(\sigma(w^T x^i + b)\big)$; (b) $-\frac{1}{N}\log\big(\sigma(-w^T x^i - b)\big)$.

   Solutions:

$$L_i(w,b) = -\frac{1}{N}\left(y^i \log(\hat{y}^i) + (1-y^i)\log(1-\hat{y}^i)\right).$$

   For $y^i = 0$, we have

$$\begin{aligned} L_i(w,b) &= -\frac{1}{N}(1-y^i)\log(1-\hat{y}^i)). \\ &= -\frac{1}{N}\log(1-\sigma(w^T x^i + b)) \\ &= -\frac{1}{N}\log(\sigma(-w^T x^i - b)) \end{aligned}$$

   Here, we have used $1 - \sigma(z) = \sigma(-z)$

Problem 3. (Polynomial embedding and cross-validation) 4 points

We are given a data set $\{(x^n, y^n)\}_{n=1}^{50}$ with $x^n \in \mathbb{R}^2, y^n \in \mathbb{R}$. We aim to map the independent variables x^n using an appropriate feature vector $\Phi(x) = \{\Phi_i(x)\}_{i=1}^{p}$, where $\Phi_i : \mathbb{R}^2 \to \mathbb{R}$.

1. Construct the feature vector $\Phi(x)$ as a polynomial of degree 2. Write all the features $\{\Phi_i(x)\}_{i=1}^{p}$. Hint: there should be 6 features.

   Solution: The number of features is $p = 6$, and they are:

$$\Phi_1(x) = 1,\ \Phi_2(x) = x_1,\ \Phi_3(x) = x_2,\ \Phi_4(x) = x_1^2,\ \Phi_5(x) = x_2^2,\ \Phi_6(x) = x_1 x_2.$$

2. The explicit expression of the predictor is $w^T\Phi(x) = \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$

   Solution: The predictor $w^T\Phi(x)$ using the features $\{\Phi_i(x)\}_{i=1}^{p}$ from the previous part can be written as:

$$y = \sum_{i=1}^{6} w_i \Phi_i(x).$$

We performed feature selection with 40 training points, and 10 test points. We trained the model on different subsets of the 6 features above and evaluated the performance on the test set in each case. A particular subset yielded the best accuracy on our 10 test points, but the model performs poorly on a new set of test points. To address the issue, we apply 4-fold cross-validation to the 40 training points, which yields 4 validation errors: $\{e_i\}_{i=1}^{4}$.

3. What is the best prediction of the error on unseen data?

   (a) $\frac{1}{4}\sum_{i=1}^{4} e_i$; (b) $\max_{i\in\{1,2,3,4\}} e_i$; (c) $\min_{i\in\{1,2,3,4\}} e_i$.

   Solution: The average error over the four folds provides the best estimate of the test error. It accounts for the variability across different splits of the data and gives a more reliable measure of model performance.