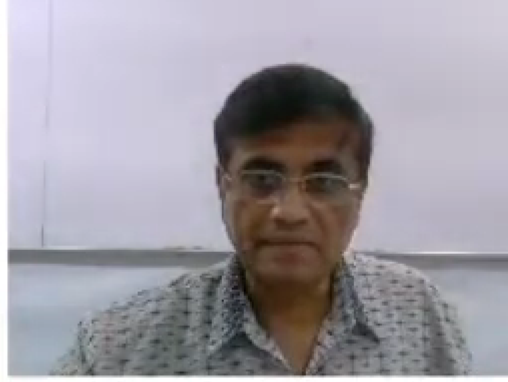
Naïve Bayes Classifier – Concept of Laplace Smoothing



- So far you know how to represent feature vector X_j
- How to create Vocabulary.
- And how to develop a model using Naïve Bayes assumption.
- Next task is to get an e mail and predict using my model whether it is a spam or non spam mail.

$P(x_j / y)$ & $P(y)$

When an unknown word appears, what will happen to my predictor?

$$P(y=1/x) = \frac{\prod_{j=1}^{n} P(x_j / y=1) P(y=1)}{\prod_{j=1}^{n} P(x_j / y=1) P(y=1) + \prod_{j=1}^{n} P(x_j / y=0) P(y=0)}$$

$$= \frac{0}{0}$$

$$X = \begin{array}{cl} \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}_{10000 \times 1} & \begin{array}{l} a \\ absen \\ abra \\ \vdots \\ lottery \\ zen \end{array} \end{array}$$

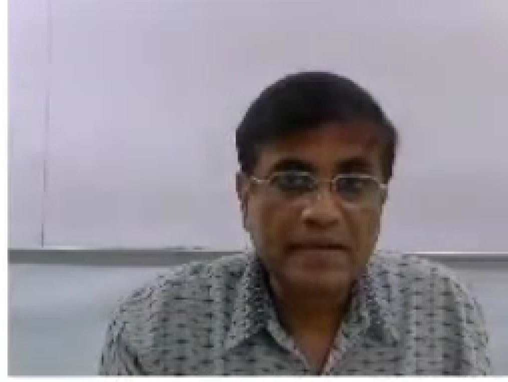
Laplace Smoothing

It is a bad idea to put $P(X_{1,00}/Y=1)=0$
How to fix it?

Take the example of Cricket tournament with following teams: India, Pakistan, Sri Lanka, Bangladesh, Australia, New Zealand, West Indies , England, Netherland. Following is the winning record of Netherland:

Date	Afghanistan's winning record	Win
15/09	Played with India	0
17/09	Played with Australia	0
19/09	Played with England	0
20/09	Played with Pakistan	0
22/09	Played with Sri Lanka	0
23/09	Going to play with Bangladesh	?

The maximum likelihood estimate of $\phi_{j/Y=1} = \frac{\#1 + 1}{\#0 + 1 + \#1 + 1} = \frac{0+1}{5+1+0+1} = \frac{1}{7}$



Laplace Smoothing

So In general the Laplace smoothing for Multivariate Bernoulli event model :

$$\phi_{j|y=1} = \frac{\sum_{i=1}^{m} 1\{x_j^{(i)} = 1 \wedge y^{(i)} = 1\} + 1}{\sum_{i=1}^{m} 1\{y^{(i)} = 1\} + 2}$$

$$\phi_{j|y=0} = \frac{\sum_{i=1}^{m} 1\{x_j^{(i)} = 0 \wedge y^{(i)} = 0\} + 1}{\sum_{i=1}^{m} 1\{y^{(i)} = 0\} + 2}$$

And for multinomial event model

$$\phi_j = \frac{\sum_{i=1}^{m} 1\{z^{(i)} = j\} + 1}{m + k},$$

Where Z is a multinomial random variable - parametrized by $\phi_j = P(Z=j)$. Given a set of m independent observations $[z^{(1)}, \dots z^{(m)}]$, the maximum likelihood estimate are given by :

$$\phi_j = \frac{\sum_{i=1}^{m} 1\{z^{(i)} = j\}}{m}.$$