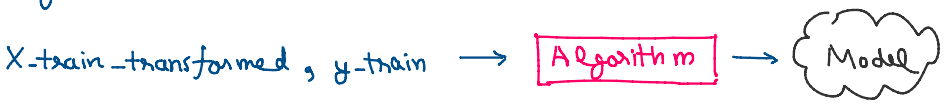
Author - Kanav Bansal
You can find me on **Linkedin** -
www.linkedin.com/in/kanavbansal
Or http://www.thataiguy.com

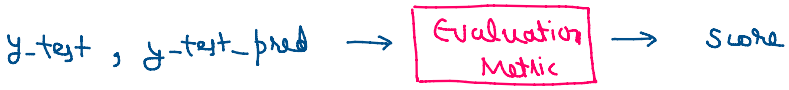
Building a Machine Learning Model

1. Training

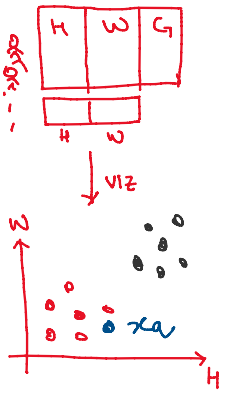


2. Predicting using unseen data

3. Evaluation



- Given Data containing Height, Weights & Gender of some individuals.



* Predict the Gender for x_q

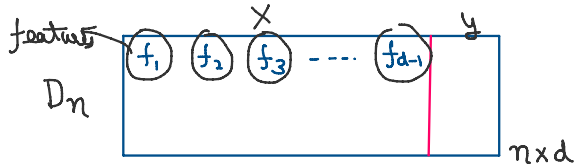
Type → Task
? ?

Evaluation Metric → ?

1. Target Variable → Gender (y)
   Gender ∈ {M, F}
2. Task → Classification
3. Evaluation Metrics → Accuracy
4. Algorithms → (How?) Probabilistic Approach
   Naive Bayes Algorithm

Naive Bayes Algorithm (AKA Maximum A Posterioxi)

Type – Supervised Learning
Task – Classification.

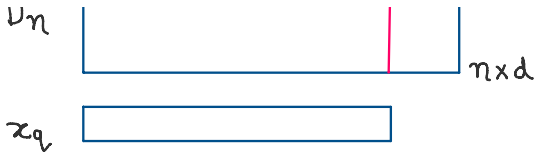


Given → $D_n = \{(x_i, y_i)_{i=1}^n \mid x_i \in \mathbb{R}^{d-1}, y_i \in \{C_1, C_2 \ldots C_k\}\}$

$x_q = f_1 \cap f_2 \cap \ldots \cap f_{d-1}$

Task → Find y_q

How? → Given x_q, compute the prob of x_q belonging to each of the classes. Whichever is maximum, assign that class to x_q.



Whichever is maximum, assign that class to x_q.

ie $P(C_1|x_q)$ $P(C_2|x_q)$ --- $P(C_K|x_q)$

$$y_q = \underset{C_1, C_2 \dots C_K}{\arg\max} \{ \underbrace{P(C_i|x_q)}_{\text{Posterior Prob.}} \}$$

From Bayes Theorem]

$$\underbrace{P(C_i|x_q)}_{\text{Posterior}} = \frac{\overbrace{P(x_q|C_i)}^{\text{Likelyhood}} * \overbrace{P(C_i)}^{\text{Prior}}}{P(x_q)\ ]\text{Marginal}}$$

Key Concepts]

→ Probability (Study of uncertainity)

Random Experiment
- It is a process for which outcome cannot be predicted with certainity.

Example]
R.E → Tossing a coin

Sample Space
- It is a set of all possible outcomes of a 'RANDOM EXPERIMENT'.

Example]
S.S = {H, T}

Event
- It is a subset of a 'SAMPLE SPACE'.

