THE HONG KONG UNIVERSITY OF SCIENCE & TECHNOLOGY

Machine Learning

Homework 1 Solutions

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 ?

Solution:

$$
\begin{aligned}
KL(p||q_\theta) &= E_p[\log p(\mathbf{x})] - E_p[\log q_\theta(\mathbf{x})] \\
H(p, q_\theta) &= -E_p[\log q_\theta(\mathbf{x})] \\
KL(p||q_\theta) &= H(p, q_\theta) - H(p), \text{ where } H(p) = -E_p[\log p(\mathbf{x})] \text{ is the entropy of } p. \\
l(\theta|\mathcal{D}) &= \sum_{i=1}^N \log q_\theta(\mathbf{x}_i).
\end{aligned}
$$

$l(\theta|\mathcal{D})$ can be viewed as an approximation of $-NH(p, q_\theta)$. Hence, maximizing the log-likelihood $l(\theta|\mathcal{D})$ amounts to minimizing the cross entropy $H(p, q_\theta)$. It also amounts to minimizing the KL divergence $KL(p||q_\theta)$ as the entropy term $H(p)$ in the third equation above does not depend on θ.

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

Solution: The design matrix and the label vector $\mathbf{y}$ are:

$$
\begin{aligned}
\mathbf{X}^\top &= \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 \end{bmatrix} \\
\mathbf{y}^\top &= \begin{bmatrix} 0 & 1 & 2 & 3 & 4 \end{bmatrix}
\end{aligned}
$$

We have

$$
\begin{aligned}
\mathbf{X}^\top\mathbf{X} &= \begin{bmatrix} 5 & 3 & 3 \\ 3 & 3 & 1 \\ 3 & 1 & 3 \end{bmatrix} \\
(\mathbf{X}^\top\mathbf{X})^{-1} &= \begin{bmatrix} 2 & -1.5 & -1.5 \\ -1.5 & 1.5 & 1 \\ -1.5 & 1 & 1.5 \end{bmatrix} \\
(\mathbf{X}^\top\mathbf{y})^\top &= \begin{bmatrix} 10 & 3 & 9 \end{bmatrix}
\end{aligned}
$$

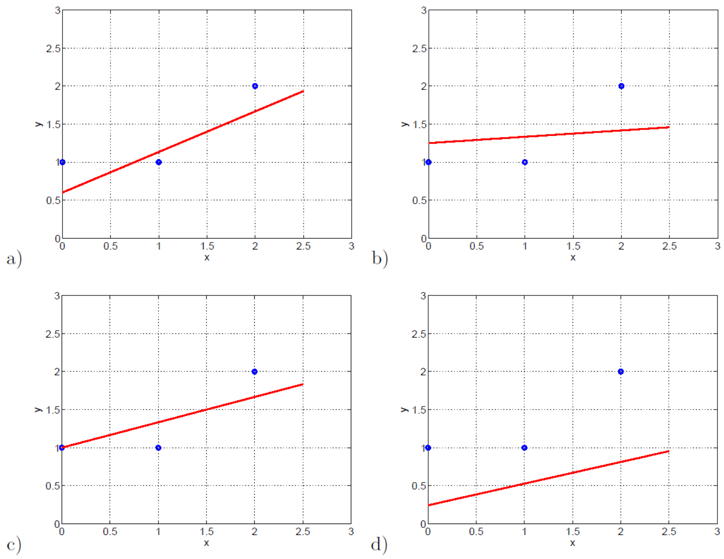
Therefore,

$$\begin{bmatrix} w_0 \\ w_1 \\ w_2 \end{bmatrix} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} = \begin{bmatrix} 2 \\ 1.5 \\ \text{-}1.5 \end{bmatrix}$$

The final regression equation is:

$$y = 2 + 1.5x_1 - 1.5x_2$$

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.

Solution: The first two objective functions regularize only w_1. The results are shown in c) and b) respectively. The line in b) has a flat slope because a large regularization constant (10) is used.

