Machine Learning Basics Cheat Sheat

Shoichiro Yamanishi

May 22, 2020

1 Sample Distribution

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

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

This can be considered an inifinitely 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 $Y = \{y_1, y_2, \cdots y_N, \}$, the sample conditional distribution $p_{sample}(Y|X)$ is defined as follows, assuming i.i.d.

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

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 $p_{real}(\boldsymbol{x})$ and the model $p_{model}(\boldsymbol{x})$ is formed as below.

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

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

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

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

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

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

and

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

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

For the labeled training data set,

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

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

2.1 Mutual Information and Descriminiative Training

The mutual information is defined as follows.

$$
\begin{aligned}
I(\boldsymbol{x},\boldsymbol{y}) &= \int\int p(\boldsymbol{x},\boldsymbol{y}) \ln \frac{p(\boldsymbol{x},\boldsymbol{y})}{p(\boldsymbol{x})p(\boldsymbol{y})} d\boldsymbol{x} d\boldsymbol{y} \\
&= \int\int p(\boldsymbol{x},\boldsymbol{y}) \left(\ln \frac{p(\boldsymbol{x},\boldsymbol{y})}{p(\boldsymbol{x})} - p(\boldsymbol{y}) \right) d\boldsymbol{x} d\boldsymbol{y} \\
&= \int\int p(\boldsymbol{x},\boldsymbol{y}) \left(\ln \frac{p(\boldsymbol{x}|\boldsymbol{y})p(\boldsymbol{y})}{\int p(\boldsymbol{x}|\boldsymbol{y})p(\boldsymbol{y})d\boldsymbol{y}} - p(\boldsymbol{y}) \right) d\boldsymbol{x} d\boldsymbol{y} \\
&= \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}\sim p(\boldsymbol{x},\boldsymbol{y})} \left[\ln \frac{p(\boldsymbol{x}|\boldsymbol{y})p(\boldsymbol{y})}{\int p(\boldsymbol{x}|\boldsymbol{y})p(\boldsymbol{y})d\boldsymbol{y}} - p(\boldsymbol{y}) \right]
\end{aligned} \tag{8}
$$

For some discriminative tasks, you want to maximize the following instead of the maximum likelihood.

$$\underset{\boldsymbol{\theta}}{\operatorname{argmax}} \left\{ \ln p_{model}(Y|X, \boldsymbol{\theta}) \right\} \tag{9}$$

where $X = (\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N)$ are the observed inputs and $Y = (\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_N)$ are the observed outputs such as transcription. After the training and hence $\boldsymbol{\theta}$ is fixed, $\boldsymbol{y}$ is inferred from a given $\boldsymbol{x}$.

The conditional in equation 9 is expanded as follows.

$$\begin{aligned} \ln p_{model}(Y|X, \boldsymbol{\theta}) &= \sum_{i=1}^{N} \ln p_{model}(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{\theta}) \\ &= \sum_{i=1}^{N} \ln \left(\frac{p_{model}(\boldsymbol{x}_i|\boldsymbol{y}_i, \boldsymbol{\theta}) p(\boldsymbol{y}_i)}{\int p_{model}(\boldsymbol{x}_i|\boldsymbol{y}^*, \boldsymbol{\theta}) p(\boldsymbol{y}^*) d\boldsymbol{y}^*} \right) \\ &\approx \mathbb{E}_{\boldsymbol{x}, \boldsymbol{y} \sim p_{sample}(\boldsymbol{x}, \boldsymbol{y})} \left[\ln \frac{p_{model}(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta}) p(\boldsymbol{y})}{\int p_{model}(\boldsymbol{x}|\boldsymbol{y}^*, \boldsymbol{\theta}) p(\boldsymbol{y}^*) d\boldsymbol{y}^*} \right] \end{aligned} \tag{10}$$

With similarity to the definition to the mutual information, equation 9 is called *Maximum Mutual Information Estimate*. It is used for speech recognition etc.

As a regularizer, a factor κ is applied to the conditional probability density as follows.

$$\ln p_{model}(Y|X, \boldsymbol{\theta}) \approx \mathbb{E}_{\boldsymbol{x}, \boldsymbol{y} \sim p_{sample}(\boldsymbol{x}, \boldsymbol{y})} \left[\ln \frac{p_{model}(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta})^{\kappa} p(\boldsymbol{y})}{\int p_{model}(\boldsymbol{x}|\boldsymbol{y}^*, \boldsymbol{\theta})^{\kappa} p(\boldsymbol{y}^*) d\boldsymbol{y}^*} \right] \tag{11}$$

where $\kappa \approx 0.1$.

