机器学习

Machine Learning

主讲人：张敏 清华大学长聘副教授

ML 机器学习

MACHINE LEARNING-MIN ZHANG

Unit.05

贝叶斯学习（II）

*图片均来自网络或已发表刊物

回顾

- 贝叶斯定理
  - 用先验概率来推断后验概率
- Max A Posterior, MAP, h_{MAP}，极大后验假设
- Maximum Likelihood, ML, h_{ML}，极大似然假设

$$P(h \mid D) = \frac{P(D \mid h)P(h)}{P(D)}$$

$$P(h \mid D) = \frac{P(D \mid h)P(h)}{P(D)}$$

回顾：ML 举例 —— 抛硬币问题

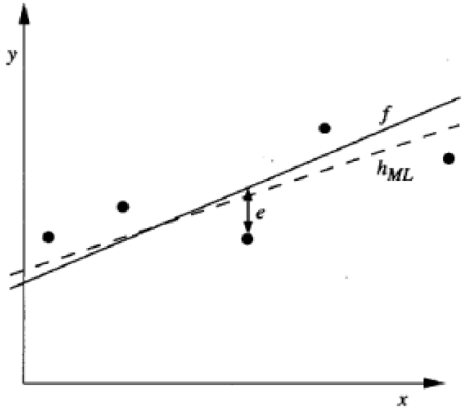
- “简单” 估计: ML (maximum likelihood，极大似然)
  - 抛一个(p,1-p)硬币 m 次，得到k 次 H 和 m-k 次 T

$$
\begin{aligned}
\log L(D \mid p) &= \log P(D \mid p) \\
&= \log(p^k (1-p)^{m-k}) \\
&= k \log p + (m-k) \log(1-p)
\end{aligned}
$$

  - 求最大值，对 p 求导令导数为 0: $\dfrac{d(\log L(D \mid p))}{dp} = \dfrac{k}{p} - \dfrac{m-k}{1-p} = 0$
  - 求解 p，得到: $p = k/m$

极大似然 & 最小二乘

- 训练数据: <x_i, d_i>
- $d_i = f(x_i) + e_i$
  - d_i : 独立的样本.
  - $f(x_i)$: 没有噪声的目标函数值
  - e_i : 噪声，独立随机变量，正态分布 $N(0, \sigma^2)$



→ d_i : 正态分布 $N(f(x_i), \sigma^2)$

极大似然 & 最小二乘

$$
\begin{aligned}
h_{ML} &= \underset{h \in H}{\operatorname{argmax}}\, p(D|h) \\
&= \underset{h \in H}{\operatorname{argmax}} \prod_{i=1}^{m} p(d_i|h) \\
&= \underset{h \in H}{\operatorname{argmax}} \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{d_i - h(x_i)}{\sigma}\right)^2} \\
&= \underset{h \in H}{\operatorname{argmax}} \sum_{i=1}^{m} \ln \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2}\left(\frac{d_i - h(x_i)}{\sigma}\right)^2
\end{aligned}
$$

独立样本

正态分布

*ln*函数的单调性

极大似然 & 最小二乘

常量

$$
\begin{aligned}
\underline{h_{ML}} &= \underset{h\in H}{\operatorname{argmax}} \sum_{i=1}^{m} \ln\frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2}\left(\frac{d_i - h(x_i)}{\sigma}\right)^2 \\
&= \underset{h\in H}{\operatorname{argmax}} \sum_{i=1}^{m} -\frac{1}{2}\left(\frac{d_i - h(x_i)}{\sigma}\right)^2 \\
&= \underset{h\in H}{\operatorname{argmax}} \sum_{i=1}^{m} -\left(d_i - h(x_i)\right)^2 \\
&= \underline{\underset{h\in H}{\operatorname{argmin}} \sum_{i=1}^{m} \left(d_i - h(x_i)\right)^2}
\end{aligned}
$$

- 独立随机变量，正态分布噪声 $N(0, \sigma^2)$, $h_{ML} = h_{LSE}$

概览

- 贝叶斯定理
  - 用先验概率来推断后验概率
- Max A Posterior, MAP, h_{MAP}， 极大后验假设
- Maximum Likelihood, ML, h_{ML}, 极大似然假设
  - ML vs. LSE (最小二乘，Least Square Error)
- Naïve Bayes, NB, 朴素贝叶斯

$$P(h \mid D) = \frac{P(D \mid h) P(h)}{P(D)}$$

Naïve Bayesian Classifier (朴素贝叶斯分类器)

- *假设目标函数* $f: X \rightarrow V$，其中每个样本 $x = (a_1, a_2, \ldots, a_n)$. 那么最有可能的 $f(x)$ 的值是:

$$v_{\text{MAP}} = \underset{v_j \in V}{argmax}\, \text{P}(x|v_j)\text{P}(v_j)$$

- 朴素贝叶斯假设:

