6. Naive Bayes

Naive Bayes is best suited for ML applications wherein the feature vectors $\bar{x}$ are discrete.

Example: ML-based e-mail SPAM filter.

Consider a feature vector $\bar{x}$ of size N, where N is the number of words in the English Language dictionary.

The labels $y = 0, 1$ indicate genuine, spam emails respectively.

$x_j = 1$, if the email contains the j^{th} word of the dictionary

$x_j = 0$, else

$$\bar{x} = \begin{bmatrix} \vdots \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \\ \vdots \end{bmatrix} \begin{array}{l} \vdots \\ \text{able} \\ \text{above} \\ \text{abroad} \\ \text{abort} \\ \text{access} \\ \vdots \end{array}$$

Occurs in email

Not occurs in email

Prior Probabilities

The quantities $p(x_j = v_j \mid y = u)$ are the prior probabilities.

$v_j = 1, 0$ and $y = 1, 0$

Genuine email

Spam email

How to calculate these?

Consider the availability of M training pairs $(\bar{x}(i), y(i))$

The various prior probabilities can now be calculated as follows.

Probability of j^{th} word in SPAM e-mail

$$= \frac{\text{No. of SPAM emails with } j^{th} \text{ word}}{\text{No. of SPAM emails}}$$

$$p(x_j=1|y=1) = \frac{\sum_{i=1}^{M} 1(x_j(i)=1, y(i)=1)}{\sum_{i=1}^{M} 1(y(i)=1)}$$

The various other probabilities can now be calculated as follows.

$$p(x_j=1|y=0) = \frac{\sum_{i=1}^{M} 1(x_j(i)=1, y(i)=0)}{\sum_{i=1}^{M} 1(y(i)=0)}$$

$$p(y=1) = \frac{\sum_{i=1}^{M} 1(y(i)=1)}{M}$$

Problem 1 : Words used in SPAM emails.

Word 1 = "BUMPER"

Word 2 = "OFFER"

Consider the table below.

	$x_1=0$	$x_1=1$
$y=0$	60	20
$y=1$	40	80

Evaluate the following :

(i) Probability of x_1 does not occur in genuine email

$$p(x_1=0|y=0) = \frac{60}{60+20} = 3/4$$

(ii) $$p(x_1=1|y=1) = \frac{80}{80+40} = 2/3$$

(iii) $$p(y=0) = \frac{60+20}{60+20+40+80} = \frac{80}{200} = \frac{2}{5}$$

(iv) $$p(y=1) = \frac{40+80}{60+20+40+80} = 3/5$$

(or) $1 - \frac{2}{5} = 3/5$

Also note,

$$p(x_j=0|y=1) = 1-p(x_j=1|y=1)$$

$$p(x_j=0|y=0) = 1-p(x_j=1|y=0)$$

$$\boxed{\begin{array}{l} p(y=0) = 1-p(y=1) \\ \text{(ie) Prob. of Genuine email} = 1 - \text{Prob. of SPAM email} \end{array}}$$

Naive Bayes assumption.

The different words are conditionally independent, given the label y.

$$p(\bar{x}=\bar{v}|y=u) = p(x_1=v_1,\ldots,x_N=v_N \mid y=u)$$

$$= p(x_1=v_1|y=u) \times \ldots\ldots \times p(x_N=v_N|y=u)$$

$$= \prod_{j=1}^{N} p(x_j=v_j|y=u)$$

Posterior Probabilities.

The posterior probabilities are calculated as follows.

$$p(y=1|\bar{x}=\bar{v}) = \frac{p(\bar{x}=\bar{v}|y=1) \times p(y=1)}{p(\bar{x}=\bar{v})}$$

$$p(y=0|\bar{x}=\bar{v}) = \frac{p(\bar{x}=\bar{v}|y=0) \times p(y=0)}{p(\bar{x}=\bar{v})}$$

Spam Classification.

Email is classified as SPAM if

$$p(y=1|\bar{x}=\bar{v}) > p(y=0|\bar{x}=\bar{v})$$

$$\Rightarrow \frac{p(\bar{x}=\bar{v}|y=1) \times p(y=1)}{p(\bar{x}=\bar{v})} > \frac{p(\bar{x}=\bar{v}|y=0) \times p(y=0)}{p(\bar{x}=\bar{v})}$$

$$\Rightarrow p(\bar{x}=\bar{v}|y=1) \times p(y=1) > p(\bar{x}=\bar{v}|y=0) \times p(y=0)$$

This implies, choose C_1 SPAM if

$$\underbrace{\prod_{j=1}^{N} p(x_j = v_j | y=1) \times p(y=1)}_{Q_1} > \underbrace{\prod_{j=1}^{N} p(x_j = v_j | y=0) \times p(y=0)}_{Q_0}$$

$$\boxed{\begin{array}{l} Q_1 > Q_0 \rightarrow \text{SPAM email} \\ Q_1 < Q_0 \rightarrow \text{Genuine email} \end{array}}$$

Laplace Smoothing

- Naive Bayes has a problem.
- Lets say a new word "IITK" appears in your email, which is not present in any training emails.
- Consider the index of "IITK" in the dictionary is j.

The prior probabilities are

$$p(x_j = 1 | y=1) = \frac{\sum_{i=1}^{M} \mathbb{1}(x_j(i) = 1, y(i) = 1)}{\sum_{i=1}^{M} \mathbb{1}(y(i) = 1)} = 0$$

$$p(x_j = 1 | y=0) = \frac{\sum_{i=1}^{M} \mathbb{1}(x_j(i) = 1, y(i) = 0)}{\sum_{i=1}^{M} \mathbb{1}(y(i) = 0)} = 0$$

These cause problems in computation of the Posterior probabilities.

- Therefore, we use the following prior probabilities instead.

$$p(x_j = 1 | y=1) = \frac{1 + \sum_{i=1}^{M} \mathbb{1}(x_j(i) = 1, y(i) = 1)}{2 + \sum_{i=1}^{M} \mathbb{1}(y(i) = 1)}$$

$$p(x_j = 1 | y=0) = \frac{1 + \sum_{i=1}^{M} \mathbb{1}(x_j(i) = 1, y(i) = 0)}{2 + \sum_{i=1}^{M} \mathbb{1}(y(i) = 0)}$$

- This is termed as Laplace Smoothing.

Problem 2 : Given

	$x_1 = 0$	$x_1 = 1$
$y=0$	60	20
$y=1$	40	80

Evaluate $p(x_1=0 \mid y=0)$, $p(x_1=1 \mid y=1)$ with Laplace smoothing.

(i) $p(x_1=0 \mid y=0) = \dfrac{1+60}{1+82} = \dfrac{61}{82}$

(ii) $p(x_1=1 \mid y=1) = \dfrac{1+80}{2+40+80} = \dfrac{81}{122}$