Random Variables & Probability Distributions

Variable · whose value can change.

- Deterministic [governed by deterministic equations]
- Random [Value is not pre-determined] can change in unpredictable ways but has a probability distribution.
  - Discrete
  - Continuous

Discrete R.V: Coin Toss $X = \begin{cases} 1 & \text{Heads} \\ 0 & \text{Tails} \end{cases}$

Bernoulli Distribution: $P(X=1) = p$, $P(X=0) = q = 1-p$, $0 \le p \le 1$

Dice Roll.

Binomial Distribution : N coin tosses

$$P_r(X=k) = f(k,n,p) = {}^nC_k\, p^k (1-p)^{n-k}$$

WRITE CODE to plot this for large 'n'.

Poisson Distribution : $f(k;\lambda) = \dfrac{\lambda^k e^{-\lambda}}{k!} = P_r(X=k)$ [Number of events in a (fixed) time interval]

$$E[X] = \lambda = \text{Var}[X]$$

$$\Omega = \{x_1, x_2, x_3, \cdots, x_n\}$$

$$0 \le P(X = x_i) \le 1$$

$$\sum_{i=1}^{n} P(X = x_i) = 1.$$

Continuous R.V.

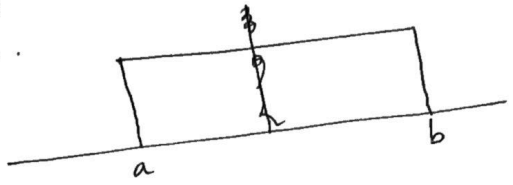
Any measuring device has finite precision & range

when outcome can be any real number [uncountable].

probability of X taking any one value is ZERO.

Practical : when no. of possibilities is very large

Uniform Distribution :



Distribution $f(x=x) = \frac{1}{b-a}$

$$P(X=x) = 0$$

$$P(x \le X \le x+\Delta x) = f(X=x)\,\Delta x, \text{ if } \Delta x \approx 0$$

$$\int_{-\infty}^{\infty} f(X=x)\,dx = 1$$

$$f(X=x) \ge 0$$

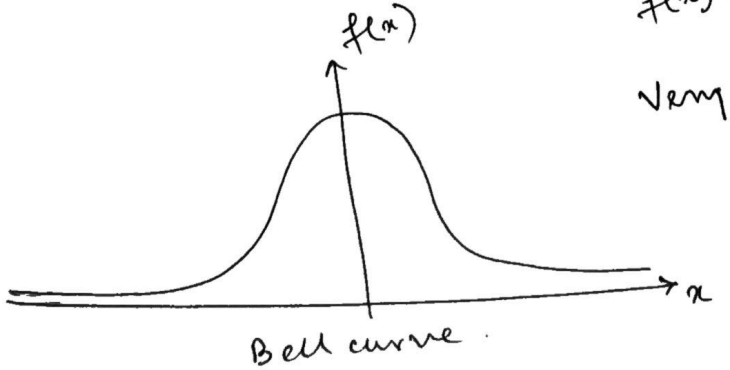
~~$0 \le f(X=x) \le 1$~~ $f(X=x)$ can be greater than 1 over a finite range of x values.

Gaussian :

Velocity of gas molecules
Noise in communication channels
Height of a large population.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

Very crucial in ML too!!



Bell curve.

Expectation, Moments & CLT

Expectation: $\mathbb{E}[g(x)] = \int_{-\infty}^{\infty} g(x) f(x)\, dx$

Mean: $\mathbb{E}[x] = \int_{-\infty}^{\infty} x f(x)\, dx = \mu$

Variance: $\sigma^2 = \mathbb{E}[(x-\mu)]^2 = \mathbb{E}[x^2] - \mathbb{E}[x]^2$

Moments: $m_n = \mathbb{E}[x^n] = \int_{-\infty}^{\infty} x^n f(x)\, dx$

Characteristic function: $\Phi_x(\omega) = \int_{-\infty}^{\infty} f(x)\, e^{j\omega x}\, dx = \mathbb{E}[e^{j\omega x}]$

Moment Generating function (MGF): $M_x(s) = \int_{-\infty}^{\infty} f(x) e^{sx}\, dx = \mathbb{E}[e^{sx}]$

$M_x^{(n)}(0) = \mathbb{E}[x^n] = m_n$

If all moments are finite & the following series converges absolutely, then

$M_x(s) = \sum_{n=0}^{\infty} \frac{m_n}{n!} s^n$ [does not work for Log-normal distribution]

If X & Y are independent random variables,

$\Phi_{X+Y}(\omega) = \mathbb{E}[e^{j\omega(x+y)}]$

$= \mathbb{E}[e^{j\omega x}]\,\mathbb{E}[e^{j\omega y}] = \Phi_x(\omega)\,\Phi_y(\omega)$

