Deep Learning Technology and Application

Ge Li

Peking University

关于 *Loss Function*

代价函数设计的基本方法

“Rather than guessing that some function might make a good estimator and then analyzing its bias and variance, we would like to have some principle from which we can derive specific functions that are good estimators for different models.”
“The most common such principle is the maximum likelihood principle.”

最大似然估计

- Consider: a set of m examples $\mathbb{X} = x^{(1)}, ..., x^{(m)}$, drawn independently from the true but unknown data generating distribution $p_{data}(x)$.
- Let $p_{model}(x, \theta)$ be a parametric family of probability distributions over the same space indexed by θ.
- Then, the maximum likelihood estimator for θ is defined as:

$$\theta_{ML} = \underset{\theta}{\operatorname{argmax}}\, p_{model}(\mathbb{X}; \theta) \tag{1}$$

$$= \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^{m} p_{model}(x^{(i)}; \theta) \tag{2}$$

最大似然估计

To obtain a more convenient but equivalent optimization problem, we transform a product into a sum:

$$\theta_{ML} = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{m} \log p_{model}(x^{(i)}; \theta) \tag{3}$$

Because the arg max does not change when we rescale the cost function, we can divide by m to obtain a version of the criterion that is expressed as an expectation with respect to the empirical distribution $p_{\hat{data}}$ defined by the training data:

$$\theta_{ML} = \underset{\theta}{\operatorname{argmax}} \mathbb{E}_{x \sim p_{\hat{data}}} \log p_{model}(x; \boldsymbol{\theta}) \tag{4}$$

最大似然估计的合理性

- One way to interpret maximum likelihood estimation is to view it as minimizing the ***dissimilarity*** between the empirical distribution $p_{\hat{data}}$ defined by the training set and the model distribution.
- The degree of dissimilarity between the two measured by the KL divergence:

$$D_{ML}(p_{\hat{data}}||p_{model}) = \mathbb{E}_{x\sim p_{\hat{data}}}[\log p_{\hat{data}}(x) - \log p_{model}(x)] \tag{5}$$

- The term on the left is a function only of the training data, not the model.(not include θ) This means when we try to minimize the KL divergence, we need only minimize:

$$-\mathbb{E}_{x\sim p_{\hat{data}}}[\log p_{model}(x)] \tag{6}$$

This is the same as the maximization in equation (**??**).

信息熵、交叉熵

- 信息熵，表示一个信源发出的信号的不确定程度。在信源发出的信号中，某信号出现的概率越大，熵越小；反之越大。
- 因此，信息熵的估算需要满足两个条件：
  1. 单调递减性，信息熵的值是信号 i 出现概率 p_i 的单调递减函数
  2. 可加性，两个独立符号所对应的不确定程度应等于各自不确定程度之和
- Shannon 用 log 函数定义信号 i（样本 i）的信息熵：

$$f(p(i)) = \log \frac{1}{p(i)} = -\log p(i)$$

- 则，包含 n 个样本的样本集合的信息熵定义为：

$$E(P) = \sum_{i}^{n} p(i) \log \frac{1}{p(i)}$$

注意：这里的 $p(i)$ 表示样本的真实分布；

信息熵、交叉熵

- 然而，在机器学习中，我们通常使用模型分布 $q(i)$ 来逼近真实分布，这时，样本集合的信息熵为：

$$E(P,Q)=\sum_{i}^{n}p(i)\log\frac{1}{q(i)}=-\sum_{i}^{n}p(i)\log q(i)$$

- 根据 Gibbs 不等式，有： $E(P,Q)\geq E(P)$
- 我们将 $E(P,Q)$ 与 $E(P)$ 的差，定义为“使用模型分布 Q 来逼近真实分布 P 时的 相对熵”（又称 KL 散度）：

$$D(P\parallel Q)=E(P,Q)-E(P)=\sum P(i)\log\frac{P(i)}{Q(i)}$$

- KL 散度，表示 2 个概率分布的差异程度；

最大似然估计的合理性

- One way to interpret maximum likelihood estimation is to view it as minimizing the ***dissimilarity*** between the empirical distribution $p_{\hat{data}}$ defined by the training set and the model distribution.
- The degree of dissimilarity between the two measured by the KL divergence:

$$D_{ML}(p_{\hat{data}}||p_{model}) = \mathbb{E}_{x\sim p_{\hat{data}}}[\log p_{\hat{data}}(x) - \log p_{model}(x)] \tag{7}$$

- The term on the left is a function only of the training data, not the model.(not include θ) This means when we try to minimize the KL divergence, we need only minimize:

$$-\mathbb{E}_{x\sim p_{\hat{data}}}[\log p_{model}(x)] \tag{8}$$

This is the same as the maximization in equation (**??**).

