2.1 概率论基本概念

2.1.1 概率

$P(A) \geq 0$ $P(\Omega) = 1$ $P(\bigcup_{i=0}^{\infty} A_i) = \sum_{i=0}^{\infty} P(A_i)$

2.1.2 最大似然估计

用频率作为概率的估计值

样本空间 $\{s_1, s_2, \ldots, s_n\}$

样本 s_k $(1 \leq k \leq n)$ 的次数是 $n_N(s_k)$，那么 s_k 在这 N 次试验中的

相对频率为 $q_N(s_k) = \frac{n_N(s_k)}{N}$

N 越来越大时，$\lim_{N \to \infty} q_N(s_k) = P(s_k)$

2.1.3 条件概率

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

一般形式为 $P(A_1 \cap A_2 \cap \cdots A_n) = P(A_1) P(A_2|A_1) \cancel{P(A_3)P(A_1)} P(A_3|A_1 \cap A_2) \cdots P(A_n | \bigcap_{i=1}^{n-1} A_i)$

2.1.4 贝叶斯法则

$$P(B|A) = \frac{P(B \cap A)}{P(A)} = \frac{P(A|B)P(B)}{P(A)}$$

将 $P(A)$ 看作普通常量

$$\arg\max_B \frac{P(A|B)P(B)}{P(A)} = \arg\max_B P(A|B)P(B)$$

argmax 代表求使后面的值最大的参数

事件A的概率计算公式

$$P(A \cap B) = P(A|B)P(B)$$
$$P(A \cap \bar{B}) = P(A|\bar{B})P(\bar{B})$$

因此 $P(A) = P(A \cap B) + P(A \cap \bar{B}) = P(A|B)P(B) + P(A|\bar{B})P(\bar{B})$

推广：假设B是样本空间Ω的一个划分，即 $\bigcup B_i = \Omega$

如果 $A \subseteq \bigcup B_i$，并且 B_i 互不相交，那么 $A = \bigcup_i B_i A$

即 $P(A) = \sum_i P(B_i A)$

即 $P(A) = \sum P(A|B_i) P(B_i)$　　全概率公式

写出贝叶斯法则的精确描述:

假设A为样本空间Ω的事件，$B_1, B_2, \ldots, B_n$ 为Ω的一个划分，如果 $A \subseteq \bigcup_{i=1}^{n} B_i$，$P(A)>0$ 并且 $i \neq j$，$B_i \cap B_j = \emptyset$，$P(B_i) > 0$ $(i=1,2,\ldots,n)$ 则

$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{P(A)} = \frac{P(A|B_j)P(B_j)}{\sum_{i=1}^{n} P(A|B_i)P(B_i)}$$

例 2-1. G: 确实使用　T: 判断使用

求 $P(G|T)$　已知: $P(T|G)$　$P(G)$

还需知道 $P(T)$

$$P(T) = P(T|G) + P(T|\bar{G})$$

2.1.5 随机变量

设X为一离散型随机变量，其全部可能的值为 $\{a_1, a_2, \ldots\}$，那么

$p_i = P(X=a_i)$，$i=1,2,\ldots$ 称为X的概率函数

分布函数: $P(X \le x) = F(x)$，$-\infty < x < \infty$

分布函数 $F(x)$ 在x处的函数值表示X落在 $(-\infty, x]$ 上的概率.

2.1.6 二项式分布

$$p_i = C_n^i p^i (1-p)^{n-i} \qquad X \sim B(n,p)$$

独立伯努利试验.

[在自然语言处理中，一般以句子为处理单位，为了简化问题，通常假设一个句子的出现独立于它前面的其他语句，句子的概率分布近似地被认为符合二项式分布]

2.1.7 联合概率分布和条件概率分布

(X_1, X_2) 的联合分布为 $p_{ij} = P(X_1=a_i, X_2=b_j)$，$i=1,2,\ldots$　$j=1,2,\ldots$

考虑 X_1 在给定 $X_2=b_j$ 条件下的概率分布，其实就是求条件概率 $P(X_1=a_i|X_2=b_j)$

即 $$P(X_1=a_i|X_2=b_j) = \frac{P(X_1=a_i, X_2=b_j)}{P(X_2=b_j)} = \frac{p_{ij}}{P(X_2=b_j)} = \frac{p_{ij}}{\sum_k p_{kj}}$$

2.1.8 贝叶斯决策理论

假设所研究的分类问题有 c 个类别，各类别的状态用 w_i 表示

先验概率：$P(w_i)$ （例如：训练的表达）

在特征空间已观察到某一向量 $x = [x_1, x_2, \cdots, x_d]^T$ 是 d 维特征空间上的某一点

条件概率密度函数 $P(x|w_i)$

可得后验概率 $P(w_i|x)$ 如下：

$$P(w_i|x) = \frac{P(x|w_i)P(w_i)}{\sum_{j=1}^{c} P(x|w_j)P(w_j)}$$

基于最小错误率的贝叶斯决策规则为：

如果 $P(w_i|x) = \max_{j=1,2,\cdots,c} P(w_j|x)$，那么 $x \in w_i$

如果 $P(x|w_i)P(w_i) = \max_{j=1,2,\cdots,c} P(x|w_j)P(w_j)$，则 $x \in w_j$

如果 $c = 2$，且 $l(x) = \frac{p(x|w_1)}{p(x|w_2)} > \frac{p(w_2)}{p(w_1)}$，则 $x \in w_1$，否则 $x \in w_2$

其中 $l(x)$ 称为似然比，而 $\frac{p(w_2)}{p(w_1)}$ 称为似然比阈值

2.1.9 期望和方差

$$E(X) = \sum_{k=1}^{\infty} x_k p_k$$

$$Var(X) = E((X - E(X))^2) = E(X^2) - E^2(X)$$