MSBD 5012 Machine Learning **(Due: 1st October)**

Written Assignment #1 Solution

Instructor: Nevin L. Zhang *Name:* studentname, *ID:* studentID

Problem 1: Likelihood, KL Divergence, Cross Entropy

Answer:

$$KL(p||q_\theta) = \sum_{\mathbf{x}_i} p(\mathbf{x}_i) \log \frac{p(\mathbf{x}_i)}{q_\theta(\mathbf{x}_i)}$$

$$H(p, q_\theta) = \sum_{\mathbf{x}_i} p(\mathbf{x}_i) \log \frac{1}{q_\theta(\mathbf{x}_i)}$$

$$KL(p||q_\theta) = -H(p) + H(p, q_\theta)$$

$$\text{When } p \text{ is fixed, } \min KL(p||q_\theta) \Leftrightarrow \min H(p, q_\theta)$$

$$\log L(\theta|\mathcal{D}) = \log p(\mathcal{D}|\theta) = \log \left(\prod_{i=1}^{N} q_\theta(\mathbf{x}_i) \right) = \sum_{i=1}^{N} \log q_\theta(\mathbf{x}_i)$$

$$\min KL(p||q_\theta) \Leftrightarrow \min H(p, q_\theta) \Leftrightarrow \min - \sum_{\mathbf{x}_i} p(\mathbf{x}_i) \log q_\theta(\mathbf{x}_i) \Leftrightarrow \max \sum_{\mathbf{x}_i} p(\mathbf{x}_i) \log q_\theta(\mathbf{x}_i)$$

$$\text{When } p \text{ is fixed, } \min KL(p||q_\theta) \Leftrightarrow \min H(p, q_\theta) \Leftrightarrow \max \sum_{\mathbf{x}_i} \log q_\theta(\mathbf{x}_i) \Leftrightarrow \max \log \text{likelihood}$$

Problem 2: Least Square Solution

Answer:

$$\text{Let } \mathbf{X} = (\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2), \text{ where } \mathbf{x}_0 = (1,1,1,1,1)^T,\ \mathbf{x}_1 = (0,0,1,1,1)^T,\ \mathbf{x}_2 = (1,1,1,0,0)^T$$

$$\text{Let } \mathbf{y} = (0,1,2,3,4)^T$$

$$\text{The linear equation is: } \mathbf{y} = \boldsymbol{w}^T \mathbf{X}$$

$$\text{The loss function is: } J(\boldsymbol{w}) = \frac{1}{5}(\mathbf{y} - \mathbf{X}\boldsymbol{w})^T(\mathbf{y} - \mathbf{X}\boldsymbol{w})$$

$$\nabla J(\boldsymbol{w}) = 0 \Rightarrow \boldsymbol{w} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

$$(\mathbf{X}^T\mathbf{X})^{-1} = \begin{pmatrix} 2 & -1.5 & -1.5 \\ -1.5 & 1.5 & 1 \\ -1.5 & 1 & 1.5 \end{pmatrix}$$

$$\boldsymbol{w} = (2, 1.5, -1.5)^T$$

Problem 3: Match the Regularizations

Answer:

Regularization can fine a certain item in the formula. After adding a regularization on a certain item its value will decrease accordingly. If we add a larger regularization, the corresponding value will decrease more. Thus we can match the results according to what I mention above.
Formula 3. and 4. add the regularization on w_0, which related to the y-intercept, so the result should have a small y-intercept, and formula 4. should have a smaller y-intercept with a larger regularization.

Thus formula 3. matches picture a. and formula 4. matches picture d.

Formula 1. and 2. add the regularization on w_1, which related to the slope, so the result should have a small slope, and formula 2. should have a smaller slope with a larger regularization.

Thus formula 1. matches picture c. and formula 2. matches picture b.

Problem 4: Logistic Regression – Batch GD

Answer:

$$\text{Let } \mathbf{X} = (\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2), \text{ where } \mathbf{x}_0 = (1,1,1,1)^T,\ \mathbf{x}_1 = (0,0,1,1)^T,\ \mathbf{x}_2 = (0,1,0,1)^T$$

$$\text{Let } \mathbf{y} = (0,0,0,1)^T, \boldsymbol{w} = (w_0, w_1, w_2)^T$$

The loss function for logistic regression is: (i represents the i-th sample)

$$J(\boldsymbol{w}) = -\frac{1}{N}\sum_{i=1}^{N}\left(y_i \log \sigma(\boldsymbol{w}^T\boldsymbol{x}_i) + (1-y_i)\log(1-\sigma(\boldsymbol{w}^T\boldsymbol{x}_i)\right)$$

