Logistic 回归 & Softmax 回归

⑬ 熵，KL 距离与 LR 回归

相对熵（概率⑱）中，$p(x)$ 是样本分布，$q(x)$ 是我们建模时定义的一个简单分布，如何度量 $q(x)$ 与 $p(x)$ 的接近程度呢？

→ 使用相对熵（K-L距离）来定义

$$KL = \sum_x p(x) \ln \frac{p(x)}{q(x)}$$

$$= \sum_x p(x) \left(\ln p(x) - \ln q(x)\right) = \sum_x p(x) \ln p(x) - \sum_x p(x) \ln q(x)$$

$$= \underbrace{\cancel{H(x)}}_{\text{样本的信息熵，固定值，可忽略}} \boxed{- \sum_x p(x) \ln q(x)}$$

⇐ 优化目标，希望 $-\sum_x p(x) \ln q(x)$ 取最小值.

→ 如何求 $-\sum_x p(x) \ln q(x)$ 的最小值

在LR中，似然概率 $q(x) = \underbrace{g}_{\text{sigmoid函数}}(\theta^T x) = \dfrac{1}{1+e^{-\theta^T x}}$

而 $-\sum_x \underbrace{p(x)}_{\text{样本概率}} \ln \underbrace{q(x)}_{\text{似然概率}} = \underbrace{-\sum_i y^{(i)} \cdot \ln \dfrac{1}{1+e^{-\theta^T x^{(i)}}}}$

也就是 $\left\{\begin{array}{l}\text{正例 } y^{(i)}=1 \\ \text{负例 } y^{(i)}=0\end{array}\right\}$ 时的负对数似然函数（来自正例）

→ 相对熵取最小值 ⇔ 负对数似然取最小值

⇔ 正对数似然函数取最大值

$\boxed{\text{参考 LR\&Softmax ③}}$

→ ~~取 {正例 $y^{(i)}=1$，负例 $y^{(i)}=-1$} 的似然对数似然函数~~

同理 把正例负例都考虑进来 得到正对数似然函数

$$\underbrace{l(\theta)}_{\text{Likelyhood}} = \ln L(\theta) = \sum_{i=1}^{m} \left(y^{(i)} \ln \frac{1}{1+e^{-\theta^T x^{(i)}}} + (1-y^{(i)}) \ln \left(1 - \frac{1}{1+e^{-\theta^T x^{(i)}}}\right)\right)$$

→ 因此，求解这个正对数似然函数的最大值，

就等价于求解 相对熵的最小值.

④拉格朗日对偶性

(a) 用途：约束条件下求极值，支持向量机，最大熵模型

(b) 原始问题：求解目标　约束条件

① $\min_{x\in R^n} f(x)$　s.t. $c_i(x) \le 0$; $h_j(x) = 0$
$i=1,\cdots,k$　$j=1,2,\cdots,l$

$f(x)$, $c_i(x)$, $h_j(x)$ 都在 R^n 上连续可微

② 引入拉格朗日函数：

$$L(x,\alpha,\beta) = f(x) + \sum_{i=1}^{k} \alpha_i \boxed{c_i(x)} + \sum_{j=1}^{l} \beta_j \boxed{h_j(x)}$$

≤ 0　$=0$

拉格朗日乘子 $\alpha_i \ge 0$

拉格朗日乘子（对正负无要求）

这一项 ≤ 0　这一项 $=0$

因此可推导出

$$L(x,\alpha,\beta) \le f(x)$$

(*) 定义：$\boxed{\theta_p(x) = \max_{\alpha,\beta;\alpha_i\ge 0} L(x,\alpha,\beta)}$，则要求 $\min_x f(x)$ 等价于求 $\min_x \theta_p(x)$

即：原始问题最优解为

$$p^* = \min_x \theta_p(x)$$

☆ 原问题 $\min_x \theta_p(x) = \min_x \max_{\alpha,\beta:\alpha_i\ge 0} L(x,\alpha,\beta) = \min_x f(x)$

(c) 对偶问题：

(*) 定义 $\boxed{\theta_D(\alpha,\beta) = \min_x L(x,\alpha,\beta)}$，再对这个定义式求极大值.

☆ 对偶问题 $\max_{\alpha;\beta;\alpha_i\ge 0} \theta_D(\alpha,\beta) = \max_{\alpha,\beta:\alpha_i\ge 0} \min_x L(x,\alpha,\beta)$ ← 原问题的对偶问题.

(*) 如果原问题，对偶问题都有最优解，则有

$$d^* = \max_{\alpha;\beta:\alpha_i\ge 0} \min_x L(x,\alpha,\beta) \quad \text{// 对偶问题 } \theta_D(\alpha,\beta)$$

$$\le \min_x \max_{\alpha,\beta:\alpha_i\ge 0} L(x,\alpha,\beta) = p^* \quad \text{// 原问题 } \theta_p(x)$$

证明：

$\because \theta_D(\alpha,\beta) = \min_x L(x,\alpha,\beta) \le L(x,\alpha,\beta) \le \max_{\alpha;\beta:\alpha_i\ge 0} L(x,\alpha,\beta) = \theta_p(x)$

$\therefore \theta_D(\alpha,\beta) \le \theta_p(x)$

又∵ $\theta_D(\alpha,\beta)$ 和 $\theta_P(x)$ 都有最优解

∴ $\max\limits_{\alpha,\beta;\alpha_i\ge 0} \theta_D(\alpha,\beta) \le \min\limits_{x} \theta_P(x)$

<d> KKT条件：为了使对偶问题的最优解，也是原始问题最优解，

必须满足KKT条件（充要条件）

条件(1~3)：$f(x)$，$c_i(x)$，$h_j(x)$ 全部连续可导，

这样才能对 $L(x,\alpha,\beta)=f(x)+\sum\limits_{i=1}^{k}\alpha_i c_i(x)+\sum\limits_{j=1}^{l}\beta_j h_j(x)$ 关于 x,α_i,β_j

求偏导，令 $\nabla_x L(x^*,\alpha^*,\beta^*)=0$，$\nabla_\alpha L(x^*,\alpha^*,\beta^*)=0$，

$\nabla_\beta L(x^*,\alpha^*,\beta^*)=0$

条件(4)：存在 α_i^*，x^*，使得对 $\forall i\in[1,k]$，$\alpha_i^* c_i(x^*)=0$。

$\alpha_i^*\ge 0$，$c_i(x^*)\le 0$（约束条件），同时要让不等式 $\le$ 取等号 $\sum\limits_{i=1}^{k}\alpha_i^* c_i(x^*)$

必须等于0，又因为 推出 $\alpha_i^* c_i(x^*)\le 0$，因此只能让每个 $\alpha_i^* c_i(x^*)=0$.

条件(5,7)：$c_i(x^*)\le 0$，$h_j(x^*)=0$ 原始约束条件

条件(6)：$\alpha_i\ge 0$ 来自拉格朗日乘子法定义

<e> 为什么要用对偶问题求原始问题的最优解.

当对偶问题求解比原始问题简单时.