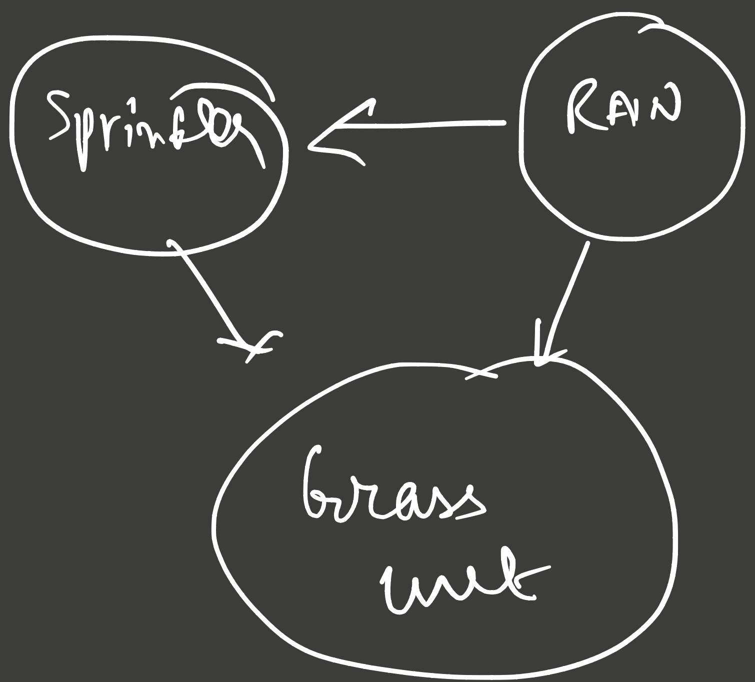
Bayesian Networks

Nodes: Random Variables

Edges: Direct Impact.



Classic Example

① Grass could be wet due to
- RAIN
- SPRINKLER

② If it RAINS, SPRINKLER MAY NOT BE USED

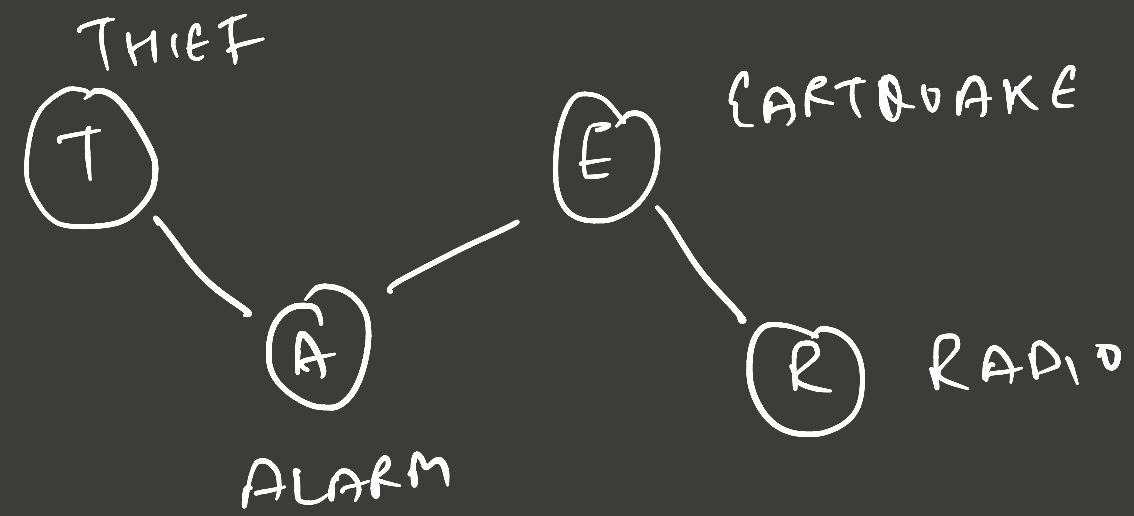
$$\underbrace{P(X_1 \ldots X_n)}_{\text{JOINT PROBABILITY}} = \prod_{k=1}^{N} P(X_k \mid \text{Parents}(X_k))$$

