# Домашнее задание №8

Асланов Андрей, ИУ9-11М

Даны документы и их классы

$D1 = (X1, X2, X3)$ C1
$D2 = (X1, X2, X4)$ C1
$D3 = (X4, X5, X6)$ C2

Определить класс документа на основе Байесовского классификатора
$D4 = (X1, X4, X5)$

Использовать 2 способа: MultiNomial и Bernulli

## Multinomial

В многомерной модели наивного Байесовского классификатора документ $D_i$ – это вектор признаков $X_i$, показывающих, встретилось ли в том или ином документе слово.

$$C_{NB} = \underset{c_j \in C}{\operatorname{argmax}} \left\{ P(C_j) \prod_{i=1}^{n} P(X_i, C_j) \right\}$$ ← общая формула для Байесовского классификатора

$$C_{NB} = \operatorname{argmax} \{ P(D4|C1)\,P(C1),\ P(D4|C2)\,P(C2) \}$$

$$C_{NB} = \operatorname{argmax} \{ \underbrace{P(X1|C1)\,P(X4|C1)\,P(X5|C1)}_{P(D4|C1)}\,P(C1),\ \underbrace{P(X1|C2)\,P(X4|C2)\,P(X5|C2)}_{P(D4|C2)}\,P(C2) \}$$

$$P(X_i, C_j) = \frac{N(X_i, C_j) + 1}{N(C_i) + k}$$ ← с учётом сглаживания по Лапласу

- $N(C_i)$ – общее число неуникальных признаков, принадлежащих классу $C_i$
- $k$ – суммарное число признаков обучающей выборки
- $N(X_i, C_j)$ – Сколько раз встречался признак $X_i$ при условии класса $C_j$

Multinomial

| Признак/(слово) документа | $P(x_i \mid C_1)$ | $P(x_i \mid C_2)$ |
|---|---|---|
| $X_1$ | $\frac{2+1}{6+6} = \frac{1}{4}$ | $\frac{0+1}{3+6} = \frac{1}{9}$ |
| $X_2$ | $\frac{2+1}{6+6} = \frac{1}{4}$ | $\frac{0+1}{3+6} = \frac{1}{9}$ |
| $X_3$ | $\frac{1+1}{6+6} = \frac{1}{6}$ | $\frac{0+1}{3+6} = \frac{1}{9}$ |
| $X_4$ | $\frac{1+1}{6+6} = \frac{1}{6}$ | $\frac{1+1}{3+6} = \frac{2}{9}$ |
| $X_5$ | $\frac{0+1}{6+6} = \frac{1}{12}$ | $\frac{1+1}{3+6} = \frac{2}{9}$ |
| $X_6$ | $\frac{0+1}{6+6} = \frac{1}{12}$ | $\frac{1+1}{3+6} = \frac{2}{9}$ |

$$C_{NB} = \operatorname{argmax}\left\{ \frac{1}{4} \cdot \frac{1}{6} \cdot \frac{1}{12} \cdot \frac{2}{3} \;,\; \frac{1}{9} \cdot \frac{2}{9} \cdot \frac{2}{9} \cdot \frac{1}{3} \right\}$$

$$C_{NB} = \operatorname{argmax}\{ \underset{\parallel}{0,00231} ; \; \underset{\parallel}{0,001828} \} \Rightarrow C_1 .$$

$P$(Д4 - это класс $C_1$) $\quad$ $P$(Д4 - это класс $C_2$)

Bernulli:

$N(X_i, C_j) + 1$ → Потому что по Th. Байеса предполагаем, что мы как бы сделали +2 лишних испытания, одно из которых удачно

$N(C_i) + 2$ → сколько раз встречался признак $X_i$ при условии класса $C_j$

$N(C_i)$ → общее число признаков, принадлежащих классу $C_i$ (неуникальных)

| Признак (слово) документа | | $P(X_i \mid C_1)$ | $P(X_i \mid C_2)$ |
|---|---|---|---|
| X1 | | $\frac{2+1}{6+2} = \frac{3}{8}$ | $\frac{0+1}{3+2} = \frac{1}{5}$ |
| X2 | | $\frac{2+1}{6+2} = \frac{3}{8}$ | $\frac{0+1}{3+2} = \frac{1}{5}$ |
| X3 | | $\frac{1+1}{6+2} = \frac{1}{4}$ | $\frac{0+1}{3+2} = \frac{1}{5}$ |
| X4 | | $\frac{1+1}{6+2} = \frac{1}{4}$ | $\frac{1+1}{3+2} = \frac{2}{5}$ |
| X5 | | $\frac{0+1}{6+2} = \frac{1}{8}$ | $\frac{1+1}{3+2} = \frac{2}{5}$ |
| X6 | | $\frac{0+1}{6+2} = \frac{1}{8}$ | $\frac{1+1}{3+2} = \frac{2}{5}$ |

$$C_{NB} = \operatorname{argmax}\{ P(X1|C1)\,P(X4|C1)\,P(X5|C1)\cdot P(C1),\ P(X1|C2)P(X4|C2)P(X5|C2)P(C2)\}$$

$$C_{NB} = \operatorname{argmax}\left\{ \frac{2}{3}\cdot\left(\frac{3}{8}\cdot\left(1-\frac{3}{8}\right)\left(1-\frac{1}{4}\right)\cdot\frac{1}{4}\cdot\frac{1}{8}\cdot\left(1-\frac{1}{8}\right)\right),\right.$$

$$\left.\frac{1}{3}\cdot\left(\frac{1}{5}\cdot\left(1-\frac{1}{5}\right)\left(1-\frac{1}{5}\right)\cdot\frac{2}{5}\cdot\frac{2}{5}\cdot\left(1-\frac{2}{5}\right)\right)\right\}$$

$$C_{NB} = \operatorname{argmax}\left\{ \frac{2}{3}\cdot\frac{3}{8}\cdot\frac{5}{8}\cdot\frac{3}{4}\cdot\frac{1}{4}\cdot\frac{1}{8}\cdot\frac{7}{8},\ \frac{1}{3}\cdot\frac{1}{5}\cdot\frac{4}{5}\cdot\frac{4}{5}\cdot\frac{2}{5}\cdot\frac{2}{5}\cdot\frac{3}{5}\right\}$$

$$C_{NB} = \operatorname{argmax}\left\{ \frac{630}{196608},\ \frac{192}{46875}\right\}$$

$$C_{NB} = \operatorname{argmax}\{ 0{,}003204;\ 0{,}004096\} \Rightarrow \text{более вероятен класс } C_2.$$

Вывод: по многокомпонентному Байесовскому классификатору более вероятен класс $C_1$, а по методу Бернулли – класс $C_2$.