

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

IIP am happy (x_0)

perturbed Data:

x_1 I happy

x_2 not sad.

$f(x)$ = Naive bayes classifier

$g(x)$ = average of $f(x)$ along the label.

I Calculate the probability ($f(x)$) for perturbed data.

$$x_1 \quad I \quad happy \quad \begin{cases} P & p(P|x_1) \\ N & p(N|x_1) \end{cases}$$

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

From
Dataset

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

Use Laplace law to handle zero probability problem.

Given $|B| = 15$

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

$$= \frac{3}{18} * \frac{2}{17} * \frac{3}{17}$$

$$p(p/x_1) = 0.166 * 0.117 * 0.176 = \boxed{0.00341}$$

$$*) p(N/x_1) = p(\text{Negative}) * p(I/N) * p(\text{happy}/N)$$

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

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

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

Similarly calculate for $x_2 \rightarrow$ not, sad $\begin{matrix} \nearrow P \\ \searrow N \end{matrix}$

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

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

$$= \frac{3}{18} * \frac{2}{17} * \frac{1}{17} = 0.1666 * 0.1176 * 0.0588$$

$$= \boxed{0.00152}$$

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

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

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

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

$$= \boxed{0.000868}$$

$f(x)$ for perturbed data.

0.0625

		P	N
x_1	I happy.	0.0034	0.00086
x_2	not sad	0.0015	0.00086

Normalize the values.

x_1	$\frac{p_1}{p_1+n_1}$	$\frac{n_1}{p_1+n_1}$
x_2	$\frac{p_2}{p_2+n_2}$	$\frac{n_2}{p_2+n_2}$

$$\frac{p_1}{p_1+n_1} = \frac{0.0034}{0.0034 + 0.00086} = \frac{0.0034}{0.0042} = \boxed{0.82}$$

$$\frac{n_1}{p_1+n_1} = \frac{0.0008}{0.0042} = \boxed{0.19}$$

$$\frac{p_2}{p_2+n_2} = \frac{0.00152}{0.00152 + 0.0008} = \boxed{0.652}$$

$$\frac{n_2}{p_2+n_2} = \frac{0.00086}{0.00232} = \boxed{0.373}$$

After Normalization $f(x)$

	(P)	(N)
$f(x)$	0.8	0.19
	0.652	0.361

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

	(P)	(N)
$g(x)$		
	0.72	0.277
	0.72	0.277

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

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

III calculate weight

$$w = e^{-\frac{(1x_1 - x_0)^2}{2 \cdot 0.2}}$$

For text, we will use jaccard similarity

$x_0 = \text{" am happy " (I/p)}.$

$x_1 = \text{I happy (perturbed)}.$

jaccard similarity = $\frac{|x \cap y|}{|x \cup y|} \rightarrow \text{happy is similar} = 1$
 $\rightarrow \text{unique. words} = 3$

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

$$w_1 = e^{-\left(\frac{(1/3)^2}{2 \cdot 0^2}\right)} = e^{-\frac{(1/9)}{2}}$$

$$w_1 = 0.945$$

$x_0 = \text{ am happy (I/p)}$

$x_2 = \text{ not sad (perturbed)}.$

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

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

Finally

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

calculate w $f(x)$ $g(x)$

(4)

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

$$= 0.9459 \left[(0.8 - 0.72)^2 + (0.199 - 0.27)^2 \right]$$

$$+ 1 \cdot \left[(0.65 - 0.72)^2 + (0.361 - 0.27)^2 \right]$$

$$L(g) = 0.025$$

Note:

$L(g)$ is minimum in this case.

so we can trust the model.