

Machine Learning Basics Cheat Sheet

Shoichiro Yamanishi

May 16, 2020

1 Sample Distribution

Let $X = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$ be N i.i.d. samples drawn from certain (unknown) distribution $p(\mathbf{x})$, which we want to model with $p_{model}(\mathbf{x}|\boldsymbol{\theta})$. We define the joint probability distribution as follows.

$$\begin{aligned} p_{sample}(X) &= p_{sample}(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N) \\ &= \delta(\mathbf{x} - \mathbf{x}_1) \delta(\mathbf{x} - \mathbf{x}_2) \cdots \delta(\mathbf{x} - \mathbf{x}_N) \end{aligned} \quad (1)$$

This can be considered an infinitely overfitted probability model. This is used as a building block to derive MLE and the loss function for inference.

For a labeled training data set with additional labels $Y = \{y_1, y_2, \dots, y_N\}$, the sample conditional distribution $p_{sample}(Y|X)$ is defined as follows, assuming i.i.d.

$$\begin{aligned} p_{sample}(Y|X) &= p_{sample}(y_1, y_2, \dots, y_N | \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N) \\ &= \delta(y - y_1) \delta(y - y_2) \cdots \delta(y - y_N) \end{aligned} \quad (2)$$

Please note that $p_{sample}(X)$ and $p_{sample}(Y|X)$ uses Dirac's delta function, which exists only inside an integral.

2 KL Divergence and Cross Entropy

The KL divergence of the real (unknown) distribution $p_{real}(\mathbf{x})$ and the model $p_{model}(\mathbf{x})$ is formed as below.

$$\begin{aligned} D_{KL}(p_{real}(\mathbf{x}) || p_{model}(\mathbf{x}|\boldsymbol{\theta})) &= \int p_{real}(\mathbf{x}) (\ln p_{real}(\mathbf{x}) - \ln p_{model}(\mathbf{x}|\boldsymbol{\theta})) d\mathbf{x} \\ &= - \int p_{real}(\mathbf{x}) \ln p_{model}(\mathbf{x}|\boldsymbol{\theta}) d\mathbf{x} + const \end{aligned} \quad (3)$$

where we put $\int p_{real}(\mathbf{x}) \ln p_{real}(\mathbf{x}) d\mathbf{x}$ into a constant. This means minimizing the KL divergence is equal to minimizing the cross entropy $-\int p_{real}(\mathbf{x}) \ln p_{model}(\mathbf{x}|\boldsymbol{\theta}) d\mathbf{x}$.

Since we don't know $p_{real}(\mathbf{x})$, we substitute it with the sample distribution.

$$\begin{aligned} - \int p_{sample}(\mathbf{x}) \ln p_{model}(\mathbf{x}|\boldsymbol{\theta}) d\mathbf{x} &= - \int \delta(\mathbf{x} - \mathbf{x}_i) \ln p_{model}(\mathbf{x}|\boldsymbol{\theta}) d\mathbf{x} \\ &= - \ln p_{model}(\mathbf{x}_i|\boldsymbol{\theta}) \end{aligned} \quad (4)$$

For $X = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$, by i.i.d. assumption, $p_{sample}(X)$ and $p_{model}(X)$ are factorized into:

$$\begin{aligned} p_{sample}(X) &= \prod_{i=1}^N p_{sample}(\mathbf{x}_i) \\ p_{model}(X) &= \prod_{i=1}^N p_{model}(\mathbf{x}_i|\boldsymbol{\theta}) \end{aligned} \quad (5)$$

and

$$\begin{aligned} - \int p_{sample}(X) \ln p_{model}(X|\boldsymbol{\theta}) dX &= - \int \prod_{i=1}^K \delta(\mathbf{x}_k^* - \mathbf{x}_i) \sum_{i=1}^K \ln p_{model}(\mathbf{x}_k^*|\boldsymbol{\theta}) d\mathbf{x}_1^* d\mathbf{x}_2^* \dots d\mathbf{x}_N^* \\ &= - \sum_{i=1}^K \ln p_{model}(\mathbf{x}_i|\boldsymbol{\theta}) \end{aligned} \quad (6)$$

So, minimizing $-\sum_{i=1}^K \ln p_{model}(\mathbf{x}_i|\boldsymbol{\theta})$ w.r.t $\boldsymbol{\theta}$ is equivalent to the Maximum Likelihood Estimate.

