Conditional Probability Revision.

[Prob. 1]

$$P(A/B) = \frac{P(A \cap B)}{P(B)} \quad \therefore \quad P(A \cap B) = P(A/B) \cdot P(B)$$

Conditional Independence

$$P(A,B/C) = P(A \cap B \cap C) / P(C)$$
$$= P(A/B,C)\ P(B \cap C) / P(C)$$

Now if A, B are conditionally independent:

$$P(A,B/C) = P(A/C) \cdot P(B/C)$$

Thus for 'n' values we have.

$$P(A_1, A_2, \ldots A_n/C) = P(A_1/C)\ P(A_2/C) \cdots P(A_n/C)$$

[Baye's Theorem Revisited.]

Consider the given scenario:

'h' → Customer will buy computer

'D' → 35 yr old cust, income = $ 50,000 PA

$P(h/D)$ = Customer will buy computer, given age, income.

$P(h)$ = Any customer will buy computer regardless of age, income.

$P(D)$ → Person is 35 yr old, earns = $ 50 K PA

$$P(h/D) = \frac{P(D/h)\ P(h)}{P(D)}.$$

Rob-2

Data table

Sl Nos	Age	Income	Buys computer
1	35	Med	Yes
2	30	High	NO
3	40	Low	NO
4	35	Med	Yes
5	45	Low	Yes
6	35	high	Yes
7	35	Med	NO
8	25	Low	NO
9	28	high	NO
10	35	Med.	Yes

Assume
Med inc = $50K

From the data we have.

$P(\text{buys computer}) = 5/10 = 0.5$

$P(\text{Age}=35, \text{Earns}=\text{Med}) = 4/10 = 0.4$

Now given that: Buys a computer = Yes.

$$P(\underbrace{35\text{yrs} + \text{Med}}_{3} / \underbrace{\text{buys} = \text{Yes}}_{5\text{ Nos}}) = 3/5 = 0.6 \quad 5/10$$

$$P(h_1/E) = \frac{P(E/h_1)\,P(h_1)}{P(E)} = \frac{0.6 \times 0.5}{0.4} = 0.75$$

h_1 = Cust buys a computer

E = Earns Med Aged 35.

Naive Baye's Classifier

Usage] Classify docs as spam/ Legit etc.

[Assumption] - Each feature is independent of the value of the other.

Naive Bayes] Considers each feature to contribute independent irrespective of correlation between the features

The model]

Baye's Theorem: $P(H)$ Probability of the hypothesis.

$P(H/E)$ Probability of the hypothesis after getting the evidence.

$$P(H/E) = \frac{P(E/H)\,P(H)}{P(E)}$$

Typical Applications:

- Spam vs No spam
- Doc subject classification
- Movie Reviews.

Text classification.

Input a document 'd'

Fixed set of classes: $C = (c1, c2, \cdots cj)$

output : Predicted class

Supervised m/c learning.

input: a doc 'd'
Fixed set of classes $C = \{C_1, C_2, C_3 \cdots C_d\}$
Training set of 'm' hand labelled docs.
$(d_1, c_1) \rightarrow (d_m, c_m)$

Output
a learned classifier $y: d \rightarrow c$.

Applying Bayes Theorem to doc class for doc 'd' and class 'c'

$$P(c/d) = \frac{P(d/c)\,P(c)}{P(d)}$$

Now we have:

C_{Map} = 'Maximum a posteriori' most likely class.

$$C_{Map} = \text{argmax } P(c/d)$$

$$= \text{argmax } P(d/c)\,P(c) \quad \text{dropping denominator.}$$

$$= \text{argmax } P(\underbrace{x_1, x_2, \cdots x_n}_{\substack{\text{doc represented as features} \\ x_1, x_2 \cdots x_n}}/c)\,P(C)$$

Multinomial Naive Bayes Independence Assumptions

Consider $P(d/c) = P(x_1, x_2 \cdots x_n/c)$

Considering independence of features, we get:

$$P(x_1, \cdots x_n/c) = P(x_1/c) \cdot P(x_2/c) \cdots P(x_n/c)$$

Thus

$$C_{NB} = \underset{x \in X}{\text{argmax}}\ P(c_j) \prod P(x/c)$$

Naive Baye's Learning.