← Random Variables

$$\therefore \quad P(S, R, G) = P(G \mid S, R)\, P(S \mid R)\, P(R)$$

Q) what is $P(G = \text{wet}, \text{Rain} = \text{True}, \text{Sprinkler} = \text{True})$

$$= P(G = W \mid S = T, R = T)\, P(S = T \mid R = T)\, P(R = T)$$



KNOWN

$$P(T),\ P(E),\ P(A \mid T, E),\ P(R \mid E)$$

Q)

$$P(A \mid T) = \frac{P(A, T)}{P(A)} = \frac{P(A, T, E) + P(A, T, \bar{E})}{P(A, T, E) + P(A, T, \bar{E}) + P(\bar{A}, T, E) + P(\bar{A}, T, \bar{E})}$$

Medical Diagnosis

① you tested +ve for a disease

② Test is 99% accurate $\Rightarrow p(\text{Test}=+\text{ve} \mid \text{Diseas}=\text{True}) = .99$

$= p(\text{Test}=-\text{ve} \mid \text{Disease}=\text{False})$

③ Rare disease (1 in 10,000 people)

Q) what is probability you have disease?

From ② $p(T \mid D) = .99$; $p(\bar{T} \mid \bar{D}) = .99 \Rightarrow p(T \mid \bar{D}) = .01$

From ③ $P(D) = 10^{-4}$; $P(\bar{D}) = 1 - 10^{-4}$

Q is: $p(D \mid T) = ?$ $\quad p(D \mid T) = \dfrac{p(T \mid D)\, p(D)}{p(T)}$

$$p(D|T) = \frac{p(T|D)\,p(D)}{p(T)}$$

$$= \frac{p(T|D)\,p(D)}{p(T|D)\,p(D) + p(T|\bar{D})\,p(\bar{D})}$$

$$= \frac{(.99)(10^{-4})}{(.99)(10^{-4}) + (.01)(1-10^{-4})}$$

$$= \frac{(.99)(.0001)}{(.99)(.0001) + (.01)(.9999)} \quad << .99$$

SPAM EMAIL CLASSIFICATION (USING NAIVE BAYES)

