评估假设

Outline

Topic

动机

- 当数据十分充足时，假设精度的估计相对容易。
- 然而当给定的数据集非常有限时，要学习一个概念并估计其将来的精度，存在两个很关键的困难：
  - 估计的偏差
  - 估计的方差

估计的偏差 (Bias in the estimate)。

- 学习到的概念在训练样例上的观察精度通常不能很好地用于估计在将来样例上的精度。
- 因为假设是从这些样例中得出的，因此对将来样例的精度估计通常偏于乐观。
- 尤其在学习器采用了很大的假设空间，并过度拟合训练样例时，这一情况更可能出现。
- 要对将来的精度进行无偏估计，典型的方法是选择与训练样例和假设无关的检验样例，在这个样例集合上检验假设。

估计的方差 (Variance in the estimate)。

- 即使假设精度在独立的无偏检验样例上测量，得到的精度仍可能与真实精度不同,(取决于特定检验样例集合的组成)。
- 检验样例越少，产生的方差越大。

Topic

真实错误率

- 定义：
  假设 h 关于目标函数 f 和分布 D 的真实错误率 (由 $error_D(h)$ 表示)(即：按分布 D 随机抽取的实例在假设 h 下被误分类的概率)：

$$error_{\mathcal{D}}(h) \equiv \Pr_{x \in \mathcal{D}}[f(x) \neq h(x)]$$

  这里，记号 $\Pr_{x \in \mathcal{D}}$ 表示概率在实例分布 $\mathcal{D}$ 上计算。

样本错误率

- 定义：假设 h 关于目标函数 f 和数据样本 S 的样本错误率（标记为 $error_S(h)$ ）为：

$$error_S(h) \equiv \frac{1}{n} \sum_{x \in S} \delta(f(x) \neq h(x))$$

其中当 $f(x) \neq h(x)$ 时 $\delta(f(x) \neq h(x))$ 为 1，其它时为 0.

偏差与方差

- 偏差 (Bias) :
  - 当 S 为训练集, $error_S(h)$ 为：

$$bias \equiv E[error_S(h)] - error_{\mathcal{D}}(h)$$

  - 对于无偏估计，h 与 S 必需独立选取。
- 方差 (Variance)
  - 即使对于无偏的 S , $error_S(h)$ 仍与 $error_{\mathcal{D}}(h)$ 不同

示例：

- 假设 h 误分类 s 中 40 个样例中的 12 个

$$error_S(h) = \frac{12}{40} = 0.3$$

  $error_{\mathcal{D}}(h)$?
- 估计量：
  - 依概率 $\mathcal{D}$ 选取容量 n 的样本 s
  - $error_S(h)$ 是随机变量
  - $error_S(h)$ 是 $error_{\mathcal{D}}(h)$ 的无偏估计

离散值假设的置信区间

基于某离散值假设 h 在样本 S 上观察到的样本错误率，估计它的真实错误率，其中：

- 样本 S 包含 n 个样例，它们的抽取按照概率分布 $\mathcal{D}$，抽取过程是相互独立的，并且不依赖于 h
- $n \geq 30$
- 假设 h 在这 n 个样例上犯了 r 个错误（$error_S(h) = r/n$）

已知这些条件，统计理论可给出以下断言：

- 没有其他信息的话，$error_{\mathcal{D}}(h)$ 最可能的值为 $error_S(h)$
- 有大约 95% 的可能性，真实错误率 $error_{\mathcal{D}}(h)$ 处于区间：

$$error_S(h) \pm 1.96\sqrt{\frac{error_S(h)(1-error_S(h))}{n}}$$

- 有大约 N% 的可能性，真实错误率 $error_{\mathcal{D}}(h)$ 处于区间：

$$error_S(h) \pm z_N\sqrt{\frac{error_S(h)(1-error_S(h))}{n}}$$

其中

$N\%$:	50%	68%	80%	90%	95%	98%	99%
z_N:	0.67	1.00	1.28	1.64	1.96	2.33	2.58

Topic

错误率估计和二项比例估计

- $error_S(h)$ 是一个随机变量。
- 随机抽取容量为 n 的样本 s，观测到 r 个误分类样例的概率为：

$$P(r) = \frac{n!}{r!(n-r)!} \, error_{\mathcal{D}}(h)^r (1 - error_{\mathcal{D}}(h))^{n-r}$$