3 Simple Linear Regression

$$\begin{aligned} p_{model}(y|\boldsymbol{x}) &= \mathcal{N}(y|\boldsymbol{w}^T\boldsymbol{x}, \sigma^2) \\ \ln p_{model}(y|\boldsymbol{x}) &\propto -\frac{1}{2\sigma^2}(y - \boldsymbol{w}^T\boldsymbol{x})^2 \\ \nabla_{\boldsymbol{w}} \ln p_{model}(y|\boldsymbol{x}) &= -\frac{1}{\sigma^2}(y - \boldsymbol{w}^T\boldsymbol{x})\boldsymbol{x} \\ \nabla_{\boldsymbol{w}} \nabla_{\boldsymbol{w}} \ln p_{model}(y|\boldsymbol{x}) &= H_{\boldsymbol{w}} \left(ln p_{model}(y|\boldsymbol{x}) \right) \\ &= \frac{1}{\sigma^2}\boldsymbol{x}\boldsymbol{x}^T \end{aligned} \tag{12}$$

4 Discrinative Binary Classification

$$
\begin{aligned}
y &\in \{0, 1\} \\
\sigma(\boldsymbol{w}^T\boldsymbol{x}) &= \frac{1}{1+\exp(-\boldsymbol{w}^T\boldsymbol{x})} \\
\frac{\partial \sigma(\boldsymbol{w}^T\boldsymbol{x})}{\partial \boldsymbol{w}} &= \sigma(\boldsymbol{w}^T\boldsymbol{x})(1-\sigma(\boldsymbol{w}^T\boldsymbol{x}))\boldsymbol{x} \\
p_{model}(y|\boldsymbol{x}) &= Bern\left(y|\sigma(\boldsymbol{w}^T\boldsymbol{x})\right) \\
&= \sigma(\boldsymbol{w}^T\boldsymbol{x})^y + (1-\sigma(\boldsymbol{w}^T\boldsymbol{x}))^{1-y} \\
\\
\ln p_{model}(y|\boldsymbol{x}) &= y \ln \sigma(\boldsymbol{w}^T\boldsymbol{x}) + (1-y)\ln(1-\sigma(\boldsymbol{w}^T\boldsymbol{x})) \\
\\
\nabla_{\boldsymbol{w}} \ln p_{model}(y|\boldsymbol{x}) &= \frac{y}{\sigma(\boldsymbol{w}^T\boldsymbol{x})}\sigma(\boldsymbol{w}^T\boldsymbol{x})(1-\sigma(\boldsymbol{w}^T\boldsymbol{x}))\boldsymbol{x} - \frac{1-y}{1-\sigma(\boldsymbol{w}^T\boldsymbol{x})}\sigma(\boldsymbol{w}^T\boldsymbol{x})(1-\sigma(\boldsymbol{w}^T\boldsymbol{x}))\boldsymbol{x} \\
&= y(1-\sigma(\boldsymbol{w}^T\boldsymbol{x}))\boldsymbol{x} - (1-y)\sigma(\boldsymbol{w}^T\boldsymbol{x})\boldsymbol{x} \\
&= (y - y\sigma(\boldsymbol{w}^T\boldsymbol{x}) - \sigma(\boldsymbol{w}^T\boldsymbol{x}) + y\sigma(\boldsymbol{w}^T\boldsymbol{x}))\boldsymbol{x} \\
&= (y - \sigma(\boldsymbol{w}^T\boldsymbol{x}))\boldsymbol{x} \\
\\
\nabla_{\boldsymbol{w}}\nabla_{\boldsymbol{w}} \ln p_{model}(y|\boldsymbol{x}) &= H_{\boldsymbol{w}}\left(lnp_{model}(y|\boldsymbol{x})\right) \\
&= \nabla_{\boldsymbol{w}}(y - \sigma(\boldsymbol{w}^T\boldsymbol{x}))\boldsymbol{x} \\
&= -\sigma(\boldsymbol{w}^T\boldsymbol{x})(1-\sigma(\boldsymbol{w}^T\boldsymbol{x}))\boldsymbol{x}\boldsymbol{x}^T
\end{aligned}
\tag{13}
$$