$$y \in \{\underbrace{0}_{\text{NOT SPAM}}, \underbrace{1}_{\text{SPAM}}\}$$

WORDS FROM ALL EMAILS

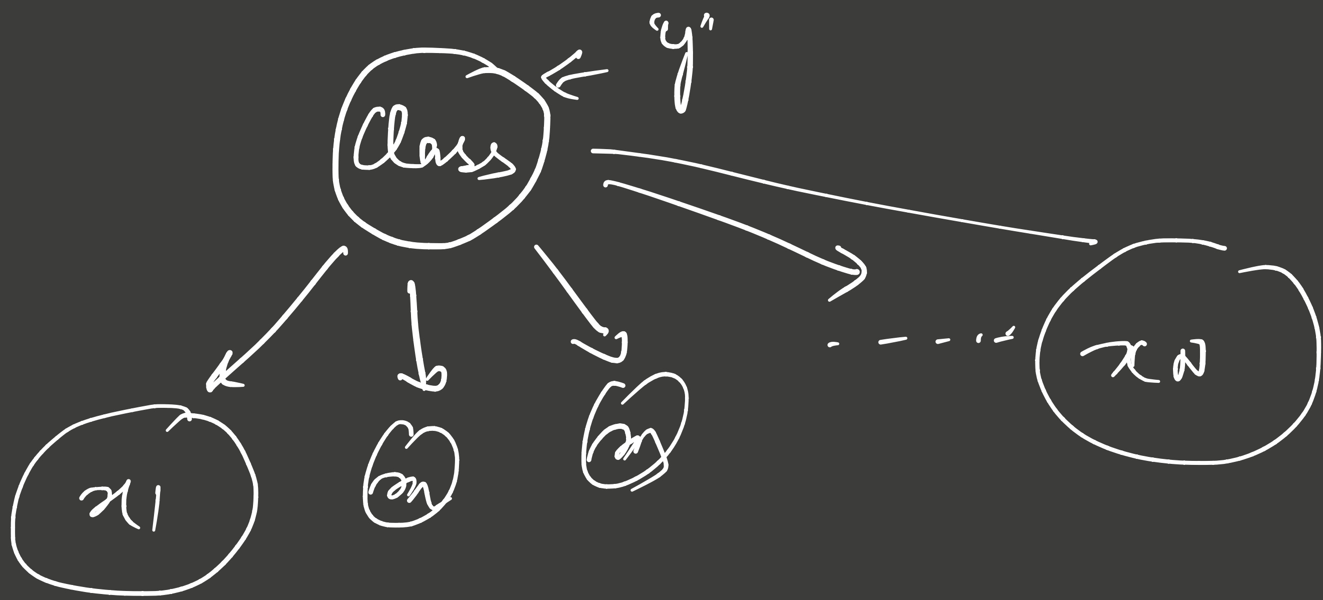
$$\left.\begin{bmatrix} \text{a} \\ \text{an} \\ \vdots \\ \text{computer} \\ \vdots \\ \text{lottery} \\ \vdots \\ \text{zoo} \end{bmatrix}\right\} \text{N words.} \qquad \leftarrow \text{Dictionary.}$$

New email.
Construct vector X

$$\begin{array}{r} \\ \text{"on" in email} \\ \\ \\ \\ \\ \end{array} \left.\begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \\ 1 \\ \vdots \\ 1 \end{bmatrix}\right\}$$

Naive Bayes

* Classification method.
* Scalable.
* Generative.



$$P(x_1, x_2 \dots x_N | y) = P(x_1|y)\, P(x_2|x_1, y) \cdots P(x_N|x_1 \dots x_{N-1}, y)$$

$\boxed{\text{Naive}}$: Naive assumption x_i and x_{i+1} are independent given y

i.e. $P(x_2|x_1, y) = P(x_2|y)$.

$\therefore$

$$p(x_1, x_2 \ldots x_N \mid y) = p(x_1 \mid y)\, p(x_2, y) \cdots p(x_N \mid y)$$

Q) what do we want to predict?

$$p(y \mid x_1, x_2 \ldots x_N)$$

$$= \frac{p(x_1, x_2 \ldots x_N \mid y)\, p(y)}{p(x_1, x_2 \ldots x_N)}$$

Parameters

$\Rightarrow$ Probability of x_i being in an spam email

$$p(x_i = 1 \mid y = 1) = \frac{\text{Count}(x_i = 1 \text{ and } y = 1)}{\text{Count}(y = 1)}$$

$\Downarrow$

$$p(x_i = 0 \mid y = 1) = \frac{\text{Count}(x_i = 0 \text{ and } y = 1)}{\text{Count}(y = 1)}$$

$\Rightarrow$ Probability of an email being spam / non-spam

$$p(y = 1) = \frac{\text{Count}(y = 1)}{\text{Count}(y = 1) + \text{Count}(y = 0)}$$

Example

Dictionary is: $[w_1\ w_2\ w_3]$

TRAIN SET

	w_1	w_2	w_3	y
1	0	0	0	1
2	0	0	0	0
3	0	0	0	1
4	1	0	0	0
5	1	0	1	1
6	1	1	1	0
7	1	1	1	1
8	1	1	0	0
9	0	1	1	0
10	0	0	1	1

	w_1	w_2	w_3	y
1	0	0	0	1
2	0	0	0	0
3	0	0	0	1
4	1	0	0	0
5	1	0	1	1
6	1	1	1	0
7	1	1	1	1
8	1	1	0	0
9	0	1	1	0
10	0	0	1	1