二项分布

若一次掷币出现正面的概率 $p = \Pr(heads)$，则 n 次掷硬币出现 r 次正面的概率 $P(r)$：

$$P(r) = \frac{n!}{r!(n-r)!} p^r (1-p)^{n-r}$$

- 期望：

$$E[X] \equiv \sum_{i=0}^{n} iP(i) = np$$

- 方差：

$$Var(X) \equiv E[(X - E[X])^2] = np(1-p)$$

- 标准差：

$$\sigma_X \equiv \sqrt{E[(X - E[X])^2]} = \sqrt{np(1-p)}$$

正态分布逼近二项分布

- $error_S(h)$ 服从二项分布：
  - 期望：$\mu_{error_S(h)} = error_{\mathcal{D}}(h)$
  - 标准差：$\sigma_{error_S(h)} = \sqrt{\frac{error_{\mathcal{D}}(h)(1-error_{\mathcal{D}}(h))}{n}}$
- 用正态分布逼近：
  - 期望：$\mu_{error_S(h)} = error_{\mathcal{D}}(h)$
  - 标准差：$\sigma_{error_S(h)} \approx \sqrt{\frac{error_S(h)(1-error_S(h))}{n}}$
- 得正态分布：

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

置信区间计算

置信区间 (confidence interval)：某个参数 p 的 N% 置信区间是一个以 N% 的概率包含 p 的区间。

- X 落入区间 (a, b) 的概率为：$\int_a^b p(x)dx$
- 期望：$E[X] = \mu$
- 方差：$Var(X) = \sigma^2$
- 标准差：$\sigma_X = \sigma$
- 80% 的概率位于 $\mu \pm 1.28\sigma$ 之间
- N% 的概率位于 $\mu \pm z_N\sigma$ 之间

$N\%$:	50%	68%	80%	90%	95%	98%	99%
z_N:	0.67	1.00	1.28	1.64	1.96	2.33	2.58

置信区间分析

- S 包含 n 个独立抽取的样例，且独立于 h
- $n \geq 30$

则
近似 95% 的概率，$error_S(h)$ 在区间

$$error_{\mathcal{D}}(h) \pm 1.96\sqrt{\frac{error_{\mathcal{D}}(h)(1-error_{\mathcal{D}}(h))}{n}}$$

同样，$error_{\mathcal{D}}(h)$ 在区间

$$error_S(h) \pm 1.96\sqrt{\frac{error_{\mathcal{D}}(h)(1-error_{\mathcal{D}}(h))}{n}}$$

近似有

$$error_S(h) \pm 1.96\sqrt{\frac{error_S(h)(1-error_S(h))}{n}}$$

双侧和单侧边界

- 双侧边界：给出了估计量的上界和下界
- 单侧边界：
  - 只限定 h 的最大借误率

Topic

中心极限定理

考虑独立同分布的随机变量 $Y_1 \ldots Y_n$ 的集合，它们服从一任意的概率分布，均值为 μ，有限方差 σ^2。定义样本均值

$$\bar{Y} \equiv \frac{1}{n} \sum_{i=1}^{n} Y_i$$

则当 $n \to \infty$ 时 $\bar{Y}$ 服从正态分布，均值为 μ 且标准差为 $\frac{\sigma^2}{n}$。

计算置信区间

通用的过程包含以下步骤：

- 确定基准总体中要估计的参数 p，例如 $error_D(h)$ 。
- 定义一个估计量 Y (如 $error_S(h)$) 它的选择应为最小方差的无偏估计量。
- 确定估计量所服从的概率分布 D_Y，包括其均值和方差。
- 确定 N% 置信区间，通过寻找阈值 L 和 U 以使这个按 D_Y 分布的随机变量有 $N\%$ 机会落入 L 和 U 之间。

