

Statistical Learning 2024 Fall

Homework 5

Chenchuan He

1. For logistic regression, show that

$$\nabla E_{\text{in}}(\mathbf{w}) = -\frac{1}{N} \sum_{n=1}^N \frac{y_n \mathbf{x}_n}{1 + e^{y_n \mathbf{w}^T \mathbf{x}_n}} = \frac{1}{N} \sum_{n=1}^N -y_n \mathbf{x}_n \theta(-y_n \mathbf{w}^T \mathbf{x}_n)$$

Argue that a 'misclassified' example contributes more to the gradient than a correctly classified one.

Solution

Since for a logistic regression with function θ , we have

$$\theta(s) = \frac{e^s}{1 + e^s} = \frac{1}{1 + e^{-s}}$$

In this case, we have

$$\theta(\mathbf{w}^T \mathbf{x}_n) = \ln(1 + e^{\mathbf{w}^T \mathbf{x}_n})$$

Thus for binary classification, we have

$$P(y_n = 1 | \mathbf{x}_n) = \theta(\mathbf{w}^T \mathbf{x}_n) = \frac{1}{1 + e^{-\mathbf{w}^T \mathbf{x}_n}}$$

$$P(y_n = -1 | \mathbf{x}_n) = 1 - \theta(\mathbf{w}^T \mathbf{x}_n) = \frac{1}{1 + e^{\mathbf{w}^T \mathbf{x}_n}}$$

Thus we have

$$P(y_n | \mathbf{x}_n) = \frac{1}{1 + e^{-y_n \mathbf{w}^T \mathbf{x}_n}}$$

Suppose we are using Log-Loss Function, we have

$$L(y_n, \hat{y}_n | \mathbf{x}_n) = -\ln(\hat{y}_n) = \ln(1 + e^{-y_n \mathbf{w}^T \mathbf{x}_n})$$

Thus, we have in-sample error

$$E_{\text{in}}(\mathbf{w}) = \frac{1}{N} \sum_{n=1}^N \ln(1 + e^{-y_n \mathbf{w}^T \mathbf{x}_n})$$

Taking gradient, we have

$$\nabla E_{\text{in}}(\mathbf{w}) = \frac{1}{N} \sum_{n=1}^N \frac{e^{-y_n \mathbf{w}^T \mathbf{x}_n} (-y_n \mathbf{x}_n)}{1 + e^{-y_n \mathbf{w}^T \mathbf{x}_n}} = -\frac{1}{N} \sum_{n=1}^N \frac{y_n \mathbf{x}_n}{1 + e^{y_n \mathbf{w}^T \mathbf{x}_n}}$$

Thus, we get

$$\nabla E_{\text{in}}(\mathbf{w}) = \frac{1}{N} \sum_{n=1}^N -y_n \mathbf{x}_n \theta(-y_n \mathbf{w}^T \mathbf{x}_n)$$

For a misclassified case, we have $-y_n \mathbf{w}^T \mathbf{x}_n > 0 \Leftrightarrow y_n \mathbf{w}^T \mathbf{x}_n < 0$,

$$\Rightarrow \frac{1}{1 + e^{y_n \mathbf{w}^T \mathbf{x}_n}} \text{ large}$$

On the other side, if it is a correctly classified case, we have $\frac{1}{1 + e^{y_n \mathbf{w}^T \mathbf{x}_n}}$ small.

As a result, we can conclude that misclassified examples contribute more to the gradient.

2. **Maximum likelihood estimator.** Let x_1, x_2, \dots, x_n be i.i.d. with Poisson distribution $P(\lambda)$:

$$f(k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k = 0, 1, 2, \dots, \infty \quad (1)$$

Find the maximum likelihood estimate of λ .

Solution

Likelihood function

$$L(\lambda) = \prod_1^n f(k) = \prod_1^n \frac{\lambda^k e^{-\lambda}}{k!}$$

Take \ln

$$\Rightarrow \ell(\lambda) = \ln(L(\lambda)) = \sum_1^n x_n \ln(\lambda) - \lambda - \ln(x_n!)$$

Get λ by solving the optimization problem: $\arg \max_{\lambda} \ell(\lambda)$

Take derivative,

$$\frac{d\ell(\lambda)}{d\lambda} = \sum_1^n \left(\frac{x_n}{\lambda} - 1\right) = \frac{1}{\lambda} \sum_1^n x_n - n = 0$$

$$\Rightarrow \lambda = \frac{1}{n} \sum_1^n x_n$$

To conclude, MLE of λ is:

$$\hat{\lambda} = \frac{1}{n} \sum_1^n x_n$$