关于交叉熵与 KL 散度

Minimizing this KL divergence corresponds exactly to minimizing the crossentropy between the distributions.

- Many authors use the term "cross-entropy" to identify specifically the negative log-likelihood of a Bernoulli or softmax distribution, but that is a misnomer.

Any loss consisting of a negative log-likelihood is a cross entropy between the empirical distribution defined by the training set and the probability distribution defined by model.

- For example, mean squared error is the cross-entropy between the empirical distribution and a Gaussian model.

关于高斯分布

- Many distributions we wish to model are truly close to being normal distributions.
- Out of all possible probability distributions with the same variance, the normal distribution encodes the maximum amount of uncertainty over the real numbers.
- The central limit theorem shows that the sum of many independent random variables is approximately normally distributed.
- This means that in practice, many complicated systems can be modeled successfully as normally distributed noise, even if the system can be decomposed into parts with more structured behavior.
- We can think of the normal distribution as being the one that inserts the least amount of prior knowledge into a model.

关于高斯分布

- 若随机变量 X 服从一个位置参数为 μ、尺度参数为 σ 的概率分布，且其概率密度函数为:

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{(x-\mu)^2}{2\sigma^2}) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- 其中 μ 为随机变量 X 的数学期望，决定了正态分布的位置，σ^2 为随机变量 X 的方差，σ 为随机变量 X 的标准差，决定了正态分布的幅度。且正态分布函数在 $\mu \pm \sigma$ 处，有拐点。
- 记为：$N \sim (\mu, \sigma^2)$

Multivariate Gaussian Distribution

若 p 维随机向量 $X=(X_1,X_2,...X_p)'$ 的概率密度函数为：

$$f(x_1,x_2,...x_p)=\frac{1}{(\sqrt{2\pi})^p|\Sigma|^{\frac{1}{2}}}exp\{-\frac{1}{2}(X-\mu)'\Sigma^{-1}(X-\mu)\} \tag{9}$$

其中：μ 为 p 维向量，是 X 的均值向量，Σ 是 $p\times p$ 维协方差矩阵，即 p 阶正定矩阵，Σ^{-1} 是 Σ 的逆矩阵，$|\Sigma|$ 是 Σ 的行列式；

则称：$X=(X_1,X_2,...X_p)'$ 服从 p 维正态分布，记为 $X\sim N_p(\mu,\Sigma)$

特别的：当 $p=1$ 时，Σ 成为一个 1×1 的矩阵，$|\Sigma|^{\frac{1}{2}}$ 也就是标准差 σ，Σ^{-1} 也就是 σ^{-2}，而 $(X-\mu)'(X-\mu)$ 也变成 $(X-\mu)^2$，因此，多元正态分布就变成了一元正态分布。

同分布中心极限定理

设 $X_1, X_2, ..., X_n, ...$ 是独立同分布的随机变量序列，$EX_i = \mu, DX_i = \sigma^2 \neq 0 (i = 1, 2, ...)$，则对于任意的实数 x，有：

$$\lim_{n \to \infty} P\left(\frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma} < x \right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} \mathrm{d}t \qquad (10)$$

定理说明：

- 对独立同分布的随机变量序列 $X_1, X_2, ..., X_n, ...$，当 n 无限增大时，$Y_n = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma}$ 服从标准正态分布 $N(0, 1)$，而 $\sum_{i=1}^{n} X_i = \sqrt{n}\sigma Y_n + n\mu$ 则服从正态分布 $N(n\mu, n\sigma^2)$.
- 可见：当 n 足够大时，n 个独立同分布的随机变量之和，服从正态分布。

非同分布的李雅普诺夫定理

设 $X_1, X_2, ..., X_n, ...$ 是互相独立的随机变量，且 $EX_i = \mu, DX_i = \sigma^2 \neq 0(i = 1, 2, ...)$，若存在 $\delta > 0$，使得:

$$\lim_{n\to\infty} \frac{1}{B_n^{2+\delta}} \sum_{i=1}^{n} E\left|X_i - \mu_i\right|^{2+\delta} = 0 \text{ 其中: } B_n^2 = \sum_{i=1}^{n} \sigma_i^2 \tag{11}$$