5 Discrinative Multi-Class Classification

$$
\begin{aligned}
&\boldsymbol{y} \in \{0,1\}^K, \quad \sum_{i=1}^{K} y_i = 1 \quad \text{(one-of-K vector)} \\
&p_{model}(y|\boldsymbol{x}) = \frac{\exp(\boldsymbol{w}_j^T \boldsymbol{x})}{\sum_{k=1}^{K} \exp(\boldsymbol{w}_k^T \boldsymbol{x})} \quad \text{where } \ y_j = 1 \\
\\
&\ln p_{model}(y|\boldsymbol{x}) = \boldsymbol{w}_j^T \boldsymbol{x} - ln \sum_{k=1}^{K} \exp(\boldsymbol{w}_k^T \boldsymbol{x}) \quad \text{where } \ y_j = 1 \\
\\
\nabla_{\boldsymbol{w}_i} \ln p_{model}(y|\boldsymbol{x}) &= \delta_{i,j} \boldsymbol{x} - \nabla_{\boldsymbol{w}_i} \ln \sum_{k=1}^{K} \exp(\boldsymbol{w}_k^T \boldsymbol{x}) \\
&= \delta_{i,j} \boldsymbol{x} - \frac{\exp(\boldsymbol{w}_i^T \boldsymbol{x}) \boldsymbol{x}}{\sum_{k=1}^{K} \exp(\boldsymbol{w}_k^T \boldsymbol{x})} \\
&= \left(\delta_{i,j} - \frac{\exp(\boldsymbol{w}_i^T \boldsymbol{x})}{\sum_{k=1}^{K} \exp(\boldsymbol{w}_k^T \boldsymbol{x})} \right) \boldsymbol{x} \quad \text{where } \ y_j = 1 \\
\\
\nabla_{\boldsymbol{w}_i} \nabla_{\boldsymbol{w}_i} \ln p_{model}(y|\boldsymbol{x}) &= -\boldsymbol{x} \left(\frac{\exp(\boldsymbol{w}_i^T \boldsymbol{x}) \boldsymbol{x}^T}{\sum_{k=1}^{K} \exp(\boldsymbol{w}_k^T \boldsymbol{x})} - \exp(\boldsymbol{w}_i^T \boldsymbol{x}) \frac{\exp(\boldsymbol{w}_i^T \boldsymbol{x}) \boldsymbol{x}^T}{\left(\sum_{k=1}^{K} \exp(\boldsymbol{w}_k^T \boldsymbol{x}) \right)^2} \right) \\
&= -\boldsymbol{x}\boldsymbol{x}^T \left(\frac{\exp(\boldsymbol{w}_i^T \boldsymbol{x})}{\sum_{k=1}^{K} \exp(\boldsymbol{w}_k^T \boldsymbol{x})} \right) \left(1 - \frac{\exp(\boldsymbol{w}_i^T \boldsymbol{x})}{\sum_{k=1}^{K} \exp(\boldsymbol{w}_k^T \boldsymbol{x})} \right) \\
\\
\nabla_{\boldsymbol{w}_m} \nabla_{\boldsymbol{w}_i} \ln p_{model}(y|\boldsymbol{x}) &= -\boldsymbol{x} \left(\exp(\boldsymbol{w}_i^T \boldsymbol{x}) \frac{- \exp(\boldsymbol{w}_m^T \boldsymbol{x}) \boldsymbol{x}^T}{\left(\sum_{k=1}^{K} \exp(\boldsymbol{w}_k^T \boldsymbol{x}) \right)^2} \right) \\
&= \boldsymbol{x}\boldsymbol{x}^T \left(\frac{\exp(\boldsymbol{w}_i^T \boldsymbol{x})}{\sum_{k=1}^{K} \exp(\boldsymbol{w}_k^T \boldsymbol{x})} \right) \left(\frac{\exp(\boldsymbol{w}_i^T \boldsymbol{x})}{\sum_{k=1}^{K} \exp(\boldsymbol{w}_k^T \boldsymbol{x})} \right) \\
&\text{where } \ m \neq i
\end{aligned} \tag{14}
$$