概率模型的数学基础

1. Bayes 公式 (appendix_probability_and_information_theory.ipynb)

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}$$

> 问题：掷硬币，得到8次正面朝上，4次正面朝下，求硬币正反面的概率?

2. 数学推导

> 问题： 已知经验数据 $S = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$， 若此后给定一个新的 x 怎样预测它对应的 y 值?

我们把 $[x_1, x_2, \dots, x_n]$ 和 $[y_1, y_2, \dots, y_n]$ 分别看做是n维随机变量 $\mathcal{X} \equiv [X_1, \cdots X_n]$ 和 $\mathcal{Y} \equiv [Y_1, \cdots Y_n]$ 的一次采样结果 ((注意：各个 X_i, Y_i 本身也可以是多维的)

基于**i.i.d** (independent, identical distribution)假设, 我们认为描述$\mathcal{X}, \mathcal{Y}$ 的**"真实"**概率分布 $\mathcal{Q}(\mathcal{X}, \mathcal{Y})$ 满足:

$$\mathcal{Q}(\mathcal{X}, \mathcal{Y}) \equiv \prod_i q(X_i, Y_i)$$

不难证明，此时我们也有

$$\mathcal{Q}(\mathcal{Y}|\mathcal{X}) \equiv \prod_i q(Y_i|X_i)$$

而若给定一个新的x, 它对应的y预测值，由如下期望值决定:

$$\langle y \rangle = \int y * q(y|x)\mathrm{d}y \qquad (1)$$

我们试图用某种理论模型 $\mathcal{P}(\mathcal{Y}|\mathcal{X}, \theta) \equiv \prod_i p(Y_i|X_i, \theta)$ 来逼近 $\mathcal{Q}(\mathcal{Y}|\mathcal{X})$，即

$$\mathcal{P}(\mathcal{Y}|\mathcal{X}, \theta) \equiv \prod_i p(Y_i|X_i, \theta) \rightarrow \mathcal{Q}(\mathcal{Y}|\mathcal{X}) \equiv \prod_i q(Y_i|X_i) \qquad (2)$$

显然当上述条件满足, 可用$p(y|x, \theta)$ 来近似 (1)式中的 $q(y|x)$,

定义

$$\mathcal{P}(\theta|\mathcal{X}, \mathcal{Y}) \equiv \prod_i p(\theta|X_i, Y_i)$$

它表示在选定理论模型下，若已知$\{X_i\}, \{Y_i\}$, 那么 θ 的取值概率是多少。

Bayes inference

由于我们已知一组经验数据$\{(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)\}$，因此在已知的信息条件下，$\theta$的概率为

$$\mathcal{P}(\theta|\{x_i\},\{y_i\}) \equiv \prod_i p(\theta|x_i,y_i) = \prod_i \left(\frac{1}{C_i}p(y_i|x_i,\theta)p(\theta)\right) \qquad (3)$$

其中C_i 是归一化系数，定义为

$$C_i = \int p(y_i|x_i,\theta)p(\theta)\mathrm{d}\theta$$

因此，用 $p(y|x,\theta)$ 近似 $q(y|x)$，同时考虑到 θ 本身的概率性，(1)式可化为：

$$\langle y\rangle \approx \int y * p(y|x,\theta)\mathcal{P}(\theta|\{x_i\},\{y_i\})\mathrm{d}y\mathrm{d}\theta = \frac{1}{C}\int y * p(y|x,\theta)\prod_i\left[p(y_i|x_i,\theta)p(\theta)\right]\mathrm{d}y\mathrm{d}\theta \quad (4)$$

其中 $C \equiv \prod_i C_i$ 是整体的归一化系数, $p(\theta)$ 是选定的 prior distribution. 一般来说，我们在选择$p(\theta)$时应当考虑以下几个方面：

- **能够反映我们关于θ 的 naive belief**
- **数学表达式具有较好的拟合能力（capacity）**
- **数学表达式在推解析导中较容易计算 （详见概率论中conjugate prior (appendix_probability_and_information_theory.ipynb)的概念）**

