Homework - Text Class by NB

giovedì 14 aprile 2022 16:50

TEXT CLASSIFICATION BY NB *(group: 404 name not found)*

	parliament	italian	loose	win	soccer	championship	government	Class
d1	0	1	0	1	1	1	0	sport
d2	0	0	1	0	1	1	1	sport
d3	0	0	0	0	0	1	0	sport
d4	1	1	1	0	0	0	1	politcs
d5	1	1	0	1	0	0	1	politcs

1. **Feature selection by using entropy – select the first k=5 best features.**

PARLIAMENT

$E(S, parliament) = -p(sport \mid parliament) * \log_2(sport \mid parliament) - p(politics \mid parliament) * \log_2(politics \mid parliament)$
$= (-0 * 0) * (-1 * 0) = 0$

$p(sport \mid parliament) = 0$

$\log_2 p(sport \mid parliament) = 0$

$p(politics \mid parliament) = 1$

$\log_2 p(politics \mid parliament) = 0$

ITALIAN

$E(S, italian) = -p(sport \mid italian) * \log_2(sport \mid italian) - p(politics \mid italian) * \log_2(politics \mid italian) =$
*= (-0.33 * -1.60)*(-0.66 *-0.60) = 0.528 * 0.396 = 0.209088*

$p(sport \mid italian) = 0.33$

$\log_2 p(sport \mid italian) = -1.60$

$p(politics \mid italian) = 0.66$

$\log_2 p(politics \mid italian) = -0.60$

LOOSE

$E(S, loose) = -p(sport \mid loose) * \log_2(sport \mid loose) - p(politics \mid loose) * \log_2(politics \mid loose) =$
= (-0.5-1)*(-0.5*-1) = 0.5 * 0.5 = 0.25*

$p(sport \mid loose) = 0.5$

$\log_2 p(sport \mid loose) = -1$

$p(politics \mid loose) = 0.5$

$\log_2 p(politics \mid loose) = -1$

WIN

$E(S, win) = -p(sport \mid win) * \log_2(sport \mid win) - p(politics \mid win) * \log_2(politics \mid win) =$
= (-0.5-1)*(-0.5*-1) = 0.5 * 0.5 = 0.25*

$p(sport \mid win) = 0.5$

$\log_2 p(sport \mid win) = -1$

$p(politics \mid win) = 0.5$

$\log_2 p(politics \mid win) = -1$

SOCCER

$E(S, soccer) = -p(sport \mid soccer) * \log_2(sport \mid soccer) - p(politics \mid soccer) * \log_2(politics \mid soccer) =$
*= (-1*0)*(0*0) = 0*

$p(sport \mid soccer) = 1$

$\log_2 p(sport \mid soccer) = 0$

$p(politics \mid soccer) = 0$

$\log_2 p(politics \mid soccer) = 0$

CHAMPIONSHIP

$E(S, championship) = -p(sport \mid championship) * \log_2(sport \mid championship) - p(politics \mid championship) * \log_2(politics \mid championship) =$
*= (-1*0)*(0*0) = 0*

$p(sport \mid championship) = 1$

$\log_2 p(sport \mid championship) = 0$

$p(politics \mid championship) = 0$

$\log_2 p(politics \mid championship) = 0$

GOVERNMENT
$E(S, government) = -p(sport \mid government) * \log_2(sport \mid government) \ - p(politics \mid government) * \log_2(politics \mid government) =$ $= (0*0)*(-0.66*-0.60) = 0*0.396 = 0$
$p(sport \mid government) = 0$
$\log_2 p(sport \mid government) = 0$
$p(politics \mid government) = 0.66$
$\log_2 p(politics \mid government) = -0.60$

So, the first k=5 best features are:

i. *Parliament*
ii. *Soccer*
iii. *Championship*
iv. *Government*
v. *italian*

2. Prior probabilities estimates, over the (reduced) training set, by using both fractions and m-estimates.

$P_{c1} = \frac{Wc1}{|Voc|} = \frac{3}{5} = 0.6$
$P_{c2} = \frac{Wc2}{|Voc|} = \frac{3}{5} = 0.6$

$P(w \mid c) = p(parliament|sport) = \frac{d_{w1,c1}+W_{c1}}{d_{c1}+|Voc|} = \frac{0+3}{3+5} = 0.375$
$P(w \mid c) = p(parliament|politics) = \frac{d_{w1,c2}+W_{c2}}{d_{c2}+|Voc|} = \frac{2+3}{3+5} = 0.625$

$P(w \mid c) = p(soccer|sport) = \frac{d_{w2,c1}+W_{c1}}{d_{c1}+|Voc|} = \frac{2+3}{3+5} = 0.625$
$P(w \mid c) = p(soccer|politics) = \frac{d_{w2,c2}+W_{c2}}{d_{c2}+|Voc|} = \frac{0+3}{3+5} = 0.375$

$P(w \mid c) = p(championship|sport) = \frac{d_{w3,c1}+W_{c1}}{d_{c1}+|Voc|} = \frac{3+3}{3+5} = 0.75$
$P(w \mid c) = p(championship|politics) = \frac{d_{w3,c2}+W_{c2}}{d_{c2}+|Voc|} = \frac{0+3}{3+5} = 0.375$

$P(w \mid c) = p(government|sport) = \frac{d_{w4,c1}+W_{c1}}{d_{c1}+|Voc|} = \frac{1+3}{3+5} = 0.5$
$P(w \mid c) = p(government|politics) = \frac{d_{w4,c2}+W_{c2}}{d_{c2}+|Voc|} = \frac{2+3}{3+5} = 0.625$

$P(w \mid c) = p(italian|sport) = \frac{d_{w5,c1}+W_{c1}}{d_{c1}+|Voc|} = \frac{1+3}{3+5} = 0.5$
$P(w \mid c) = p(italian|politics) = \frac{d_{w5,c2}+W_{c2}}{d_{c2}+|Voc|} = \frac{2+3}{3+5} = 0.625$

3. Classification of the following docs:

- **d1 = "The team of Juventus will likely not win the Italian soccer championship".**

	Parliament	Italian	Soccer	Championship	Government	CLASS
d1	**0**	**1**	**1**	**1**	**0**	**SPORT**

*p(sport|{italian, soccer , championship}) = p(sport) * p(italian|sport) * p(soccer|sport) * p(championship|sport) = 0.6 * 0.5 * 0.625 * 0.75 =* ***0.14***

*p(politics|{italian, soccer , championship}) = p(politics) * p(italian|politics) * p(soccer|politics) * p(championship|politics) = 0.4 * 0.625 * 0.375 * 0.375 =* ***0.04***

- **d2 = "the Italian parliament, based on the initiative of the government, has approved a law for Italian soccer teams".**

	Parliament	Italian	Soccer	Championship	Government	CLASS
d2	**1**	**1**	**1**	**0**	**1**	**POLITICS**

*p(sport|{parliament , italian , soccer , government}) = p(sport) * p(parliament|sport) * p(italian|sport) * p(soccer|sport) * p(government|sport) =*
*= 0.6 * 0.375 * 0.5 * 0.625 * 0.5 =* ***0.036***

*p(politics|{parliament , italian , soccer , government}) = p(politics) * p(parliament|politics) * p(italian|politics) * p(soccer|politics) * p(government|politics) =*
*= 0.4 * 0.625 *0.625 * 0.375 * 0.625 =* ***0.037***