Example]
Event - Getting a Head
{H}

→ Probability →

$$P(\text{Event}) = \frac{\text{No. of favourable Outcomes}}{\text{Total no. of Outcomes}} = \frac{|\text{Event}|}{|\text{Sample Space}|}$$

→ Example-1 :

R.E → Tossing 2 coins
S.S → {HH, HT, TH, TT}
Event]
A : Getting atleast 1 Head
A → {HH, HT, TH}

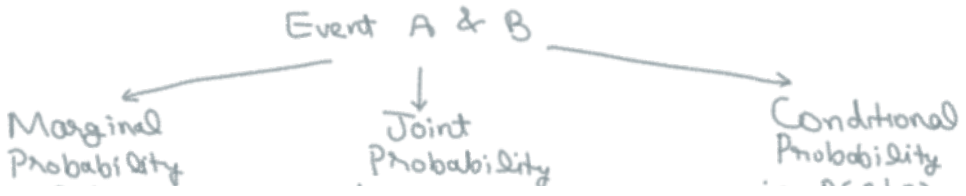
P(A) = 3/4

→ Example-2 :

R.E → Rolling a dice
S.S → {1, 2, 3, 4, 5, 6}
Event]
B : Getting a Prime no.
B → {2, 3, 5}

P(B) = 3/6

→ In case of two or more events]



Marginal Probability i.e. $P(A)$ & $P(B)$

Joint Probability i.e. $P(A \cap B)$

Conditional Probability i.e. $P(A|B)$

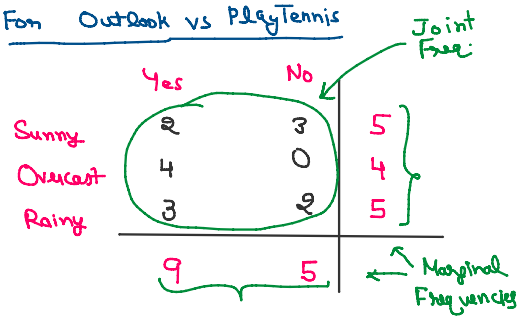
* Next we will see how to compute these probabilities with the help of 'FREQUENCY TABLE'.

Consider this binary classification data set:

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Frequency Table

For Outlook vs PlayTennis



Marginal Prob →

$P(\text{PlayTennis} = \text{Yes}) = \frac{9}{14}$

$P(\text{PlayTennis} = \text{No}) = \frac{5}{14}$

$P(\text{Outlook} = \text{Sunny}) = \frac{5}{14}$

Joint Prob →

$P(\text{Outlook} = \text{Sunny} \cap \text{PlayTennis} = \text{Yes}) = \frac{2}{14}$

$P(\text{Outlook} = \text{Overcast} \cap \text{PlayTennis} = \text{No}) = \frac{0}{14}$

$P(\text{Outlook} = \text{Rainy} \cap \text{PlayTennis} = \text{No}) = \frac{2}{14}$

Conditional Prob →

$$P(\text{Outlook} = \text{Sunny} \mid \text{PlayTennis} = \text{Yes}) = \frac{P(A \cap B)}{P(B)} = \frac{2/14}{9/14} = \frac{2}{9}$$

Bayes Theorem →

Acc^{n} to Conditional Prob → $P(A|B) = \frac{P(A \cap B)}{P(B)}$ — (1)

$P(B|A) = \frac{P(B \cap A)}{P(A)}$

$\because B \cap A = A \cap B$

$P(B|A) = \frac{P(A \cap B)}{P(A)}$

$P(A \cap B) = P(B|A)P(A)$ — (2)

Substituting (2) in eq^{n} (1) :

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Likelyhood

Substituting ② in eqn ①:

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

Likelyhood, Prior, Marginal, Posteriori

Independent Events

If given events A & B are know to be independent:

$$P(A|B) = P(A)$$
$$P(B|A) = P(B)$$

Similarly:

$$P(A \cap B) = P(A) * P(B)$$

Data →

$$D_n = \left\{ (x_i, y_i)_{i=1}^{n} \mid x_i \in \mathbb{R}^{d-1},\ y_i \in \{0,1\} \right\}$$

f_1	f_2	----	f_{d-1}	y
← x_1 →				y_1
← x_2 →				y_2
!				⋮
← x_n →				y_n

$$x_i = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_{d-1} \end{bmatrix}$$

Naive Bayes Algorithm

Task → Classification

How?

