最近在学习最大熵模型，看到极大似然估计这部分，没有看明白条件概率分布$p(y|x)$的对数似然函数。上网查了很多资料都没有一个合理的解释。基本直接给出对数似然函数的一般形式:

$$L_{\overline{p}} = \prod_{x} p(x)^{\overline{p}(x)}.$$

其实并没有解决问题。为了方便以后其他人的学习和理解，我结合自己的理解给出完整的解释。
其实第一眼之所以不理解，因为这是最大似然函数的另外一种形式。一般书上描述的最大似然函数的一般形式是各个样本集X中各个样本的联合概率:

$$L(x_1, x_2, ..., x_n; \theta) = \prod_{i=1}^{n} p(x_i; \theta).$$

其实这个公式和上式是等价的。$x_1, x_2, ..., x_n$是样本具体观测值。随机变量X是离散的，所以它的取值范围是一个集合，假设样本集的大小为n，X的取值有k个，分别是$v_1, v_2, ..., v_k$。用$C(X = v_i)$表示在观测值中样本v_i出现的频数。所以$L(x_1, x_2, ..., x_n; \theta)$可以表示为:

$$L(x_1, x_2, ..., x_n; \theta) = \prod_{i=1}^{k} p(v_i; \theta)^{C(X=v_i)}.$$

对等式两边同时开n次方，可得

$$L(x_1, x_2, ..., x_n; \theta)^{\frac{1}{n}} = \prod_{i=1}^{k} p(v_i; \theta)^{\frac{C(X=v_i)}{n}}.$$

因为经验概率$\overline{p}(x) = \frac{C(X=v_i)}{n}$，所以简写得到:

$$L(x_1, x_2, ..., x_n; \theta)^{\frac{1}{n}} = \prod_{x} p(x; \theta)^{\overline{p}(x)}.$$

很明显对$L(x_1, x_2, ..., x_n; \theta)$求最大值和对$L(x_1, x_2, ..., x_n; \theta)^{\frac{1}{n}}$求最大值的优化的结果是一样的。整理上式所以最终的最大似然函数可以表示为:

$$L(x; \theta) = \prod_{x} p(x: \theta)^{\overline{p}(x)}.$$

忽略θ，更一般的公式就是本文的第一个公式。结合公式一，参考v_JULY_v博客中的最大熵模型中的数学推导（http://m.blog.csdn.net/v_july_v/article/details/40508465），可得到联合概率密度的似然函数，即最大熵中的对数似然函数:

$$\begin{aligned}
L_{\overline{p}} &= \log \prod_{x,y} p(x, y)^{\overline{p}(x,y)} \\
&= \sum_{x,y} \overline{p}(x, y) \log p(x, y) \\
&= \sum_{x,y} \overline{p}(x, y) \log[\overline{p}(x) p(y|x)] \\
&= \sum_{x,y} \overline{p}(x, y) \log p(y|x) + \sum_{x,y} \overline{p}(x, y) \log \overline{p}(x)
\end{aligned}$$

上述公式第二项是一个常数项(都是样本的经验概率)，一旦样本集确定，就是个常数，可以忽略。所以最终的对数似然函数为:

$$L_{\overline{p}} = \sum_{x,y} \overline{p}(x, y) \log p(y|x).$$

上式就是最大熵模型中用到的对数似然函数。