Compute the gradient: (j represents the j-th row)

$$\frac{\partial J(\boldsymbol{w})}{\partial w_j} = -\frac{1}{N}\sum_{i=1}^{N}\left(y_i - \sigma(\boldsymbol{w}^T\boldsymbol{x}_i)\right)x_{ij}$$

According to Gradient Descent,

$$w_j = w_j + \alpha\frac{1}{N}\sum_{i=1}^{N}\left(y_i - \sigma(\boldsymbol{w}^T\boldsymbol{x}_i)\right)x_{ij}$$

Thus,

$$w_0 = -2 + 0.1 \times \frac{1}{4} \times [(0-\sigma(-2)) \times 1 + (0-\sigma(-1)) \times 1 + (0-\sigma(-1)) \times 1 + (1-\sigma(0)) \times 1]$$

$$= -2 + 0.025 \times (1 - \sigma(0) - 2\sigma(-1) - \sigma(-2)) = -2.004$$

$$w_1 = 1 + 0.1 \times \frac{1}{4} \times [(0-\sigma(-1)) \times 1 + (1-\sigma(0)) \times 1] = 1 + 0.025 \times (1 - \sigma(0) - \sigma(-1)) = 1.006$$

$$w_2 = 1 + 0.1 \times \frac{1}{4} \times [(0-\sigma(-1)) \times 1 + (1-\sigma(0)) \times 1] = 1 + 0.025 \times (1 - \sigma(0) - \sigma(-1)) = 1.006$$

Sample labels:

$$P(y_1 = 1|\mathbf{X}, \boldsymbol{w}) = \sigma(-2.004) = 0.119, P(y_1 = 0|\mathbf{X}, \boldsymbol{w}) = 1 - P(y_1 = 1|\mathbf{X}, \boldsymbol{w}) = 1 - 0.119 = 0.881$$

$$P(y_1 = 1|\mathbf{X}, \boldsymbol{w}) < P(y_1 = 0|\mathbf{X}, \boldsymbol{w})$$

$$y_1 \text{ is classified class 0.}$$

$$P(y_2 = 1|\mathbf{X}, \boldsymbol{w}) = \sigma(-2.004 + 1.006) = 0.269, P(y_2 = 0|\mathbf{X}, \boldsymbol{w}) = 1 - P(y_2 = 1|\mathbf{X}, \boldsymbol{w}) = 1 - 0.269 = 0.731$$

$$P(y_2 = 1|\mathbf{X}, \boldsymbol{w}) < P(y_2 = 0|\mathbf{X}, \boldsymbol{w})$$

$$y_2 \text{ is classified class 0.}$$

$$P(y_3 = 1|\mathbf{X}, \boldsymbol{w}) = \sigma(-2.004 + 1.006) = 0.269, P(y_3 = 0|\mathbf{X}, \boldsymbol{w}) = 1 - P(y_3 = 1|\mathbf{X}, \boldsymbol{w}) = 1 - 0.269 = 0.731$$

$$P(y_3 = 1|\mathbf{X}, \boldsymbol{w}) < P(y_3 = 0|\mathbf{X}, \boldsymbol{w})$$

$$y_3 \text{ is classified class 0.}$$

$$P(y_4 = 1|\mathbf{X}, \boldsymbol{w}) = \sigma(-2.004 + 1.006 + 1.006) = 0.502, P(y_4 = 0|\mathbf{X}, \boldsymbol{w}) = 1 - P(y_4 = 1|\mathbf{X}, \boldsymbol{w}) = 1 - 0.502 = 0.498$$

$$P(y_4 = 1|\mathbf{X}, \boldsymbol{w}) > P(y_4 = 0|\mathbf{X}, \boldsymbol{w})$$

$$y_4 \text{ is classified class 1.}$$

Training error:

$$\sum \mathbf{1}(y_i \neq y) = 0$$

Problem 5: Logistic Regression – Minimum Error

Answer:

(1).
For logistic regression, the decision rule for performing classification is:

$$\hat{y} = 1 \Leftrightarrow P(y = 1|\mathbf{x}, \boldsymbol{w}) > P(y = 0|\mathbf{x}, \boldsymbol{w}) \Leftrightarrow \boldsymbol{w}^T\mathbf{x} > 0,\ \hat{y} = 1 \Leftrightarrow \boldsymbol{w}^T\mathbf{x} < 0$$