Given x_q, compute $P(C_1|x_q)$, $P(C_2|x_q)$, ---- $P(C_K|x_q)$.

Which ever probability is maximum, prediction would be the corresponding class.

$$\hat{y}_q = \underset{i \in \{1,2 \cdots k\}}{\arg\max} \{P(C_i|x_q)\}$$ i.e maximum a posteriori

$$\rightarrow P(C_1|x_q) = \frac{P(x_q|C_1)\,P(C_1)}{P(x_q)} \Rightarrow P(C_1|x_q) \propto P(x_q|C_1)\,P(C_1)$$

$$\rightarrow P(C_2|x_q) = \frac{P(x_q|C_2)\,P(C_2)}{P(x_q)} \Rightarrow P(C_2|x_q) \propto P(x_q|C_2)\,P(C_2)$$

⋮ ⋮

$$\rightarrow P(C_K|x_q) = \frac{P(x_q|C_K)\,P(C_K)}{} \Rightarrow P(C_K|x_q) \propto P(x_q|C_K)\,P(C_K)$$

$$\rightarrow P(C_K|X_q) = \frac{P(X_q|C_K)\,P(C_K)}{P(X_q)} \Rightarrow P(C_K|X_q) \propto P(X_q|C_K)\,P(C_K)$$

Naive Bayes Derivation

Given $X_q = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_{d-1} \end{bmatrix}$ i.e. X_q is $f_1 \cap f_2 \cap f_3 \cdots \cap f_{d-1}$

Using Bayes Theorem:

$$P(C_K|X_q) = \frac{P(X_q|C_K) * P(C_K)}{P(X_q)}$$

Compute → $\boxed{P(C_K|X_q) \propto P(X_q|C_K)\,P(C_K)}$

$$\hat{y}_q = \underset{K \in \{1,2\cdots K\}}{\arg\max} \{ P(C_K|X_q) \}$$

$$\hat{y}_q = \underset{K \in \{1,2\cdots K\}}{\arg\max} \{ \boxed{P(X_q|C_K)\,P(C_K)} \}$$

↳ Lets zoom inside this term and understand how to compute it step by step.

According to Conditional Probability ↴

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

$$\Rightarrow P(A \cap B) = P(A|B)\,P(B)$$

$P(X_q|C_K)\,P(C_K)$ can be written as $P(X_q \cap C_K)$

We have already seen that X_q is $f_1 \cap f_2 \cdots \cap f_{d-1}$

$$\rightarrow P(X_q|C_K)\,P(C_K)$$

$$\rightarrow P(X_q \cap C_K)$$

$$\rightarrow P(f_1 \cap f_2 \cap f_3 \cdots f_{d-1} \cap C_K)$$

$$P(\underbrace{f_1}_{A} \cap \underbrace{f_2 \cap f_3 \cdots \cap f_{d-1} \cap C_k}_{B})$$

Using $P(A \cap B) = P(A|B)\,P(B)$

$$\rightarrow P(f_1 \mid f_2 \cap f_3 \cdots \cap f_{d-1} \cap C_k) * P(\underbrace{f_2}_{A} \cap \underbrace{f_3 \cdots \cap f_{d-1} \cap C_k}_{B})$$

Expanding this expressing using Chain Rule of conditional Probability

$$\rightarrow P(f_1 \mid f_2 \cap f_3 \cdots \cap f_{d-1} \cap C_k) * P(f_2 \mid f_3 \cap f_4 \cdots \cap f_{d-1} \cap C_k) * \vdots * P(f_{d-1} \mid C_k) * P(C_k)$$

∴ Above expression is extremly hard to compute.