每个属性独立

$$\text{P}(x|v_j) = \text{P}(a_1, a_2 \cdots a_n|v_j) = \prod_i \text{P}(a_i|v_j)$$

- 朴素贝叶斯分类器:

$$v_{\text{NB}} = \underset{v_j \in V}{argmax}\, \text{P}(v_j) \prod_i \text{P}(a_i|v_j)$$
$$= \underset{v_j \in V}{argmax} \{\log \text{P}(v_j) + \sum_i \log \text{P}(a_i|v_j)\}$$

如果满足属性之间的独立性，那么 $v_{MAP} = v_{NB}$

举例1：词义消歧 (Word Sense Disambiguation)

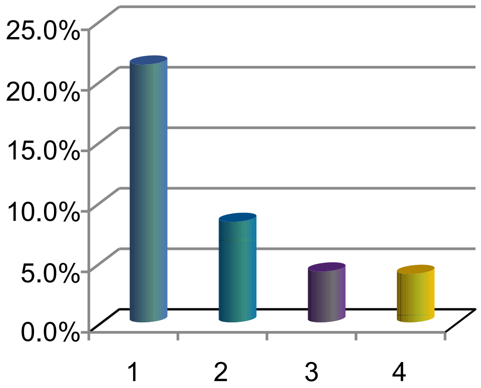
- e.g. fly =? bank = ?
- 对于单词 w，使用上下文 c 进行词义消歧
  - e.g. A fly flies into the kitchen while he fry the chicken. （他在炸鸡时一只苍蝇飞进了厨房）
  - 上下文 c: 在词 w 周围的一组词w_i(即：特征 / 属性)
  - s_i: 词 w 的第 i^{th} 个含义(即：输出标签)
- 朴素贝叶斯假设: $$P(c|s_k) = \prod_{w_i \in c} P(w_i|s_k)$$
- 朴素贝叶斯选择:

$$s = \underset{s_k}{argmax}\{\log P(s_k) + \sum_{w_i \in c} \log P(w_i|s_k)\}$$

其中: $$P(s_k) = \frac{C(s_k)}{C(w)} \qquad P(w_i|s_k) = \frac{C(w_i, s_k)}{C(s_k)}$$

举例 2: 垃圾邮件过滤

- **垃圾邮件量:** 900亿/天， 80% 来自 <200 发送者
- 第四季度主要垃圾邮件来源 (数据来自 Sophos)
  - 美国 (21.3% 垃圾信息来源， 较28.4%有所下降)
  - 俄罗斯 (8.3%, 较 4.4% 有上升)
  - 中国 (4.2%, 较 4.9% 有下降)
  - 巴西 (4.0%, 较 3.7% 有上升)



垃圾邮件过滤问题中人们学到的经验：

- 不要武断地忽略任何信息
  - E.g. 邮件头信息
- 不同的代价: **假阳性** v.s. **假阴性**
- 一个非常好的参考报告: http://www.paulgraham.com/better.html

从垃圾邮件过滤中学到的经验

(根据报告:)

早期关于贝叶斯垃圾邮件过滤的论文有两篇，于1998年发表在同一个会议

1) 作者是 Pantel 和 Lin; 2) Microsoft 研究院的一个小组

Pantel 和 Lin的过滤方法效果更好

但它只能捕捉92%的垃圾邮件，且有1.16% 假阳性错误

文章作者实现了一个贝叶斯垃圾邮件过滤器

它 能捕捉 99.5%的垃圾邮件 且 假阳性错误低于0.03%

从垃圾邮件过滤中学到的经验(续)

Subject*FREE 0.9999

Subject*free 0.9782,

free 0.6546

free!! 0.9999

(根据报告)

- 5 处不同

1.他们训练过滤器**的数据非常少**:

  - 160 垃圾邮件和466非垃圾邮件

2.最重要的一个不同可能是**他们忽略了邮件头**

3.Pantel 和 Lin **对词进行了stemming (词干化)** —— 做法有些草率了

4.计算概率的方式不同。**他们使用了全部的词，但作者只用了最显著的15个词**

5.他们**没有对假阳性做偏置**。而作者考虑了：**对非垃圾邮件中出现的词频翻倍**

概览

- 贝叶斯定理
  - 用先验概率来推断后验概率

$$P(h|D)=\frac{P(D|h)P(h)}{P(D)}$$

- Max A Posterior, MAP, h_{MAP} ，极大后验假设
- Maximum Likelihood, ML, h_{ML}，极大似然假设
  - ML vs. LSE (最小二乘，Least Square Error)
- Naïve Bayes, NB, 朴素贝叶斯
  - 独立属性/特征假设
  - NB vs. MAP

MDL (最小描述长度，Minimum Description Length)

