Maximal Entropy Model

1 Maximal Entropy Model

1.1 Background

The entropy is $H(x) = -\sum_x p(x)\log p(x) = \mathbb{E}_X[\log p(x)]$.

$$L(p(x), \lambda_0) = -\sum_x p(x)\log p(x) + \lambda_0(\sum_x p(x) - 1) \tag{1}$$

The joint entropy is,

$$H(X,Y) = -\sum_{X,Y} p(X,Y)\log p(X,Y) \tag{2}$$

The conditional entropy is,

$$\begin{aligned}
H(X|Y) &= -\sum_{y\in Y} p(y)H(X|Y=y) \\
&= -\sum_{y\in Y} p(y)\sum_{x\in X} p(x|y)\log p(x|y) \\
&= -\sum_{y\in Y}\sum_{x\in X} p(y)p(x|y)\log p(x|y) \\
&= -\sum_{y\in Y, x\in X} p(x,y)\log p(x|y) = -\mathbb{E}_{X,Y}[\log p(X|Y)] \\
&= -\sum_{y\in Y, x\in X} p(x,y)\log\frac{p(x,y)}{p(y)} \\
&= \sum_{y\in Y, x\in X} p(x,y)\log\frac{p(y)}{p(x,y)} \\
&= \sum_{y\in Y, x\in X} p(x,y)\log p(y) - \sum_{y\in Y, x\in X} p(x,y)\log p(x,y) \\
&= -H(X,Y) + \sum_{y\in Y}\sum_{x\in X} p(x,y)\log p(y) \\
&= -H(X,Y) + \sum_{y\in Y} p(y)\log p(y) \\
&= H(Y) - H(X,Y)
\end{aligned} \tag{3}$$

Bayes' rule for entropy,

$$H(X_1|X_2) = H(X_2|X_1) + H(X_1) - H(X_2) \tag{4}$$

Example of discrete distribution:

$$\begin{aligned}
L(p(x), \lambda_0) &= -\sum_x p(x)\log p(x) + \lambda_0(\sum_x p(x) - 1) \\
\frac{\partial L}{\partial p(x)} &= -\log p(x) - 1 + \lambda_0 = 0 \\
\therefore\ \ p(x) &= \exp(\lambda_0 - 1) = C,\ and \sum_x p(x) - 1 \\
\therefore\ \ p(x) &= \frac{1}{N}\ \rightarrow\ \text{uniform distribution}
\end{aligned} \tag{5}$$

Example of continuous distribution, given $\mu = \int_x xp(x)dx$. :

$$\begin{aligned} L(p(x), \lambda_0) &= -\int_x p(x)\log p(x)dx + \lambda_0(\int_x p(x)dx - 1) + \lambda_1(\int_x xp(x)dx - \mu) \\ \frac{\partial L}{\partial p(x)} &= -\log p(x) - 1 + \lambda_0 + \lambda_1 x = 0 \\ \therefore\ p(x) &= \exp(\lambda_0 - 1 + \lambda_1 x) = \frac{\exp(\lambda_1 x)}{\exp(1-\lambda_0)} \\ \because\ \int_x p(x)dx &= 1 \\ \therefore\ \int_x \frac{\exp(\lambda_1 x)}{\exp(1-\lambda_0)}dx &= \frac{\int_x \exp(\lambda_1 x)dx}{\exp(1-\lambda_0)} = 1 \\ \therefore\ \exp(1-\lambda_0) &= \int_x \exp(\lambda_1 x)dx \\ \therefore\ p(x) &= \frac{\exp(\lambda_1 x)}{\int_x \exp(\lambda_1 x)dx} \ \rightarrow \text{ exponential distribution} \end{aligned} \tag{6}$$

Example of continuous distribution, given $\sigma^2 = \int_x (x-\mu)^2 p(x)dx$:

$$\begin{aligned} L(p(x), \lambda_0) &= -\int_x p(x)\log p(x)dx + \lambda_0(\int_x p(x)dx - 1) + \lambda_1(\int_x (x-\mu)^2 p(x)dx - \sigma^2) \\ \frac{\partial L}{\partial p(x)} &= -\log p(x) - 1 + \lambda_0 + \lambda_1 x = 0 \\ \therefore\ p(x) &= \exp(\lambda_0 - 1 + \lambda_1 x) \ \rightarrow \text{ exponential distribution} \end{aligned} \tag{7}$$

1.2 Maximal Entropy Model (MEM)

Feature fucntion $f(x, y), i = 1, 2, \cdots, n$, where,

$$f(x,y) = \begin{cases} 1, \text{ if } (x,y) \text{ meets the condition,} \\ 0, \text{ otherwise.} \end{cases}$$

