THE HONG KONG UNIVERSITY OF SCIENCE & TECHNOLOGY
Machine Learning
Homework 1

Due Date: See course webpage.

Your answers should be typed, not handwritten. You can submit a Word file or a pdf file. Submissions are to be made via Canvas. Note that penalty applies if your similarity score exceeds 40. To minimize your similarity score, don't copy the questions.

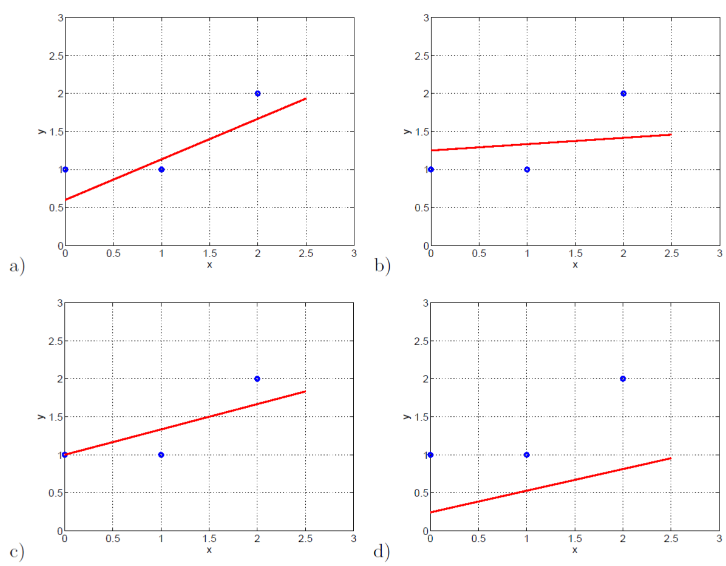
Question 1: Suppose a dataset $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^N$ is generated from some unknown distribution $p(\mathbf{x})$ and we learn from $\mathcal{D}$ a distribution $q_\theta(\mathbf{x})$ with parameters θ. What is the KL divergence $KL(p||q_\theta)$ of q_θ from p? What is the cross entropy $H(p, q_\theta)$ between p and q_θ? How are they related?

What is the log-likelihood of $l(\theta|\mathcal{D})$? How is maximizing $l(\theta|\mathcal{D})$ related to minimizing the cross entropy and the KL divergence ?

Question 2 Consider carrying out linear regression on the following dataset. Manually compute the ordinary least squares solution.

x_1	0	0	1	1	1
x_2	1	1	1	0	0
y	0	1	2	3	4

Question 3 The following figures show linear regression results on a dataset of only three data points (marked blue).



The results were obtained using following regularization schemes:

1. $\frac{1}{3}\sum_{i=1}^{3}(y_i - w_0 - w_1 x_i)^2 + \lambda w_1^2$ where $\lambda = 1$.
2. $\frac{1}{3}\sum_{i=1}^{3}(y_i - w_0 - w_1 x_i)^2 + \lambda w_1^2$ where $\lambda = 10$.
3. $\frac{1}{3}\sum_{i=1}^{3}(y_i - w_0 - w_1 x_i)^2 + \lambda (w_0^2 + w_1^2)$ where $\lambda = 1$.

4. $\frac{1}{3}\sum_{i=1}^{3}(y_i - w_0 - w_1 x_i)^2 + \lambda(w_0^2 + w_1^2)$ where $\lambda = 10$.

Match the regularization schemes with the regress results. Briefly explain your answers.

Question 4 Consider applying logistic regression to the following dataset:

x_1	0	0	1	1
x_2	0	1	0	1
y	0	0	0	1

The target is to learn a model of the form $p(y = 1|\mathbf{x}, \mathbf{w}) = \sigma(w_0 + w_1 x_1 + w_2 x_2)$.

Suppose $w_0 = -2$, $w_1 = 1$ and $w_2 = 1$ initially and $\alpha = 0.1$. Manually run the batch gradient descent algorithm for one iteration. Give the weights and training error (i.e., fraction of misclassified examples) after the iteration.

Question 5 Consider applying logistic regression to the following dataset:

x_1	0	0	1	1
x_2	0	1	0	1
y	1	0	0	1

1. If we use raw feature x_1 and x_2, the model is
$$p(y = 1|\mathbf{x}, \mathbf{w}) = \sigma(w_0 + w_1 x_1 + w_2 x_2).$$
What is the minimum achievable training error in this case? Give weights that achieve the minimum error.

2. Next consider using an additional feature $x_1 x_2$ in addition to the raw feature x_1 and x_2. The model now is
$$p(y = 1|\mathbf{x}, \mathbf{w}) = \sigma(w_0 + w_1 x 1 + w_2 x_2 + w_3 x_1 x_2).$$
What is the minimum achievable training error in this case? Give weights that achieve the minimum error.

Question 6 Consider the gradient vector in logistic regression $\nabla J(\mathbf{w}) = (\frac{\partial J(\mathbf{w})}{\partial w_0}, \frac{\partial J(\mathbf{w})}{\partial w_1}, \ldots, \frac{\partial J(\mathbf{w})}{\partial w_D})$ where

$$\frac{\partial J(\mathbf{w})}{\partial w_j} = -\frac{1}{N}\sum_{i=1}^{N}[y_i - \sigma(z_i)]x_{i,j}.$$

Suppose the feature x_1 is binary and, in the training set, it takes value 1 only in a small number of training examples with class label 1 (i.e., $y = 1$), and it takes value 0 in all training examples with class label 0 (i.e., $y = 0$). What will happen to the weight w_1 if we update it repeatedly using the following rule:

$$w_1 \leftarrow w_1 + \alpha\frac{1}{N}\sum_{i=1}^{N}[y_i - \sigma(\mathbf{w}^\top \mathbf{x}_i)]x_{i,1}$$

What if we use the following update rule instead:

$$w_1 \leftarrow w_1 + \alpha[-\lambda w_1 + \frac{1}{N}\sum_{i=1}^{N}[y_i - \sigma(\mathbf{w}^\top \mathbf{x}_i)]x_{i,1}],$$

where λ is the regularization constant?

Self-Practice Questions: I will minimize the amount math derivations in class for the sake of the majority of the students. Those interested in understanding all the math can try to solve the following problems. This is for self-practice. **Do not include the answers in your submission.**

- Prove Theorem 1.1 in Lecture 01-2.
- Let $I(X;Y) = H(X) - H(X|Y)$. Prove that

$$I(X;Y) = KL(P(X,Y)||P(X)P(Y)).$$

- Derive the OLS solution for linear regression.
- Derive the gradient descent update rule for Softmax.