Training error is:

$$\mathbf{1}(\hat{y} \neq y_i)$$

$$\text{Let } \mathbf{X} = (\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2),\ \text{where } \mathbf{x}_0 = (1,1,1,1)^T,\ \mathbf{x}_1 = (0,0,1,1)^T,\ \mathbf{x}_2 = (0,1,0,1)^T$$

$$\text{Let } \mathbf{y} = (1,0,0,1)^T, \boldsymbol{w} = (w_0, w_1, w_2)^T$$

If we want a perfect model, we should have: (i represents the i-th sample)

$$\boldsymbol{w}^T\boldsymbol{x}_i > 0,\ i = 1, 4$$

$$\boldsymbol{w}^T\boldsymbol{x}_i < 0,\ i = 2, 3$$

So we can get:

$$\begin{cases} w_0 > 0 & (1) \\ w_0 + w_2 < 0 & (2) \\ w_0 + w_1 < 0 & (3) \\ w_0 + w_1 + w_2 > 0 & (4) \end{cases}$$

From equation (1), (2) and (3) we know that $w_0 > 0$, $w_1 < 0$, $w_2 < 0$, and $|w_0| < |w_1|$, $|w_0| < |w_2|$. So equation (4) can't be achieved.

The minimum achievable training error is 1.

A feasible solution can be:

$$w_0 = 1,\ w_1 = -2,\ w_2 = -3$$

Note that we can change the value of $\boldsymbol{w}$ to fit any three of the equations above, so we only provide a feasible solution.

(2)
We set $\boldsymbol{w} = (w_0, w_1, w_2, w_3)^T$.
Likely, we can get:

$$\begin{cases} w_0 > 0 & (1) \\ w_0 + w_2 < 0 & (2) \\ w_0 + w_1 < 0 & (3) \\ w_0 + w_1 + w_2 + w_3 > 0 & (4) \end{cases}$$

This time we can achieve all the equations above by setting proper values of $\boldsymbol{w}$.

The minimum achievable training error is 0.

A feasible solution can be:

$$w_0 = 1,\ w_1 = -2,\ w_2 = -3,\ w_3 = 10$$

Problem 6: GD and GD with Regularization

Answer:
(1)

$$w_1 = w_1 + \alpha \cdot \frac{1}{N} \sum_{i=1}^{N} \left[y_i - \sigma(\boldsymbol{w}^T x_i) \right] x_{i,1}$$

$$= w_1 + \alpha \cdot \frac{1}{N} \left[\underbrace{(0 - \sigma(\boldsymbol{w}^T x_1)) \times 0 + \cdots + (0 - \sigma(\boldsymbol{w}^T x_k)) \times 0}_{most\ of\ the\ samples} + \underbrace{(1 - \sigma(\boldsymbol{w}^T x_{k+1})) \times 1 + \cdots + (1 - \sigma(\boldsymbol{w}^T x_{k+n})) \times 1}_{small\ number\ of\ the\ samples} \right]$$

$$= w_1 + \alpha \cdot \frac{1}{N} \cdot a \left[1 - \sigma(\boldsymbol{w}^T x^*) \right]$$

Its obvious that $k >> n$, and we denote that $a \left[1 - \sigma(\boldsymbol{w}^T x^*) \right] = (1 - \sigma(\boldsymbol{w}^T x_{k+1})) \times 1 + \cdots + (1 - \sigma(\boldsymbol{w}^T x_{k+n})) \times 1$. Because $\alpha > 0$, $N > 0$, $a > 0$, $0 < \sigma(\boldsymbol{w}^T x^*) < 1$, $0 < 1 - \sigma(\boldsymbol{w}^T x^*) < 1$, w_1 will always increase by iteration. It may get an infinity.

(2)
Likely, we can get:

$$w_1 = w_1 + \alpha \cdot \left[-\lambda w_1 + \frac{1}{N} \cdot a(1 - \sigma(\boldsymbol{w}^T x^*)) \right]$$

Initially, w_1 is relatively small, $\frac{1}{N} \cdot a(1 - \sigma(\boldsymbol{w}^T x^*)) > \lambda w_1$, w_1 will increase. λw_1 becomes larger, so w_1 will increase more and more slowly.
When w_1 becomes larger, i.e. $\lambda w_1 > \frac{1}{N} \cdot a(1 - \sigma(\boldsymbol{w}^T x^*))$, w_1 will decrease. Then w_1 will fluctuate in a fixed value w_1^*:

$$\lambda w_1 = \frac{1}{N} \cdot a(1 - \sigma(\boldsymbol{w}^T x^*))$$

$$w_1^* = \frac{1}{N \cdot \lambda} \cdot a(1 - \sigma(\boldsymbol{w}^T x^*))$$

Finally w_1 will converge to w_1^*.