- 奥卡姆剃刀:
  - 偏向于最短的假设
- MDL：
  - 偏向假设 *h* 使得最小化：

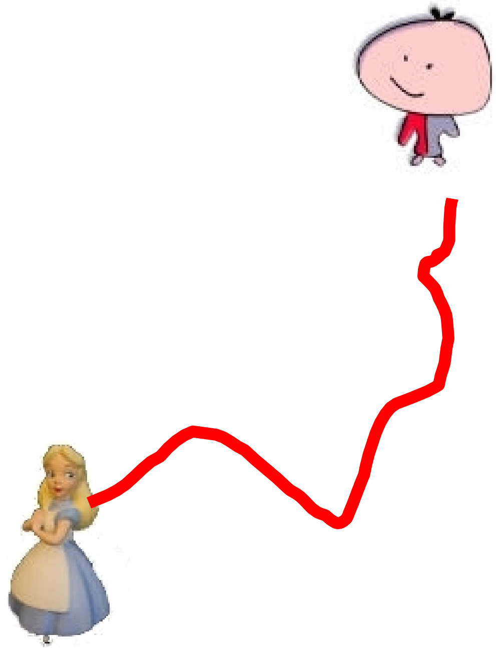
$$h_{\text{MDL}} = \underset{h \in \mathcal{H}}{argmin}\{L_{C_1}(h) + L_{C_2}(D|h)\}$$

其中 $L_C(x)$ 是 x在编码C下的 描述长度

MDL解释（基于信息理论）

- 为随机发送的信息所设计的编码
  - 遇到消息 i 的概率是 p_i
- 所需的最短编码(最小期望传输位数)是什么？
  - 为可能性较大的消息赋予较短的编码

Example: BABABABADABACAABAACABDAAAAABAAAAAAAADBCA
Binary coding: A→00, B→01, C→10, D→11, then the code:
0100010001000100110001001000000010000010000011100000000000100000000000
0000011011000,
A shorter code: Let A→0, B→10, C→110, D→111, then the code becomes
1001001001001110100110001000110010111000



- 最优编码对消息i 的编码长度为 $-\log_2 p$ 比特 [Shannon & Weaver 1949]

MDL 和 MAP

$-\log_2 p(h)$: 假设空间H最优编码下，h 的长度

$-\log_2 p(D|h)$: 最优编码下，给定h 时D 的描述长度

$$
\begin{aligned}
h_{\text{MAP}} &= \underset{h \in \mathcal{H}}{argmax}\, \mathrm{P}(D|h)\mathrm{P}(h) \\
&= \underset{h \in \mathcal{H}}{argmax}\{\log_2 \mathrm{P}(D|h) + \log_2 \mathrm{P}(h)\} \\
&= \underset{h \in \mathcal{H}}{argmin}\{\underbrace{-\log_2 \mathrm{P}(D|h)}_{L_{C2}(D|h)} \underbrace{-\log_2 \mathrm{P}(h)}_{L_{C1}(h)}\} \\
&= h_{\text{MDL}}
\end{aligned}
$$

对MDL的另一个解释

$$h_{\mathsf{MDL}} = \underset{h \in \mathcal{H}}{argmin}\{L_{C_1}(h) + L_{C_2}(D|h)\}$$

- h的长度 和 给定h编码数据的代价
  - 假设实例的序列以及编码规则对发送者和接收者来说都是已知的
  - 没有分类错误: 除h外不需要传输额外的信息
  - 如果 h 错误分类了某些样本，则需要传输:
  - 1. 哪个实例出错了?
    -- 最多 $\log_2 m$ (m: 实例的个数)
  - 2. 正确的分类结果是什么?
    -- 最多 $\log_2 k$ (k: 类别的个数)

对MDL的解释

$$h_{\mathrm{MDL}} = \underset{h \in \mathcal{H}}{argmin}\{L_{C_1}(h) + L_{C_2}(D|h)\}$$

- 权衡: 假设的复杂程度 vs. 由假设造成的错误数
  - 更偏好 一个短的且错误更少的假设
    而不是一个长的但完美分类训练数据的假设

应对*过拟合* 问题

总结

- 贝叶斯定理
  - 用先验概率来推断后验概率

$$P(h \mid D) = \frac{P(D \mid h)P(h)}{P(D)}$$

- Max A Posterior, MAP, h_{MAP} (极大后验假设)
- Maximum Likelihood, ML, h_{ML} (极大似然假设)
  - ML vs. LSE (最小二乘，Least Square Error)
- Naïve Bayes, NB, 朴素贝叶斯
  - 独立性假设
  - NB vs. MAP
- Maximum description length, MDL (最小描述长度)
  - 权衡: 假设复杂度 vs. 假设带来的错误
  - MDL vs. MAP