For the labeled training data set,

$$\begin{aligned} - \int p_{sample}(Y|X) \ln p_{model}(Y|X, \boldsymbol{\theta}) dY &= - \int \prod_{i=1}^K \delta(y_k^* - y_i) \sum_{i=1}^K \ln p_{model}(y_k^*|\mathbf{x}_i, \boldsymbol{\theta}) dy_1^* dy_2^* \dots dy_N^* \\ &= - \sum_{i=1}^K \ln p_{model}(y_i|\mathbf{x}_i, \boldsymbol{\theta}) \end{aligned} \quad (7)$$

This is the *cross entropy* loss function for training.

3 Simple Linear Regression

$$p_{model}(y|\mathbf{x}) = \mathcal{N}(y|\mathbf{w}^T \mathbf{x}, \sigma^2)$$

$$\ln p_{model}(y|\mathbf{x}) \propto -\frac{1}{2\sigma^2}(y - \mathbf{w}^T \mathbf{x})^2$$

$$\nabla_{\mathbf{w}} \ln p_{model}(y|\mathbf{x}) = -\frac{1}{\sigma^2}(y - \mathbf{w}^T \mathbf{x})\mathbf{x} \quad (8)$$

$$\begin{aligned} \nabla_{\mathbf{w}} \nabla_{\mathbf{w}} \ln p_{model}(y|\mathbf{x}) &= H_{\mathbf{w}} (\ln p_{model}(y|\mathbf{x})) \\ &= \frac{1}{\sigma^2} \mathbf{x} \mathbf{x}^T \end{aligned}$$

4 Discriminative Binary Classification

$$y \in \{0, 1\}$$

$$\sigma(\mathbf{w}^T \mathbf{x}) = \frac{1}{1 + \exp(-\mathbf{w}^T \mathbf{x})}$$

$$\frac{\partial \sigma(\mathbf{w}^T \mathbf{x})}{\partial \mathbf{w}} = \sigma(\mathbf{w}^T \mathbf{x})(1 - \sigma(\mathbf{w}^T \mathbf{x}))\mathbf{x}$$

$$\begin{aligned} p_{model}(y|\mathbf{x}) &= \text{Bern}(y|\sigma(\mathbf{w}^T \mathbf{x})) \\ &= \sigma(\mathbf{w}^T \mathbf{x})^y + (1 - \sigma(\mathbf{w}^T \mathbf{x}))^{1-y} \end{aligned}$$

$$\ln p_{model}(y|\mathbf{x}) = y \ln \sigma(\mathbf{w}^T \mathbf{x}) + (1 - y) \ln(1 - \sigma(\mathbf{w}^T \mathbf{x}))$$

$$\begin{aligned} \nabla_{\mathbf{w}} \ln p_{model}(y|\mathbf{x}) &= \frac{y}{\sigma(\mathbf{w}^T \mathbf{x})} \sigma(\mathbf{w}^T \mathbf{x})(1 - \sigma(\mathbf{w}^T \mathbf{x}))\mathbf{x} - \frac{1 - y}{1 - \sigma(\mathbf{w}^T \mathbf{x})} \sigma(\mathbf{w}^T \mathbf{x})(1 - \sigma(\mathbf{w}^T \mathbf{x}))\mathbf{x} \\ &= y(1 - \sigma(\mathbf{w}^T \mathbf{x}))\mathbf{x} - (1 - y)\sigma(\mathbf{w}^T \mathbf{x})\mathbf{x} \\ &= (y - y\sigma(\mathbf{w}^T \mathbf{x}) - \sigma(\mathbf{w}^T \mathbf{x}) + y\sigma(\mathbf{w}^T \mathbf{x}))\mathbf{x} \\ &= (y - \sigma(\mathbf{w}^T \mathbf{x}))\mathbf{x} \end{aligned}$$

$$\begin{aligned} \nabla_{\mathbf{w}} \nabla_{\mathbf{w}} \ln p_{model}(y|\mathbf{x}) &= H_{\mathbf{w}} (\ln p_{model}(y|\mathbf{x})) \\ &= \nabla_{\mathbf{w}} (y - \sigma(\mathbf{w}^T \mathbf{x}))\mathbf{x} \\ &= -\sigma(\mathbf{w}^T \mathbf{x})(1 - \sigma(\mathbf{w}^T \mathbf{x}))\mathbf{x} \mathbf{x}^T \end{aligned} \quad (9)$$