To simplify the computation lets take a NAIVE assumption that:
Features are conditionally independent of each other.
ie if $f_1, f_2, f_3 \cdots f_{d-1}$ are independent of each other,

$$P(f_1 \mid f_2 \cap f_3 \cdots \cap f_{d-1} \cap C_k) = P(f_1 \mid C_k)$$

$$\rightarrow P(f_1 \mid C_k) * P(f_2 \mid C_k) * \cdots * P(f_{d-1} \mid C_k) * P(C_k)$$

$$\rightarrow P(C_k) * \prod_{i=1}^{d-1} P(f_i \mid C_k)$$

$$\hat{y}_q = \underset{J \in \{1,2 \cdots k\}}{\arg\max} \left\{ P(X_q \mid C_J)\, P(C_J) \right\}$$

$$\downarrow$$

$$P(C_J) * \prod_{i=1}^{d-1} P(f_i \mid C_J)$$

ie $$\hat{y}_q = \underset{J \in \{1,2,\cdots k\}}{\arg\max} \left\{ P(C_J) * \prod_{i=1}^{d-1} P(f_i \mid C_J) \right\}$$

Flavours of Naive Bayes

1. Gaussian Naive Bayes

   If data contains numerical features use GaussianNB.

   Assumptions :
   - (a) X contains features which are conditionally independent
   - (b) All the numerical features follows Normal Distribution

$$P(f_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} * \exp\left\{ -\frac{(f_i - \mu_y)^2}{2\sigma_y^2} \right\}$$

2. Multinomial Naive Bayes

   If data contains text use MultinomialNB.

3. Bernoulli Naive Bayes

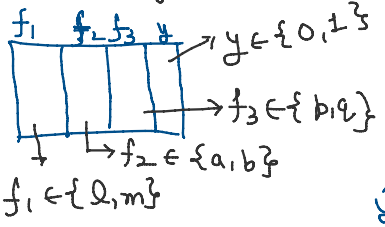
   If data contains categorical features use BernoulliNB.

Issues with Naive Bayes

1. Numerical Instability (For all flavours of NB)

2. Zero probability issue. (For MultinomialNB & BernoulliNB)

In x_q, if a feature contains a value that was never observed in the historical data, we will encounter a zero probability issue.



$x_q \rightarrow f_1 = m \;,\; f_2 = c \;,\; f_3 = p$

Acc^{n} to NB,

$$\hat{y}_q = \underset{i \in \{1,2,\ldots K\}}{\arg\max} \left\{ P(C_i) * \prod_{j=1}^{d-1} P(f_j \mid C_i) \right\}$$

i.e Find $P(C_1 = 0 \mid x_q) \propto P(C_1=0) * P(f_1 \mid C_1) * P(f_2 \mid C_1) * P(f_3 \mid C_1)$

& $P(C_2 = 1 \mid x_q) \propto P(C_2=1) * P(f_1 \mid C_2) * P(f_2 \mid C_2) * P(f_3 \mid C_2)$

Observe that → $P(f_2 \mid C_1) = 0$

& $P(f_2 \mid C_2) = 0$

Which means → $P(C_1=0 \mid x_q) = 0$
& $P(C_2=1 \mid x_q) = 0$

Now, how do we decide the class of x_q?

Solution → Apply regularization technique.

{ Here we will use → Laplace Smoothing }

ie Whenever the $P(f_J \mid C_i) = 0$, apply a smoothening technique so that the value is not exactly equal to 0.

Laplace Smoothing ⇒ $$\frac{P(f_J \mid C_i) + \alpha}{n + k * \alpha}$$

Here, α = Hyperparameter
$\alpha \in [0, +\infty]$

$n = P(C_i)$

k = # of unique values in f_J