注意这个公式允许我们在没有"模型训练"的前提下，就可以直接进行预测 （"空手套白狼"）， 预测的好坏取决于我们选择的"泛型"$p(y|x,\theta)$ 函数形式与实际问题的匹配程度

"鞍点"近似 (MAP, maximum posteriori)

实际应用中，（4）的数值计算过于困难。为简化计算，我们进一步引入"鞍点"近似，即我们仅考虑（3）中 θ^* 的贡献，其中 θ^* 对应 $\mathcal{P}(\theta|\{x_i\},\{y_i\})$ 的峰值位置，即

$$\begin{aligned}\theta^* &= \operatorname*{argmax}_\theta \mathcal{P}(\theta|\{x_i\},\{y_i\}) = \operatorname*{argmax}_\theta \frac{1}{n}\log\mathcal{P}(\theta|\{x_i\},\{y_i\}) \\ &= \operatorname*{argmax}_\theta \frac{1}{n}\log\prod_{i=1}^{n}(p(y_i|x_i,\theta)p(\theta)) \\ &= \operatorname*{argmax}_\theta \left[S(\theta)+\log p(\theta)\right] \\ S(\theta) &\equiv \frac{1}{n}\sum_{i=1}^{n}\log p(y_i|x_i,\theta)\end{aligned} \qquad (5)$$

其中 $S(\theta)$ 是 Maximum Likelihood 项, $\log p(\theta)$ 项相当于"正则化"

$$\langle y\rangle \approx \int y * p(y|x,\theta) * \delta(\theta-\theta^*)\mathrm{d}y\mathrm{d}\theta = \int y * p(y|x,\theta^*)\mathrm{d}y \qquad (6)$$

3. 重要关系

与正则化的关系

$\log p(\theta)$ 相当于正则化项，例如：取 $p(\theta)$ 为正态分布，则得到 Ridge 正则化

与 MLE的关系 （Maximum Likelihood Estimation）

当 $p(\theta)$ 是一个 trivial 的 **uniform** 分布时，有

$$\underset{\theta}{\operatorname{argmax}}\, \mathcal{P}(\theta|\{x_i\},\{y_i\}) = \underset{\theta}{\operatorname{argmax}}\, \mathcal{P}(\{y_i\}|\{x_i\},\theta)$$

即，**没有正则化项时，MAP “退化为” MLE**

与KL散度的关系 （Kullback–Leibler divergence）

$\underset{\theta}{\operatorname{argmax}}\, S(\theta)$ 可视为一个KL divergence项:

定义“经验”概率分布为 $q_e \equiv \frac{1}{n}\sum_i \delta(x-x_i)\delta(y-y_i)$ 则:

$$\begin{aligned}\underset{\theta}{\operatorname{argmax}}\, S(\theta) &= \underset{\theta}{\operatorname{argmax}} \int \mathrm{d}x\mathrm{d}y q_e(y|x)p(y|x,\theta) \\ &= \underset{\theta}{\operatorname{argmin}} \int \mathrm{d}x\mathrm{d}y q_e(y|x)\left(q_e(y|x)-p(y|x,\theta)\right) \\ &= \underset{\theta}{\operatorname{argmin}}\, KL\left(q_e\|p\right)\end{aligned}$$

与MSE 的关系 （线性回归，逻辑回归）

同样基于i.i.d.假设，我们认为单次采样的model分布具有如下高斯形式:

$$p(y|x,\theta) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(f(x,\theta)-y)^2}{2\sigma^2}\right]$$

则

$$\theta^* = \operatorname{argmin} S_\theta = \operatorname{argmin}\left(\sum_{i=1}^{n}\frac{1}{n}\|f(x_i,\theta)-y_i\|^2\right)$$

这个结论和前文model first“决定论”框架中基于 minimize training error $R_{\text{emp}}(\alpha)$ 的形式一致

求得最优参数θ^* 后，给定输入x，预测值y 为:

$$\langle y\rangle = \int y\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(f(x,\theta^*)-y)^2}{2\sigma^2}\right]\mathrm{d}y = f(x,\theta^*)$$