Statistics – Information Critierion

Yabusame

2024-05-23

1 信息熵的基础知识

定义随机变量 x 的熵(离散情况):

$$H[x] = -\sum_x p(x)\log_2 p(x)$$

(Noiseless Coding Theorem) 熵是传输随机变量状态的比特数的下界

(Maximum Entropy)

$$\widetilde{H} = -\sum_i p(x_i)\ln p(x_i) + \lambda\left(\sum_i p(x_i) - 1\right) \overset{\text{maximize}}{\Rightarrow} p(x_i) = \frac{1}{M}; H = \ln M$$

where M is the total number of the state

(Differential Entropy) 在(连续分布) $p(x)$下，观测到随机变量x_i的概率 $\Rightarrow p(x_i)\Delta$

$$\lim_{\Delta\to 0}\left\{-\sum_i p(x_i)\Delta\ln p(x_i)\right\} = -\int p(x)\ln p(x)\,\mathrm{d}x = H[x]$$

(Kullback-Leibler Divergence)

$$\begin{aligned}\mathrm{KL}(p\|q) &= -\int p(\boldsymbol{x})\ln q(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} - \left(-\int p(\boldsymbol{x})\ln p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\right)\\ &= -\int p(\boldsymbol{x})\ln\frac{q(\boldsymbol{x})}{p(\boldsymbol{x})}\,\mathrm{d}\boldsymbol{x}\end{aligned}$$

注意 $\mathrm{KL}(p\|q) \neq \mathrm{KL}(q\|p)$，所以它并不是 metric (Jensen 不等式证明等号成立的条件是 $\forall \boldsymbol{x} : p(\boldsymbol{x}) = q(\boldsymbol{x})$)

(Conditional Entropy) 考虑联合分布，在已知 x 的情况下确定 y 值所需要的信息是 $-\ln p(y|x) \Rightarrow$ 所对应的信息熵 $H[y|x] = -\iint p(\boldsymbol{x},\boldsymbol{y})\ln p(\boldsymbol{y}|\boldsymbol{x})\,\mathrm{d}\boldsymbol{y}\,\mathrm{d}\boldsymbol{x}$

$$H[x,y] = H[y|x] + H[x]$$

(Mutual Information) $\boldsymbol{x}$和$\boldsymbol{y}$之间的相互信息

$$I[\boldsymbol{x},\boldsymbol{y}] = \mathrm{KL}[p(\boldsymbol{x},\boldsymbol{y})\|p(\boldsymbol{x})p(\boldsymbol{y})] = -\iint p(\boldsymbol{x},\boldsymbol{y})\ln\left(\frac{p(\boldsymbol{x})p(\boldsymbol{y})}{p(\boldsymbol{x},\boldsymbol{y})}\right)\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{y}$$

2 (Bouns) 信息准则的几个谣传

AIC 和 BIC 的原始形式如下:

$$\mathrm{AIC}_m = -2\sum_{i=1}^{n}\log\big(p_{\hat{\theta}}(y_i)\big) + 2d_m$$

$$\mathrm{BIC}_m = -2\sum_{i=1}^{n}\log\big(p_{\hat{\theta}}(y_i)\big) + d_m\log(n)$$

2.1 AIC 适合预测, BIC 适合解释？

这种看法忽略了参数化和非参数化情形, AIC 可能在$n \to \infty$时也不能取到最好模型，在非参数情况下, BIC 选择生成数据模型的一致性也不是良好定义的。

2.2 应该使用 AIC, 因为现实情况更常见到非参数化的情形？

在实际情况中信息准则也会受到样本数的影响，例如：

1. 在样本少非参数的情况下， BIC 可以察觉到突出的模型
2. 在参数化的情况下，系数在不同的数量级上很小，并且样本数不足以 估计它们，在这种情况下选择模型使用 AIC 更加适合
3. 主要有部分参数化、部分非参数化的两种情况

2.3 penlity l_0 不如 penlity (LASSO, SCAD, MCP), 因为它是离散的？

信息准则相当于带有l_0 penlity 的回归， penlity 是否正确取决于计算的目的

- 即使是对于固定调整参数(fixed tuning parameter)，选择模型的能力与 penlity function 的连续与否没有直接关系
- 固定调整参数(fixed tuning parameter)的选择基于数据，其他 penlity 不一定会给出更好结果
- 事实上l_0 penlity 方式以最少的约束条件得到了 minimax rate 最佳值

复杂的理论看不懂…直接到建议部分：

1. 当选择模型是为了预测的时候：
   - 依据参数化指标或者交叉验证选择 AIC-Type 或 BIC-Type 方法
   - 使用适应性的信息准则(例如 BC) 结合 AIC 和 BIC 的方式
2. 当希望从后随观测的相似的样本大小中重新得到选择参数时(选择变量用以解释模型)—使用 BIC-Type 防止引入不重要的变量
3. 如果保护最坏情况预测精度非常重要，首选 AIC — 极小极大速率(minmax rate)最优性 此外对于寻找可能相关变量的探索性研究，尽管 AIC 型方法可能会过度选择，但不会遗漏重要变量，这些变量可以在大样本量的后续研究中验证
4. 当预测变量的数量 d 与样本大小 n 相比并不小，并且考虑 d 变量的所有子集时，最好使用高维 AIC 或 BIC 来解决潜在的严重选择偏差
5. 当模型选择不稳定性较高时，出于预测目的，可以考虑模型平均