则对于任意实数 x，有:

$$\lim_{n\to\infty} P\left(\frac{1}{B_n}\sum_{i=1}^{n}(X_i - \mu_i) < x\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}\mathrm{d}t} \tag{12}$$

非同分布的李雅普诺夫定理

李雅普诺夫定理表明：当 n 很大时：

$$Z_n = \frac{1}{B_n}\sum_{i=1}^{n}(X_i - \mu_i) = \frac{1}{B_n}(\sum_{i=1}^{n} X_i - \sum_{i=1}^{n}\mu_i) \tag{13}$$

近似服从正态分布 $N(0,1)$，即 $\sum_{i=1}^{n} X_i = B_n Z_n + \sum_{i=1}^{n}\mu_i$ 近似服从正态分布 $N(\sum_{i=1}^{n}\mu_i, B_n^2)$

- 不管随机变量 $X_1, X_2, ..., X_n, ...$ 各自具有怎样的分布，当 n 很大时，它们的和近似服从正态分布。
- 从理论上再次肯定了：大量随机因素叠加的结果，近似服从正态分布。

最大似然估计的合理性

- One way to interpret maximum likelihood estimation is to view it as minimizing the ***dissimilarity*** between the empirical distribution $p_{\hat{data}}$ defined by the training set and the model distribution.
- The degree of dissimilarity between the two measured by the KL divergence:

$$D_{ML}(p_{\hat{data}}||p_{model}) = \mathbb{E}_{x\sim p_{\hat{data}}}[\log p_{\hat{data}}(x) - \log p_{model}(x)] \tag{14}$$

- The term on the left is a function only of the training data, not the model.(not include θ) This means when we try to minimize the KL divergence, we need only minimize:

$$-\mathbb{E}_{x\sim p_{\hat{data}}}[\log p_{model}(x)] \tag{15}$$

This is the same as the maximization in equation (**??**).

条件概率的最大似然估计

The maximum likelihood estimator can readily be generalized to the case where our goal is to estimate a conditional probability $P(y|x;\theta)$.
If X represents all our inputs and Y all our observed targets, then the conditional maximum likelihood estimator is:

$$\theta_{ML} = \underset{\theta}{\operatorname{argmax}}\, P(Y|X;\boldsymbol{\theta}). \tag{16}$$

If the examples are assumed to be i.i.d. (independent identically distributed), then this can be decomposed into:

$$\theta_{ML} = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{m} \log p(y^{(i)}|x^{(i)};\boldsymbol{\theta}). \tag{17}$$

基于上述原则推导 MSE 的合理性

最小均方误差函数 (Mean Squared Error, MSE)：对于一组有 m 个样本的训练集，代价函数 MSE 定义如下:

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{m}(h_\theta(x_i) - y_i)^2$$

- 若，$h_\theta(x_i) = \theta^T x_i$，则为 Linear Regression.
- 此处，为不失一般性，仅假设 $h_\theta(x_i)$ 为神经网络输出层的输出值.

基于上述原则推导 MSE 的合理性

假设目标值与输入变量之间存在如下关系：

$$y_i = f(x_i;\theta) + \epsilon_i$$

由前文推导，可合理假设：$\epsilon_i \sim \mathcal{N}(0, \sigma^2)$，则其概率密度函数为：

$$p(\epsilon_i) = \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{\epsilon_i^2}{2\sigma^2}\right)$$

代入上式，并由误差的定义，可以推知：

$$p(y_i|x_i;\theta) = \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{(y_i - h_\theta(x_i))^2}{2\sigma^2}\right)$$

基于上述原则推导 MSE 的合理性

根据前文结论，求取代价函数的基本原理是：

在已知 m 个 (x_i, y_i) 的前提下，若对代价函数进行最小化，则能够使 “按照上述概率模型所得到的最大似然值” 最大化。

于是，这里，我们首先写出 “按照上述概率模型所得到的最大似然值”：

$$
\begin{aligned}
L(\theta) &= \prod_{i=1}^{m} p(y_i|x_i;\theta) \\
&= \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{(y_i - h_\theta(x_i))^2}{2\sigma^2}\right)
\end{aligned}
$$