The last two objective functions regularize both w_0 and w_1. The results are shown in a) and d) respectively. The intercepts are lower than in the other two cases.

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.

Solution: $\mathbf{w}^\top\mathbf{x}_1 = -2$, $\sigma(\mathbf{w}^\top\mathbf{x}_1) = 0.12$; $\mathbf{w}^\top\mathbf{x}_2 = -1$, $\sigma(\mathbf{w}^\top\mathbf{x}_2) = 0.27$; $\mathbf{w}^\top\mathbf{x}_3 = -1$, $\sigma(\mathbf{w}^\top\mathbf{x}_3) = 0.27$; $\mathbf{w}^\top\mathbf{x}_4 = 0$, $\sigma(\mathbf{w}^\top\mathbf{x}_4) = 0.5$.

$$
\begin{aligned}
w_0 &= -2 + 0.1 \times \frac{1}{4} \times ([0-0.12]\times 1 + [0-0.27]\times 1 + [0-0.27]\times 1 + [1-0.5]\times 1) = -2.004 \\
w_1 &= 1 + 0.1 \times \frac{1}{4} \times ([0-0.12]\times 0 + [0-0.27]\times 0 + [0-0.27]\times 1 + [1-0.5]\times 1) = 1.00575 \\
w_2 &= 1 + 0.1 \times \frac{1}{4} \times ([0-0.12]\times 0 + [0-0.27]\times 1 + [0-0.27]\times 0 + [1-0.5]\times 1) = 1.00575
\end{aligned}
$$

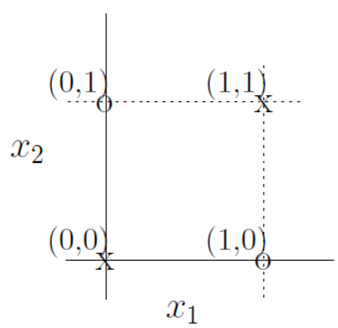
With the new parameters, we have $\mathbf{w}^\top\mathbf{x}_1 = -2.004 < 0$, and hence $\mathbf{x}_1$ is classified into class 0; $\mathbf{w}^\top\mathbf{x}_2 = -2.004 + 1.00575 < 0$, and hence $\mathbf{x}_2$ is classified into class 0; $\mathbf{w}^\top\mathbf{x}_3 = -2.004 + 1.00575 < 0$, and hence $\mathbf{x}_3$ is classified into class 0; $\mathbf{w}^\top\mathbf{x}_4 = -2.004 + 1.00575 + 1.00575 > 0$, and hence $\mathbf{x}_4$ is classified into class 1. The training error is 0.

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
$$p(y = 1|\mathbf{x}, \mathbf{w}) = \sigma(w_0 + w_1 x1 + w_2 x_2 + w_3 x_1 x_2).$$
What is the minimum achievable training error in this case? Give weights that achieve the minimum error.

Solution:



1. As shown above, the dataset is not linearly separable. The minimum achievable error using a linear classifier is 0.25. It is achieved by, for instance, the weights $w_0 = 0.5$, $w_1 = -1$ and $w_2 = -1$. In this case, the first three examples are classified correctly and the last example is classified incorrectly.
2. With the additional feature $x_1 x_2$, we can correctly classify all four examples using weights $w_0 = 0.5$, $w_1 = -1$, $w_2 = -1$ and $w_3 = 2$.

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?

Solution: Since $\sigma(\mathbf{w}^\top \mathbf{x}_i) < 1$, $\frac{1}{N} \sum_{i=1}^{N} [y_i - \sigma(\mathbf{w}^\top \mathbf{x}_i)] x_{i,1}$ is always positive. If we use the first (unregularized) update rule, the weight w_1 might increase without bound, leading to numerical instability.

If we use the second (regularized) update rule, w_1 will stop increasing when $\lambda w_1 \geq \frac{1}{N} \sum_{i=1}^{N} [y_i - \sigma(\mathbf{w}^\top \mathbf{x}_i)] x_{i,1}$. So, regularization makes logistic regression numerically stable with regard to the scenario described in this problem.