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

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

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.

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.

2. Suppose our training dataset consists of N data points. Recall that the logistic loss is written as $L(w, b) = -\frac{1}{N}\sum_{i=1}^{N}\big(y^i \log(\hat{y}^i) + (1 - y^i)\log(1 - \hat{y}^i)\big)$, 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.

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

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

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.

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

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) $\min_{i\in\{1,2,3,4\}} e_i$.