lec13.3 Zero Probabilities and Laplace Smoothing

零概率问题

零概率问题是指在概率估计中，如果某个特征值从未在某个类中出现，那么对应的条件概率估计将为零。这会导致整体概率计算结果为零，从而使得这个类的预测结果被完全忽略

Example: predicting whether to play or not (zero probabilities)

Outlook			Temperature			Humidity			Windy			Play	
	Yes	*No*		*Yes*	*No*		*Yes*	*No*		*Yes*	*No*	*Yes*	*No*
Sunny	2	3	Hot	2	2	High	3	4	False	6	2	9	5
Overcast	4	0	Mild	4	2	Normal	6	1	True	3	3		
Rainy	3	2	Cool	3	1								
Sunny	2/9	3/5	Hot	2/9	2/5	High	3/9	4/5	False	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	Mild	4/9	2/5	Normal	6/9	1/5	True	3/9	3/5		
Rainy	3/9	2/5	Cool	3/9	1/5								

Test instance $\overline{X} = (\text{Outlook} = \text{overcast}, \text{Temp} = \text{cool}, \text{Humidity} = \text{high}, \text{Windy} = \text{true})$

$$
\begin{aligned}
P(\text{Play} = \text{no} \mid \overline{X}) &\propto P(\overline{X} \mid \text{Play} = \text{no}) P(\text{Play} = \text{no}) \\
&= P(\text{Outlook} = \text{overcast} \mid \text{Play} = \text{no}) \times P(\text{Temp} = \text{cool} \mid \text{Play} = \text{no}) \\
&\quad \times P(\text{Humidity} = \text{high} \mid \text{Play} = \text{no}) \times P(\text{Windy} = \text{true} \mid \text{Play} = \text{no}) \times P(\text{Play} = \text{no}) \\
&= 0 \times 1/5 \times 4/5 \times 3/5 \times 5/14 = 0
\end{aligned}
$$

拉普拉斯平滑

拉普拉斯平滑（Laplace smoothing）是一种解决零概率问题的技术。通过在计算概率时增加一个小的常数（通常是1），来避免出现零概率。具体的公式为：

$$P(x_i = a \mid C = c) = \frac{n(a, c) + 1}{N(c) + m_i}$$

其中：

- $n(a, c)$ 是在类 c 中具有值 a 的特征 x_i 的训练对象数目
- $N(c)$ 是类 c 中总训练对象数目
- m_i 是特征 x_i 的可能取值的数量

则

- 在应用拉普拉斯平滑之前，如果特征值从未在类中出现，其概率为零
- 应用拉普拉斯平滑之后，通过增加1，使得概率变得非零，并重新计算总概率

例子

假设有一个类 c，其训练对象总数为 9，特征 x_i 有 5 个可能的取值。拉普拉斯平滑前后的概率计算如下：

平滑前：

$$
\begin{aligned}
P(x_1 = a_1 \mid C = c) &= \frac{3}{9} \\
P(x_2 = a_2 \mid C = c) &= \frac{1}{9} \\
P(x_3 = a_3 \mid C = c) &= \frac{0}{9} \\
P(x_4 = a_4 \mid C = c) &= \frac{2}{9} \\
P(x_5 = a_5 \mid C = c) &= \frac{3}{9}
\end{aligned}
$$

平滑后：

$$
\begin{aligned}
P(x_1 = a_1 \mid C = c) &= \frac{3+1}{9+5} \\
P(x_2 = a_2 \mid C = c) &= \frac{1+1}{9+5} \\
P(x_3 = a_3 \mid C = c) &= \frac{0+1}{9+5} \\
P(x_4 = a_4 \mid C = c) &= \frac{2+1}{9+5} \\
P(x_5 = a_5 \mid C = c) &= \frac{3+1}{9+5}
\end{aligned}
$$