基于上述原则推导 MSE 的合理性

对上式进行最大化求取，等价于：

$$
\begin{aligned}
\log\left(L(\theta)\right) &= log\prod_{i=1}^{m} p(y_i|x_i;\theta) \\
&= \sum_{i=1}^{m}\log\frac{1}{\sqrt{2\pi}\sigma}exp\left(-\frac{(y_i-h_\theta(x_i))^2}{2\sigma^2}\right) \\
&= m\log\frac{1}{\sqrt{2\pi}\sigma}-\frac{1}{\sigma^2}\cdot\frac{1}{2}\sum_{i=1}^{m}(y_i-h_\theta(x_i))^2
\end{aligned}
$$

要通过调整 θ，使上式最大化，只需要考虑使最后的二次项最小化即可，即最小化：

$$
\operatorname*{argmin}_{\theta}\frac{1}{2}\sum_{i=1}^{m}(y_i-h_\theta(x_i))^2 \quad \text{即：} \quad \operatorname*{argmin}_{\theta}\frac{1}{2}\sum_{i=1}^{m}(h_\theta(x_i)-y_i)^2
$$

基于伯努利分布的 Loss Function

若可假设网络输出满足如下分布：

$$p(y=1|x;\theta)=h_\theta(x)$$
$$p(y=0|x;\theta)=1-h_\theta(x)$$

上式可写为：

$$p(y|x;\theta)=(h_\theta(x))^y(1-h_\theta(x))^{1-y}$$

基于伯努利分布的 Loss Function

则，似然函数为：

$$\begin{aligned} L(\theta) &= p(Y|X;\theta) \\ &= \prod_{i=1}^{m} p(y^{(i)}|x^{(i)};\theta) \\ &= \prod_{i=1}^{m} \left((h_\theta(x))^{y^{(i)}} (1 - h_\theta(x))^{1-y^{(i)}} \right) \end{aligned}$$

进行 $\log$ 处理，得到需要最大化的 Loss Function 为：

$$\begin{aligned} l(\theta) &= \log L(\theta) \\ &= \sum_{i=1}^{m} \left(\log h(x^{(i)}) + (1 - y^{(i)}) \log (1 - h(x^{(i)})) \right) \end{aligned}$$

Logistic Regression

基于多项分布的 Loss Function

若可假设网络输出满足如下分布：

$$\begin{bmatrix} p(y^{(i)}=1|x^{(i)};\theta) \\ p(y^{(i)}=2|x^{(i)};\theta) \\ \vdots \\ p(y^{(i)}=k|x^{(i)};\theta) \end{bmatrix} = \frac{1}{\sum_{j=1}^{k} exp(f(x^{(i)};\theta_j))} \begin{bmatrix} exp(f(x^{(i)};\theta_1) \\ exp(f(x^{(i)};\theta_2) \\ \vdots \\ exp(f(x^{(i)};\theta_k) \end{bmatrix}$$

定义函数：

$$1\{\text{表达式为真}\} = 1$$

基于多项分布的 Loss Function

则，似然函数为：

$$\begin{aligned} L(\theta) &= p(Y|X;\theta) \\ &= \prod_{i=1}^{m} p(y^{(i)}|x^{(i)};\theta) \\ &= \prod_{i=1}^{m} \left(\sum_{j=1}^{k} 1\{y^{(i)} = j\} \frac{f(x^{(i)};\theta_j)}{\sum_{j=1}^{k} f(x^{(i)};\theta_j)} \right) \end{aligned}$$

进行 log 处理，得到需要最大化的 Loss Function 为：

$$\begin{aligned} l(\theta) &= \log L(\theta) \\ &= \sum_{i=1}^{m} \left(\sum_{j=1}^{k} \left(1\{y^{(i)} = j\}\right) \log \frac{f(x^{(i)};\theta_j)}{\sum_{j=1}^{k} f(x^{(i)};\theta_j)} \right) \end{aligned}$$

基于多项分布的 Loss Function

等同于最小化：

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}\left(\sum_{j=1}^{k}\left(1\{y^{(i)}=j\}\right)\log\frac{f(x^{(i)};\theta_j)}{\sum_{j=1}^{k} f(x^{(i)};\theta_j)}\right)$$

为便于计算，通常取 $f(x^{(i)};\theta_j) = exp(\theta_j^T)x^{(i)}$，得：

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}\left(\sum_{j=1}^{k}\left(1\{y^{(i)}=j\}\right)\log\frac{exp(\theta_j^T x^{(i)})}{\sum_{j=1}^{k} exp(\theta_j^T x^{(i)})}\right)$$

SoftMax

Thanks.