Machine Learning, 2024 Spring
Assignment 2

2024.3.18

Consider **multi-class softmax regression** and **maximum likelihood estimation (MLE)**. Denote the samples as $(\mathbf{x}^{(1)}, y^{(1)}), (\mathbf{x}^{(2)}, y^{(2)}), ..., (\mathbf{x}^{(M)}, y^{(M)})$, where $y \in \{1, 2, ..., K\}$, and the weights in softmax regression as $\mathbf{w} = [\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_K]^T$.

Problem 1. Assume $K = 3$. For example, we are building a model to classify a number of people into 3 groups based on their smoking rate, where the labels are 'no smoking', 'light smoking', and 'heavy smoking'. Derive the log-likelihood. (6 pts)

Solution: Denote the probability of the i-th sample as $P_{i,1}$: no smoking, $P_{i,2}$: light smoking, and $P_{i,3}$: heavy smoking, then for any i,

$$P_{i,1} + P_{i,2} + P_{i,3} = 1.$$

Let $T_{i,k} = 1$ if $y^{(i)} = k$ and 0 otherwise be the indicator function. Then

$$P(y^{(i)} \mid \mathbf{x}^{(i)}, \mathbf{w}) = \prod_{k=1}^{3} P_{i,k}^{T_{i,k}}.$$

The likelihood function is

$$L(\mathbf{w}) = \prod_{i=1}^{M} P(y^{(i)} \mid \mathbf{x}^{(i)}, \mathbf{w}).$$

So the log-likelihood is

$$\begin{aligned}\log L(\mathbf{w}) &= \sum_{i=1}^{M} \log P(y^{(i)} \mid \mathbf{x}^{(i)}, \mathbf{w}) \\ &= \sum_{i=1}^{M} \sum_{k=1}^{3} T_{i,k} \log P_{i,k} \\ &= \sum_{i=1}^{M} (T_{i,1} \log P_{i,1} + T_{i,2} \log P_{i,2} + T_{i,3} \log (1 - P_{i,1} - P_{i,2})).\end{aligned}$$

(Also acceptable to assume the samples are i.i.d., i.e. no need to distinguish between $P_{i,1}$, $P_{i,2}$, $P_{i,3}$ for different i and directly setting p_1, p_2, p_3.)

Problem 2. For general K, derive the loss function / objective. (7 pts)

Solution: In general, softmax regression gives

$$P(y = k \mid \mathbf{x}, \mathbf{w}) = \frac{\exp(\mathbf{w}_k^T \mathbf{x})}{\sum_{j=1}^{K} \exp(\mathbf{w}_j^T \mathbf{x})}.$$

Let $T_{i,k} = 1$ if $y^{(i)} = k$ and 0 otherwise be the indicator function, then

$$P(y^{(i)} \mid \mathbf{x}^{(i)}, \mathbf{w}) = \prod_{k=1}^{K} \left[\frac{\exp(\mathbf{w}_k^T \mathbf{x}^{(i)})}{\sum_{j=1}^{K} \exp(\mathbf{w}_j^T \mathbf{x}^{(i)})} \right]^{T_{i,k}}.$$

The likelihood is

$$L(\mathbf{w}) = \prod_{i=1}^{M} P(y^{(i)} \mid \mathbf{x}^{(i)}, \mathbf{w}).$$

So the log-likelihood is

$$\begin{aligned}\log L(\mathbf{w}) &= \sum_{i=1}^{M} \log P(y^{(i)} \mid \mathbf{x}^{(i)}, \mathbf{w}) \\ &= \sum_{i=1}^{M}\sum_{k=1}^{K} T_{i,k} \log \frac{\exp(\mathbf{w}_k^T \mathbf{x}^{(i)})}{\sum_{j=1}^{K} \exp(\mathbf{w}_j^T \mathbf{x}^{(i)})}.\end{aligned}$$

The objective is to maximize $\log L(\mathbf{w})$.
(Or, the loss function is the negative log-likelihood,

$$J(\mathbf{w}) = -\log L(\mathbf{w}) = \ldots$$

The objective is to minimize $J(\mathbf{w})$.)

Problem 3. Based on Problem 2, we set $\mathbf{w}_K = 0$. You need to set another $\mathbf{w}_t = 0$ (choose t randomly by yourself) and derive the loss function. (7 pts)

Solution: Setting $\mathbf{w}_K = 0$ gives

$$\begin{aligned}J(\mathbf{w}) &= -\left[\sum_{i=1}^{M} T_{i,K} \log \frac{\exp(\mathbf{w}_K^T \mathbf{x}^{(i)})}{\sum_{j=1}^{K} \exp(\mathbf{w}_j^T \mathbf{x}^{(i)})} + \sum_{i=1}^{M}\sum_{k=1}^{K-1} T_{i,k} \log \frac{\exp(\mathbf{w}_k^T \mathbf{x}^{(i)})}{\sum_{j=1}^{K} \exp(\mathbf{w}_j^T \mathbf{x}^{(i)})}\right] \\ &= -\left[\sum_{i=1}^{M} T_{i,K} \log \frac{1}{\sum_{j=1}^{K} \exp(\mathbf{w}_j^T \mathbf{x}^{(i)})} + \sum_{i=1}^{M}\sum_{k=1}^{K-1} T_{i,k} \log \frac{\exp(\mathbf{w}_k^T \mathbf{x}^{(i)})}{\sum_{j=1}^{K} \exp(\mathbf{w}_j^T \mathbf{x}^{(i)})}\right].\end{aligned}$$

Further setting $\mathbf{w}_1 = 0$ gives

$$J(\mathbf{w}) = -\left[\sum_{i=1}^{M} (T_{i,K} + T_{i,1}) \log \frac{1}{\sum_{j=1}^{K} \exp(\mathbf{w}_j^T \mathbf{x}^{(i)})} + \sum_{i=1}^{M}\sum_{k=2}^{K-1} T_{i,k} \log \frac{\exp(\mathbf{w}_k^T \mathbf{x}^{(i)})}{\sum_{j=1}^{K} \exp(\mathbf{w}_j^T \mathbf{x}^{(i)})}\right].$$

(In general, setting $\mathbf{w}_t = 0$, random $t \in \{1, 2, \ldots, K-1\}$ gives

$$J(\mathbf{w}) = -\left[\sum_{i=1}^{M} (T_{i,K} + T_{i,t}) \log \frac{1}{\sum_{j=1}^{K} \exp(\mathbf{w}_j^T \mathbf{x}^{(i)})} + \sum_{i=1}^{M}\sum_{k=1}^{K-1} \mathbb{I}_{k \neq t} T_{i,k} \log \frac{\exp(\mathbf{w}_k^T \mathbf{x}^{(i)})}{\sum_{j=1}^{K} \exp(\mathbf{w}_j^T \mathbf{x}^{(i)})}\right],$$

where $\mathbb{I}_{k \neq t}$ is the indicator function which equals 0 when $k = t$ and 1 otherwise.)
(Can also write in the form of log-likelihood, i.e. omitting '-'.)