4.1 a)

Lets say we have 'd' document available
And as the classification problem is binary (onion OR Economist)

$k=2$ eg. $y=1 \rightarrow$ Economist

$y=2 \rightarrow$ Onion

The $y^{(i)}$ represent the category (Onion or Economist) of ith document in collection.

$x_j^{(i)}$ corresponds to presence or absense (0|1) of particular word in ith document from the list of words in vocabulary V.

Naive Bayes algorithm basically finds the most probable class label (ONION or Economist) given $x_i (x_1 \ldots x_d)$

$$\hat{Y} = \underset{y}{\text{argmax}}\ P(Y=y|X)$$

$$= \underset{y}{\text{argmax}}\ P(y|x)$$

As $P(y|x) = \frac{P(x|y)\,P(y)}{P(x)}$

$$= \underset{y}{\text{argmax}}\ \frac{P(x|y)\,P(y)}{P(x)}$$

As $P(x)$ is just a scaling variable, we can ignore it.

$$= \underset{y}{\text{argmax}}\ P(x|y)\,P(y)$$

Now as we have V such feature vectors where each feature vector (x_i) is 1 if ith word appears in the article.

So,

$$\underset{y}{\text{argmax}}\ P(x|y)\,P(y)$$

$$\left(P(x_1|y)P(x_2|y)\ldots P(x_v|y)\right)P(y)$$

$$= \prod_{i=1}^{V}\left[P(x_i|y)\right]P(y)$$

$$= \prod_{i=1}^{V}\left[P(x_i|y)\right]P(Y=y)$$

4.1 b)

For Representation of $P(X|Y=y)$ where X is feature vector and y is associated label

In Naive Bayes classifier :-

→ Prior $P(Y) = k-1$ if we have 'k' classes

→ Liklihood $P(X|Y=y) = (2^d - 1)k$ for binary features

↓

Naive Bayes.
(No assumption)

If we hold the simplicity Assumption for NB classifier which states that given the class label Y, pair of feature X_i and X_j $(i \neq j)$ are conditionally independent.

$$P(X|Y=y) = \underbrace{P(x_1|Y)\,P(x_2|Y)\,P(x_3|Y)\ldots P(x_d|Y)}_{'d'}$$

So, Liklihood $(P(X|Y=y)$ Now has $dk+k-1$ parameters. (with Assumption)

Its good to make simplicity Assumption because it is a linear function $(dk+k+1)$ unlike a power-ed function that would rise exponentially if d value grows to million of features.

As a Result, if we have large amount of feature vectors in our dataset, it is good to have Simplicity Assumption true, due to computational efficiency

4.1 c) Let X_w are independent random vocabulary words where $w \in (0 \cdots v)$

$$P(X_w | Y=y) = \theta_{yw} \ \& \ P(X_w=0 | Y=y) = 1-\theta_{yw}$$

Acc to Burnoli Distribution:

$$p^{x_i}(1-p)^{1-x_i}$$

$$P(X_w | Y=y)^{x_i} \ P(X_w=0 | Y=y)^{1-x_i}$$

Liklihood function

$$\mathcal{L}(p) = \prod_{w=1}^{v} P(X_w | Y=y)^{x_w} \ P(X_w=0 | Y=y)^{1-x_w}$$

$$\mathcal{L}(p) = P(X_1 | Y=y)^{x_1} P(X=0 | Y=y)^{x_1} . P(X_2 | Y=y)^{x_2} P(X_2=0 | Y=y)^{1-x_2} . P(X_w | Y=y)^{x_w} \cdots P(X_w=0 | Y=y)^{1-x_w}$$

$$\mathcal{L}(p) = P(X_i | Y=y)^{\Sigma x_i} P(X_i=0 | Y=y)^{\Sigma(1-x_i)}$$

$$\mathcal{L}(p) = P(X_i | Y=y)^{\Sigma x_i} P(X_i=0 | Y=y^{n-\Sigma x_i}$$

To maximize liklihood $\frac{\partial \mathcal{L}(p)}{\partial p} = 0$

Log Transformation:

$$\log L(p) = \Sigma x_i \log P(X_i | Y=y) + (-1)(n-\Sigma x_i)\log P(X_i=0 | Y=y)$$

$$\frac{\partial \log(L(P))}{\partial P} = \frac{\Sigma x_i}{P(X_i | Y=y)} + \frac{n-\Sigma x_i}{P(X_i=0 | Y=y)} = 0$$