Topic

两个假设错误率间的差异

假设 h_1 在一拥有 n_1 个独立抽取样例的样本 S_1 上测试，且 h_2 在 n_2 个独立抽取样例的样本 S_2 上测试。

- 估计两个假设的真实错误率间的差异

$$d \equiv error_{\mathcal{D}}(h_1) - error_{\mathcal{D}}(h_2)$$

- 选取估计量

$$\hat{d} \equiv error_{S_1}(h_1) - error_{S_2}(h_2)$$

- 确定估计量的概率分布

$$\sigma_{\hat{d}} \approx \sqrt{\frac{error_{S_1}(h_1)(1-error_{S_1}(h_1))}{n_1} + \frac{error_{S_2}(h_2)(1-error_{S_2}(h_2))}{n_2}}$$

- 寻找区间 (L, U)，落入区间的概率为 N%

$$\hat{d} \pm z_N \sqrt{\frac{error_{S_1}(h_1)(1-error_{S_1}(h_1))}{n_1} + \frac{error_{S_2}(h_2)(1-error_{S_2}(h_2))}{n_2}}$$

Topic

学习算法比较

对于两个学习算法 L_A 与 L_B
估计:
$E_{S \subset \mathcal{D}}[error_{\mathcal{D}}(L_A(S)) - error_{\mathcal{D}}(L_B(S))]$
其中:

- $L(S)$ 代表给定训练样本 S 时学习算法 L 输出的假设
- 下标 $S \subset \mathcal{D}$ 表示期望值是在基准分布 $\mathcal{D}$ 中抽取的样本 S 上计算。

上述表达式描述的是学习算法 L_A 和 L_B 的差的期望值。

计算

- 对于有限的样本 D_0，要估计上述的量需要将 D_0 分割成训练集合 S_0 和不相交的测试集合 T_0。
  - 训练数据可以用来既训练 L_A 又训练 L_B，
  - 而测试数据则用来比较两个学习到的假设的准确度。

$$error_{T_0}(L_A(S_0)) - error_{T_0}(L_B(S_0))$$

估计两学习算法 L_A 和 L_B 错误率差异的一种方法

- 将可用数据 D_0 分割成 k 个相同大小的不相交子集 $T_1, T_2, \ldots, T_k$，其大小至少为 30。
- For i from 1 to k, do 使用 T_i 作为测试集合，而剩余的数据作为训练集合 Si
  - $S_i \leftarrow \{D_0 - T_i\}$
  - $h_A \leftarrow L_A(S_i)$
  - $h_B \leftarrow L_B(S_i)$
  - $\delta_i \leftarrow error_{T_i}(h_A) - error_{T_i}(h_B)$
- 返回值 $\bar{\delta}$，$\bar{\delta} \equiv \frac{1}{k}\sum_{i=1}^{k}\delta_i$

更合适的说法

算法对

$$E_{S \subset D_0}[error_{\mathcal{D}}(L_A(S)) - error_{\mathcal{D}}(L_B(S))]$$

进行了估计。而不是

$$E_{S \subset \mathcal{D}}[error_{\mathcal{D}}(L_A(S)) - error_{\mathcal{D}}(L_B(S))]$$

Topic

成对 t 检验

比较 h_A 与 h_B

- 将可用数据 D_0 分割成 k 个相同大小的不相交子集 $T_1, T_2, \ldots, T_k$，其大小至少为 30。
- For i from 1 to k, do
  - $\delta_i \leftarrow error_{T_i}(h_A) - error_{T_i}(h_B)$
- 返回值 $\bar{\delta}$, $\bar{\delta} \equiv \frac{1}{k}\sum_{i=1}^{k}\delta_i$

d 的 $N\%$ 置信区间:

- $\bar{\delta} \pm t_{N,k-1}\ s_{\bar{\delta}}$
- $s_{\bar{\delta}} \equiv \sqrt{\frac{1}{k(k-1)}\sum_{i=1}^{k}(\delta_i - \bar{\delta})^2}$

其中：

- δ_i 近似正态分布