The expectation of feature function $f(x, y)$ about empirical distribution $\hat{P}(X, Y)$ (obtained from the training set) is,

$$\mathbb{E}_{\hat{P}}(f) = \sum_{x,y} \hat{P}(x,y) f(x,y) \tag{8}$$

The expectation of feature function $f(x, y)$ about empirical distribution $\hat{P}(Y|X)$ is,

$$\mathbb{E}_P(f) = \sum_{x,y} \hat{P}(x) P(y|x) f(x,y) \tag{9}$$

The maximal entropy model is

$$H(P) = -\sum_{x,y} \hat{P}(x) P(y|x) \log P(y|x) \tag{10}$$

and it meets,

$$\mathbb{E}_{\hat{P}}(f) = \mathbb{E}_P(f) \ \rightarrow \ \sum_{x,y} \hat{P}(x,y) f(x,y) = \sum_{x,y} \hat{P}(x) P(y|x) f(x,y) \tag{11}$$

Having the training set $T = \{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$. Assuming a set that satisfies all constraints is $\mathcal{C} = \{P \in \mathcal{P} | \, \mathbb{E}_P(f_i) = \mathbb{E}_{\hat{P}}(f_i), i = 1, 2, \cdots, n\}$ the learning of MEM is subjective to constrained optimization problem:

$$\begin{aligned} \max_{P \in \mathcal{C}} \ & H(P) = -\sum_{x,y} \hat{P}(x) P(y|x) \log P(y|x) \\ \text{s.t.} \ & \mathbb{E}_P(f_i) = \mathbb{E}_{\hat{P}}(f_i), i = 1, 2, \cdots, n \\ & \sum_y P(y|x) = 1 \end{aligned} \tag{12}$$

This problem can change to,

$$\begin{aligned} \min_{P \in \mathcal{C}} \ & -H(P) = \sum_{x,y} \hat{P}(x) P(y|x) \log P(y|x) \\ \text{s.t.} \ & \mathbb{E}_P(f_i) = \mathbb{E}_{\hat{P}}(f_i), i = 1, 2, \cdots, n \\ & \sum_y P(y|x) = 1 \end{aligned} \tag{13}$$

Build the Lagrange function $L(P(x), \lambda),\ \lambda = \{\lambda_0, \lambda_1\}$,

$$\begin{aligned} L(P(y|x), \lambda) &= -H(P) + \lambda_0 \sum_{i=1}^{n} (\mathbb{E}_P(f_i) - \mathbb{E}_{\hat{P}}(f_i)) + \lambda_1 (\sum_{i=1}^{n} P(y|x) - 1) \\ &= \sum_{x,y} \hat{P}(x) P(y|x) \log P(y|x) + \lambda_0 (\sum_{x,y} \hat{P}(x,y) f(x,y) - \sum_{x,y} \hat{P}(x) P(y|x) f(x,y)) + \lambda_1 (\sum_{y} P(y|x) - 1) \end{aligned} \tag{14}$$

The original problem is

$$\min_{P \in \mathcal{C}} \max_{\lambda} L(P(y|x), \lambda)$$

Therefore, the dual problem is

$$\max_{\lambda} \min_{P \in \mathcal{C}} L(P(y|x), \lambda)$$

Because $L(P(y|x), \lambda)$ is a convex function, therefore the dual problem is subjective to the original problem. So, we can get the solution of the original problem via calculating the solution of the dual problem. Let $\frac{\partial L(P(y|x), \lambda)}{\partial P(y|x)} = 0$,

$$\begin{aligned} \frac{\partial L(P(y|x), \lambda)}{\partial P(y|x)} =& \sum_{x,y} \hat{P}(x) (\log P(y|x) + 1) - \lambda_0 \sum_{x,y} \hat{P}(x) f(x,y) + \sum_{y} \lambda_1 = 0 \\ & \sum_{x,y} \hat{P}(x) \left[(\log P(y|x) + 1) - \lambda_0 \sum_{x,y} f(x,y) + \lambda_1 \right] = 0 \\ & \because \ \hat{P}(x) \text{ is the prior distribution } \ \therefore \ \hat{P}(x) \geq 0 \\ & (\log P(y|x) + 1) - \lambda_0 \sum_{x,y} f(x,y) + \lambda_1 = 0 \\ & \log P(y|x) = \lambda_0 \sum_{x,y} f(x,y) - \lambda_1 - 1 \\ & P(y|x) = \exp \left[\lambda_0 \sum_{x,y} f(x,y) - \lambda_1 - 1 \right] \\ & P(y|x) = \frac{\exp(\lambda_0 \sum\limits_{x,y} f(x,y))}{\exp(\lambda_1 + 1)} \\ & \because \ \sum_{y} p(y|x) = 1 \\ & \therefore \ \sum_{y} \frac{\exp(\lambda_0 \sum\limits_{x,y} f(x,y))}{\exp(\lambda_1 + 1)} = 1 \\ & P(y|x) = \frac{\exp(\lambda_0 \sum\limits_{x,y} f(x,y))}{Z(x,y)} \\ & Z(x,y) = \sum_{y} \exp(\lambda_0 \sum_{x,y} f(x,y)) \end{aligned} \tag{15}$$