$$\Sigma x_i \, (P(X_i=0 | Y=y) = n - \Sigma x_i \, (P(X_i | Y=y)$$

$$\Sigma x_i \, (P(X_i=0 | Y=y) = nP(X_i | Y=y) - \Sigma x_i \, P(X_i | Y_i=y)$$

$$\Sigma x_i \, (P(X_i=0 | Y=y) + \Sigma x_i \, P(X_i | Y_i=y) = nP(X_i | Y=y)$$

$$\Sigma x_i(1) = nP(X_i | Y=y) \quad (\text{as } 1-\theta_{yw}+\theta_{yw}=1)$$

$$\frac{\sum_{i=0}^{v} \Sigma x_i}{n} = \hat{P}(X_i | Y=y.$$

$P(Y=1) = P$ $\quad$ $P(Y=2) = 1-P$

Burnolli Representation :-

$$(P)^{x_1}(1-P)^{x_2}$$

Likehood function:

$$L = P(Y=1)^{x_1} \; P(Y=0)^{x_2}$$

$$P^{x_1}(1-P)^{x_2}$$

Taking derivative of L to maximize

$$\frac{\partial L(\theta)}{\partial \theta} = 0$$

But we need to log transform first

$$\log(L) = x_1 \log(P) + x_2(\log(1-P)$$
$$\times(-1)$$

$$\frac{d \log(\theta)}{\partial(\theta)} = \frac{x_1}{P} - \frac{x_2}{1-P} = 0$$

$$\frac{x_1}{P} = \frac{x_2}{1-P}$$

$$x_1 - x_1 P = x_2 P$$

$$x_1 = x_2 P + x_1 P$$

$$x_1 = P(x_1 + x_2)$$

$$\boxed{P = \frac{x_1}{x_1 + x_2}}$$

4.2

4 features $= (x_1, x_2, x_3, x_4)$

3 labels $= (+1, 0, -1)$

a) For prior probability

$$\boxed{\begin{aligned} P(Y=+1) &= \frac{2}{7} \\ P(Y=0) &= \frac{2}{7} \\ P(Y=-1) &= \frac{3}{7} \end{aligned}}$$

Because

$\hat{P}(y = y_k)$

$\left\{ \frac{\text{No. of } y = y}{\text{Total values in Data.}} \right\}$

b) $P((X_1 = 1) | Y = +1) = 1/2$

$P(X_2 | Y = +1) = 1$

$P(X_3 | Y = +1) = 1/2$

$P(X_4 | Y = +1) = 1/2$

$P(X_1 = 1 | Y = 0) = 1/2$

$P(X_2 = 1 | Y = 0) = 1/2$

$P(X_3 = 1 | Y = 0) = 1$

$P(X_4 = 1 | Y = 0) = 1$

$P(X_1 = 1 | Y = -1) = 1/3$

$P(X_2 = 1 | Y = -1) = 1/3$

$P(X_3 = 1 | Y = -1) = 1/3$

$P(X_4 = 1 | Y = -1) = 2/3$

$x_i \downarrow$ $y_i \rightarrow$	$y=+1$	$y=0$	$y=-1$
$x_1=1$	$1/2$	$1/2$	$1/3$
$x_2=1$	1	$1/2$	$1/3$
$x_3=1$	$1/2$	1	$1/3$
$x_4=1$	$1/2$	1	$2/3$

c) The given data point with features x_1, x_2, x_3, x_4

$$\left(\prod_{i=1}^{v} P(x|y=y)\right) P(Y=y)$$

If $y=0$

$\rightarrow P(x_1|y=0) \times P(x_2|y=0) \times P(x_3|y=0) \times P(x_4|y=0)\ P(y=0)$

$\rightarrow 1/2 \times 1/2 \times 1 \times 1 \times 2/7$

$\rightarrow 1/14$

If $y=+1$

$\rightarrow P(x_1|y=+1) \times P(x_2|y=+1) \times P(x_3|y=+1) \times P(x_4|y=+1) P(y=1)$

$\rightarrow 1/2 \times 1/2 \times 1 \times 1/2 \times 2/7$

$\rightarrow 1/28$

If $y=-1$

$\rightarrow P(x_1|y=-1) \times P(x_2|y=-1) \times P(x_3|y=-1) \times P(x_4|y=-1) (P(y=-1)$

$\rightarrow 1/3 \times 1/3 \times 1/3 \times 2/3 \times 3/7$

$\rightarrow 2/189$

Q5.

Objective function states to classify in the class which has maximum probability

From above calculation we find that

$$1/14 > 1/28 > 2/189$$

So, The data point should be classified into

<u>$y=0$ class</u> (Max probability)