First iteration: Use frequencies in the data.

$$\hat{P}(c_j) = \frac{\text{doc. count } c = c_j}{\text{Tot Nos of docs.}}$$

$$\hat{P}(w_i/c_j) = \frac{\text{Count}(w_i, c_j)}{\sum \text{Count}(w_i, c_j)}$$

Example. Which class does document '5' belong to?

Set	Doc	Words.	class.
Train.	1	china Beijing china	c
	2	China china Shanghai	c
	3	China Macao	c
	4	Delhi India china	i
Test	5	China China Delhi	?

Problem

Find $P(c/d_5)$ and $P(i/d_5)$

Soln.

$$P(c/d_5) = \frac{P(d_5/c)\, P(c)}{P(d_5)}.$$

Step 1

Priors: $P(c) = \frac{\text{doc count}(c)}{\text{Tot Docs.}} = \frac{3}{4}$

Similarly $P(i) = 1/4.$

[STEP 1] We have the Prior Probabilities.

$P(c) = 3/4$ $P(i) = 1/4.$

[Step2] - Compute likelihood.

<u>likelihood:</u> Conditional probability of a word occuring in a document given that document belongs to a particular category.

$$P(\text{word}/\text{category}) = \frac{(\text{Nos of occurence of the word in all docs from a category} + \ldots}{(\text{All words in all docs from the Category} + \text{Total unique word count})}$$

Thus

$$P(\text{word}/\text{category}) = \frac{\text{Count}(w,c) + 1}{\text{Count } c + M}$$

Now we have the following

List of unique words:]

China Beijing Shanghai Macao Delhi India	Thus $M = 6$	For class 'c' Count $c = 8$ For class 'i' count $i = 3$

<u>Thus we have the following.</u> For class 'c'

$P(\text{China}/c) = (5+1)/(8+6) = 3/7$

$P(\text{Beijing}/c) = (1+1)/(8+6) = 1/7$

$P(\text{Shanghai}/c) = (1+1)/(8+6) = 1/7$

$P(\text{Macao}/c) = (1+1)/(8+6) = 1/7$

$P(\text{Delhi}/c) = (0+1)/(8+6) = 1/14$

$P(\text{India}/c) = ($[illegible]

Thus for class 'i' we get the following

$P(\text{China}/i) = (1+1)/(3+6) = 2/9$

$P(\text{Beijing}/i) = (0+1)/(3+6) = 1/9$

$P(\text{Shanghai}/i) = (0+1)/(3+6) = 1/9$

$P(\text{Macao}/i) = (0+1)/(3+6) = 1/9$

$P(\text{Delhi}/i) = (1+1)/(3+6) = 2/9$

$P(\text{India}/i) = (1+1)/(3+6) = 2/9$

Thus Now we can find

$$P(c/d_5) = P(c) * \left[\prod P\left(\text{word in } d_5/c\right)\right]$$

Thus

$$P(c/d_5) = \frac{3}{4} \times \left[\left(\frac{3}{7}\right)^2 \times \frac{1}{14}\right] = \frac{3}{4} \times \frac{9}{49} \times \frac{1}{14} = \frac{27}{2744} = 0.00$$

$$P(i/d_5) = \frac{1}{4} \times \left[\left(\frac{2}{9}\right)^2 \times \frac{2}{9}\right] = \frac{1}{4} \times \frac{4}{81} \times \frac{2}{9} = \frac{8}{2916} = 0.0027$$

Thus document 5 [China, China, Delhi]

↳ Classified as class 'c'