5 Discriminative Multi-Class Classification

$$\mathbf{y} \in \{0, 1\}^K, \quad \sum_{i=1}^K y_i = 1 \quad (\text{one-of-K vector})$$

$$p_{model}(y|\mathbf{x}) = \frac{\exp(\mathbf{w}_j^T \mathbf{x})}{\sum_{k=1}^K \exp(\mathbf{w}_k^T \mathbf{x})} \quad \text{where } y_j = 1$$

$$\ln p_{model}(y|\mathbf{x}) = \mathbf{w}_j^T \mathbf{x} - \ln \sum_{k=1}^K \exp(\mathbf{w}_k^T \mathbf{x}) \quad \text{where } y_j = 1$$

$$\begin{aligned} \nabla_{\mathbf{w}_i} \ln p_{model}(y|\mathbf{x}) &= \delta_{i,j} \mathbf{x} - \nabla_{\mathbf{w}_i} \ln \sum_{k=1}^K \exp(\mathbf{w}_k^T \mathbf{x}) \\ &= \delta_{i,j} \mathbf{x} - \frac{\exp(\mathbf{w}_i^T \mathbf{x}) \mathbf{x}}{\sum_{k=1}^K \exp(\mathbf{w}_k^T \mathbf{x})} \\ &= \left(\delta_{i,j} - \frac{\exp(\mathbf{w}_i^T \mathbf{x})}{\sum_{k=1}^K \exp(\mathbf{w}_k^T \mathbf{x})} \right) \mathbf{x} \quad \text{where } y_j = 1 \end{aligned}$$

$$\begin{aligned} \nabla_{\mathbf{w}_i} \nabla_{\mathbf{w}_i} \ln p_{model}(y|\mathbf{x}) &= -\mathbf{x} \left(\frac{\exp(\mathbf{w}_i^T \mathbf{x}) \mathbf{x}^T}{\sum_{k=1}^K \exp(\mathbf{w}_k^T \mathbf{x})} - \exp(\mathbf{w}_i^T \mathbf{x}) \frac{\exp(\mathbf{w}_i^T \mathbf{x}) \mathbf{x}^T}{\left(\sum_{k=1}^K \exp(\mathbf{w}_k^T \mathbf{x}) \right)^2} \right) \\ &= -\mathbf{x} \mathbf{x}^T \left(\frac{\exp(\mathbf{w}_i^T \mathbf{x})}{\sum_{k=1}^K \exp(\mathbf{w}_k^T \mathbf{x})} \right) \left(1 - \frac{\exp(\mathbf{w}_i^T \mathbf{x})}{\sum_{k=1}^K \exp(\mathbf{w}_k^T \mathbf{x})} \right) \end{aligned}$$

$$\begin{aligned} \nabla_{\mathbf{w}_m} \nabla_{\mathbf{w}_i} \ln p_{model}(y|\mathbf{x}) &= -\mathbf{x} \left(\exp(\mathbf{w}_i^T \mathbf{x}) \frac{-\exp(\mathbf{w}_m^T \mathbf{x}) \mathbf{x}^T}{\left(\sum_{k=1}^K \exp(\mathbf{w}_k^T \mathbf{x}) \right)^2} \right) \\ &= \mathbf{x} \mathbf{x}^T \left(\frac{\exp(\mathbf{w}_i^T \mathbf{x})}{\sum_{k=1}^K \exp(\mathbf{w}_k^T \mathbf{x})} \right) \left(\frac{\exp(\mathbf{w}_i^T \mathbf{x})}{\sum_{k=1}^K \exp(\mathbf{w}_k^T \mathbf{x})} \right) \end{aligned}$$

where $m \neq i$

(10)