$y = 0$

$p(w_1=0|y=0) = 3/5 = 0.6$

$p(w_2=0|y=0) = 2/5 = 0.4$

$p(w_3=0|y=0) = 3/5 = 0.6$

$p(y=0) = 0.5$

$y = 1$

$P(w_1=1|y=1) = 2/5 = 0.4$

$p(w_2=1|y=1) = 1/5 = .2$

$p(w_3=1|y=1) = 3/5 = .6$

$p(y=1) = 0.5$

Test email $\{0, 0, 1\}$

$$p(y=1 \mid w_1 = 0, w_2 = 0, w_3 = 1)$$

$$= \frac{p(w_1=0 \mid y=1)\, p(w_2=0 \mid y=1)\, p(w_3=p \mid y=1) * p(y=1)}{p(w_1=0, w_2=0, w_3=0)}$$

$$= \frac{(1 - p(w_1=1 \mid y=1))(1 - p(w_2=1 \mid y=1))(p(w_3=1 \mid y=1)) * p(y=1)}{Z}$$

$$= \frac{(0.6) * (0.8) * (0.6) * 0.5}{Z}$$

Similarly,

$$p(y=0 \mid w_1=0, w_2=0, w_3=1) = \frac{(.6) * (.4) * (0.6) * 0.5}{Z}$$

$$\frac{p(y=1 \mid w_1=0, w_2=0, w_3=1)}{p(y=0 \mid w_1=0, w_2=0, w_3=1)} = \frac{6\times 8\times 6\times 5}{6\times 4\times 6\times 5} = 2$$

$\because$ $p(y=1, \dots) > p(y=0, w\dots)$

$\therefore$ we estimate this to be a "spam" message

Gaussian Naive Bayes

Classes $C_1, \dots C_K$

Continuous attribute x

For class C_k, $\mu_K = \text{mean}(x \mid y(x) = C_k)$

$\sigma^2_k = \text{Variance}(x \mid y(x) = C_k)$

Now for x = some observation 'v'.

$$p(x = v \mid C_k) = \frac{1}{\sqrt{2\pi\sigma^2_k}} e^{\frac{-(v-\mu_k)^2}{2\sigma^2_k}}$$

Example (from wikipedia)

Height (Feet)	Weight (lbs)	Foot size (inches)	Gender
6	180	12	m
5.92	190	11	m
5.58	170	12	m
5.92	165	10	m
5	100	6	F
5.5	150	8	F
5.42	130	7	F
5.75	150	9	F

	Male	Female
Mean (height)	5.855	5.4175
Variance (height)	3.5×10^{-2}	9.7×10^{-2}
Mean (weight)	176.25	132.5
variance (weight)	1.22×10^{2}	5.58×10^{2}
mean (foot)	11.25	7.5
Variance (foot)	9.17×10^{-1}	1.67

Q) Classify person height = 6ft, weight = 130 lbs,
foot = 8 inches.

$$P(\text{male} \mid h = 6ft, \text{weight} = 130 \text{ lbs}, \text{foot} = 8 \text{ inches})$$

$$= \frac{P(\text{male}) * P(h = 6ft \mid \text{male}) * P(\text{weight} = 130 \text{ lbs} \mid \text{male}) * P(\text{foot} = 8 \mid \text{male})}{Z}$$

$$= \frac{0.5 * \frac{1}{\sqrt{2\pi \, \text{Var}(\text{height} \mid \text{male})}} e^{\frac{(6 - 5.855)^2}{2 \, \text{Variance}(\text{height} \mid \text{male})}} * \ldots}{}$$

$$= \frac{6.2 \times 10^{-9}}{Z}$$

$$P(\text{female} \mid \overset{\text{Data}}{\ldots}) = \frac{5.4 \times 10^{-4}}{2}$$

$\therefore$ Classified as female.

Generating Data | Sampling from Naive Bayes Model.