Zero probabilities and Laplace smoothing

Procheta Sen

UNIVERSITY OF LIVERPOOL

Example: predicting whether to play or not

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 $\bar{X} = (\text{Outlook} = \text{sunny}, \text{Temp} = \text{cool}, \text{Humidity} = \text{high}, \text{Windy} = \text{true})$

$$P(\text{Play} = \text{no} \,|\, \bar{X}) \propto P(\bar{X} \,|\, \text{Play} = \text{no}) P(\text{Play} = \text{no})$$
$$= P(\text{Outlook} = \text{sunny} \,|\, \text{Play} = \text{no}) \times P(\text{Temp} = \text{cool} \,|\, \text{Play} = \text{no})$$
$$\times P(\text{Humidity} = \text{high} \,|\, \text{Play} = \text{no}) \times P(\text{Windy} = \text{true} \,|\, \text{Play} = \text{no}) \times P(\text{Play} = \text{no})$$
$$= 3/5 \times 1/5 \times 4/5 \times 3/5 \times 5/14 = 0.020$$

Probability estimations

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

where

$n(a, c)$ is the number of training objects in class c with $x_i = a$,

$N(c)$ is the total number of training objects in class c

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

$$P(\text{Play} = \text{no} \,|\, \overline{X}) \propto P(\overline{X} \,|\, \text{Play} = \text{no}) P(\text{Play} = \text{no})$$

$$= P(\text{Outlook} = \text{overcast} \,|\, \text{Play} = \text{no}) \times P(\text{Temp} = \text{cool} \,|\, \text{Play} = \text{no})$$

$$\times P(\text{Humidity} = \text{high} \,|\, \text{Play} = \text{no}) \times P(\text{Windy} = \text{true} \,|\, \text{Play} = \text{no}) \times P(\text{Play} = \text{no})$$

$$= 0 \times 1/5 \times 4/5 \times 3/5 \times 5/14 = 0$$

Zero probabilities

- Issue: If the feature value a_i does not co-occur with a class value c, then the corresponding probability estimation will be 0.
- For example: Given $\text{Outlook} = \text{overcast}$, the probability of $\text{Play} = \text{no}$ is $0/5$. The other features will be ignored as the final result will be multiplied by 0.
- This is bad for our 4 feature dataset, but terrible for (say) a 1000 feature dataset.
- In text classification, we often encounter situations where a feature does not occur in a particular class.

Laplace smoothing

- We can “borrow” some probabilities from high probability features and distribute them among zero probability features to avoid having feature with zero probabilities
- This is called ***smoothing***
- There are numerous smoothing techniques based on different policies. As long as the total probability mass remains unchanged any policy of probability reassignment is valid.
- A popular method is called **Laplace smoothing**. For feature x_i and class c the probability estimates will be updated as follows:

$P(x_i = a \mid C = c) = \dfrac{n(a, c) + 1}{N(c) + m_i}$ for every value a of feature x_i,

where $n(a, c)$ is the number of training objects in class c with $x_i = a$,

$N(c)$ is the total number of training objects in class c, and

m_i is the number of possible values of feature x_i.

Laplace smoothing: before

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

$n(a_1, c)$ $n(a_2, c)$ $n(a_3, c)$ $n(a_4, c)$ $n(a_5, c)$ $N(c) = 9$

$P(x_i = a_j \mid C = c)$: $\frac{1}{9}$ $\frac{3}{9}$ $\frac{0}{9}$ $\frac{2}{9}$ $\frac{3}{9}$

Laplace smoothing: after

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

$n(a_1, c)$ $n(a_2, c)$ $n(a_3, c)$ $n(a_4, c)$ $n(a_5, c)$ $N(c) = 9 + 5 = 14$

$P(a_i \mid C = c)$: $\frac{1+1}{9+5}$ $\frac{3+1}{9+5}$ $\frac{0+1}{9+5}$ $\frac{2+1}{9+5}$ $\frac{3+1}{9+5}$