

②

problem on LIME for Text data.

Dataset:

$D_1$	I am <u>happy</u>	P
$D_2$	not <u>happy</u>	P
$D_3$	I am <u>sad</u>	N

Input: I am happy (no)

Perturbed Data: P (P) → (P, N)

$x_1$  I happy.

$x_2$  not sad.

$f_m$  = Naive Bayes classifier

$g_m$  = average of  $f_m$  along the label.

Calculate the probability ( $f_m$ ) for perturbed data.

$$x_1 \text{ I happy} \leq_N \frac{p_{(P|x_1)}}{p_{(N|x_1)}}$$

$$*) p(P|x_1) = p(\text{positive}) * (p(I|P) * p(happy|P))$$

From  
Dataset

$$= \frac{2}{3} * \frac{1}{2} * \frac{2}{2}$$

use laplace law to handle zero probability problem.

Given  $|B| = 15$

$$\Rightarrow \frac{2+1}{3+15} * \frac{1+1}{2+15} * \frac{2+1}{2+15}$$
$$= \frac{3}{18} * \frac{2}{17} * \frac{3}{17}$$

$$p(p|x_1) = 0.166 * 0.117 * 0.176 = 0.00341$$

\* )  $p(N|x_1) = p(\text{negative}) * p(I|N) * p(\text{happy}|N)$

$$= \frac{1+1}{3+15} * \frac{1+1}{1+15} * \frac{0+1}{1+15}$$

$$= \frac{2}{18} * \frac{2}{16} * \frac{1}{16} = 0.1111 * 0.125 * 0.0625$$

$$= 0.000867 = 8.67 \times 10^{-4}$$

Similarly calculate for  $x_2 \rightarrow \text{not sad} \rightarrow P_N$

\* )  $p(p|x_2) = p(\text{positive}) * p(\text{not } p) * p(\text{sad}|p)$

$$= \frac{2+1}{3+15} * \frac{1+1}{2+15} + \frac{0+1}{2+15}$$

$$= \frac{3}{18} * \frac{2}{17} * \frac{1}{17} = 0.1666 * 0.1176 * 0.0583$$

$$= 0.00152$$

$$*) p(N/x_2)$$

$$p(\text{Negative}) * p(\text{Not } N) * p(\text{sad } | N)$$

$$\frac{1+1}{3+15} * \frac{0+1}{1+15} * \frac{1+1}{1+15}$$

$$= \frac{2}{18} * \frac{1}{16} * \frac{2}{16} = 0.1111 * 0.0625 * 0.125$$

$$= \boxed{0.000868}$$

$f(n)$  for perturbed data.

	P	N
$x_1$	I happy.	0.0034
$x_2$	not sad	0.0015

Normalize the values.

$$\begin{array}{ll} x_1 & \frac{P_1}{P_1+n_1} \\ x_2 & \frac{n_1}{P_1+n_1} \\ & \frac{P_2}{P_2+n_2} \\ & \frac{n_2}{P_2+n_2} \end{array}$$

$$\frac{P_1}{P_1+n_1} = \frac{0.0034}{0.0034 + 0.00086} = \frac{0.0034}{0.0042} = \boxed{0.82}$$

$$\frac{n_1}{P_1+n_1} = \frac{0.00086}{0.0042} = \boxed{0.19}$$

$$\frac{P_2}{P_2+n_2} = \frac{0.00152}{0.00152 + 0.00086} = \boxed{0.652}$$

$$\frac{n_2}{P_2+n_2} = \frac{0.00086}{0.00238} = \boxed{0.373}$$

After Normalization  $f(x)$

(P) : (N)

$f(x) = 0.8 \quad 0.19$

$0.652 \quad 0.361$

II calculate  $g(x) \rightarrow$  average of  $f(x)$  along label.

(P) (N)

$g(x)$   $= 0.72 \quad 0.277$

$0.72 \quad 0.277$

$$\Rightarrow \frac{0.8 + 0.652}{2} = 0.72$$

$$\Rightarrow \frac{0.19 + 0.361}{2} = 0.277$$

III calculate weight

$$\hat{\omega} = e^{-\frac{(|x_1 - x_0|^2)}{2\sigma^2}}$$

For text, we will use jaccard similarly

$x_0$  = "am happy" (I/p).

$x_1$  = I happy (perturbed).

$$\text{jaccard similarity} = \frac{|x_0 \cap x_1|}{|x_0 \cup x_1|} \rightarrow \begin{matrix} \text{happy is similar} \\ = 1 \end{matrix}$$

(at most 4 different words)

unique.  
words = 3

$$J(x_0, x_1) = \frac{1}{3}$$

$$w_1 = e^{-\left(\frac{(1/3)^2}{2 \cdot 2}\right)} = e^{-\frac{(1/9)}{2}}$$

$$w_1 = 0.945$$

$x_0$  = am happy (I/p)

$x_2$  = not sad (perturbed).

$$J(x_0, x_2) = \frac{0}{4} \quad \begin{matrix} (\text{no similar}) \\ (\text{unique}) \end{matrix}$$

$$= e^{-\left(\frac{0^2}{2}\right)} = e^{-0} = 1$$

Finally

$x_1$	I happy	0.9459	$\begin{bmatrix} 0.8 & 0.199 \\ 0.65 & 0.361 \end{bmatrix}$	$\begin{bmatrix} 0.72 & 0.27 \\ 0.72 & 0.27 \end{bmatrix}$
$x_2$	not sad	1		

(calculate  $w$   $f(n)$   $g(n)$ )

$$L(g) = \sum w (f(n) - g(n))^2$$

$$= 0.9459 * [(0.8 - 0.72)^2 + (0.199 - 0.27)^2]$$

$$+ 1 * [(0.65 - 0.72)^2 + (0.361 - 0.27)^2]$$

$$L(g) = 0.025$$

Note:

$L(g)$  is minimum in this case.  
so we can trust the model.