MGF of Normal Distribution

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \qquad \int_{-\infty}^{\infty} f(x)\,dx = 1$$

$$M(s) = \mathbb{E}[e^{sx}] = \int_{-\infty}^{\infty} e^{sx} f(x)\,dx$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{sx} e^{-x^2/2}\,dx$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(x-s)^2/2} e^{s^2/2}\,dx$$

$$= e^{s^2/2} \left[\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(x-s)^2/2}\,dx \right] = e^{s^2/2}$$

Central Limit Theorem

$$\bar{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n} \qquad \{X_i\} \text{ i.i.d. rv}$$

Let $Y_i = X_i - \mu$ & $Z_n = \dfrac{Y_1 + Y_2 + \cdots + Y_n}{\sqrt{n\sigma^2}}$

$$M_Y(s) = 1 + m_1 s + \frac{m_2}{2!} s^2 + \cdots = 1 + \frac{\sigma^2}{2} s^2 + \cdots$$

$$M_{Z_n}(s) = \left[M_Y\left(\frac{s}{\sqrt{n\sigma^2}} \right) \right]^n = \left[1 + \frac{\sigma^2}{2} \left(\frac{s}{\sqrt{n\sigma^2}} \right)^2 + \cdots \right]^n$$

$$= \left[1 + \frac{s^2}{2n} + \cdots \right]^n \sim e^{s^2/2} \quad \text{for large } n$$

BAYES' THEOREM

$U = [A_1, A_2, \cdots, A_n]$ is a partition of S, meaning A_i & A_j are mutually exclusive & $\sum_i P(A_i) = 1$

Let B be an arbitrary event:

$$\therefore B = B \cap S = B \cap [A_1 \cup A_2 \cup A_3 \cdots \cup A_n]$$

$$= (B \cap A_1) \cup (B \cap A_2) \cdots \cup (B \cap A_n)$$

BA_i & BA_j are mutually exclusive $\therefore P(B) = P(B \cap A_1) + P(B \cap A_2) + \cdots + P(B \cap A_n)$

Total probability theorem: $P(B \cap A_i) = P(B/A_i)\, P(A_i)$

$$= P(A_i/B_i)\, P(B_i)$$

$$\Rightarrow P(A_i/B) = \frac{P(B/A_i)\, P(A_i)}{P(B)}$$

$$= \frac{P(B/A_i)\, P(A_i)}{\sum_i P(B/A_i)\, P(A_i)}$$

BAYES' THEOREM.

Example: A test for cancer is known to be 95% accurate. A person submits to test & is found positive. Let the person belong to a population of 1L people out of which 2000 actually have the disease. whats the prob. of the person being actually infected?

$P(T_+/C) = 0.95$ $\quad P(T_-/H) = 0.95$

$P(H) = 0.98$ $\quad P(C) = 0.02$

$$P(C/T_+) = \frac{P(T_+/C)\, P(C)}{P(T_+)} = \underline{0.278}.$$

$$P(T_+) = P(T_+/C) P(C) + P(T_+/H) P(H)$$

~~307~~

Example: Papoulis page 34

$$P(D/B_1) = \frac{100}{2000} = 0.05 \qquad P(D/B_2) = \frac{200}{500} = 0.4$$

$$P(D/B_3) = \frac{100}{1000} = 0.1 \qquad P(D/B_4) = \frac{100}{1000} = 0.1$$

$$P(B_1) = P(B_2) = P(B_3) = P(B_4) = 1/4$$

$$P(D) = 0.1625$$

$$P(B_2/D) = \frac{0.4 \times 0.25}{0.1625} = 0.615$$

Example Papoulis page 104

n coin tosses k heads

Prob of head in $(n+1)$th toss?

Ans: $\frac{k+1}{n+2}$

Naive Bayes & Gaussian Naive Bayes

Bayes' Theorem

$U = [A_1, A_2, \ldots, A_n]$ is a partition of S
& B is an arbitrary event.

$$P(A_i/B) = \frac{P(B/A_i)\,P(A_i)}{\sum_i P(B/A_i)\,P(A_i)} = \frac{P(B/A_i)P(A_i)}{P(B)}$$

Dataset of 1L emails with 70% Normal & 30% Spam.
Find total count of each occuring words in each category

BAG of words model

w_1 = work
w_2 = deadline
w_3 = hand
w_4 = bonus
w_5 = luxury
w_6 = Money
(SPAM TRIGGER WORDS)
Total 100 words in each category

#words / #docs — Two ways to calculate

$p(w_1/N) = \frac{22}{100} = 0.22$
$p(w_2/N) = \frac{26}{100} = 0.26$
$p(w_3/N) = \frac{19}{100} = 0.19$
$p(w_4/N) = \frac{13}{100} = 0.13$
$p(w_5/N) = \frac{8}{100} = 0.08$
$p(w_6/N) = \frac{12}{100} = 0.12$

$p(w_1/S) = \frac{11}{100} = 0.11$
$p(w_2/S) = \frac{4}{100} = 0.04$
$p(w_3/S) = \frac{14}{100} = 0.14$
$p(w_4/S) = \frac{31}{100} = 0.31$
$p(w_5/S) = \frac{21}{100} = 0.21$
$p(w_6/S) = \frac{19}{100} = 0.19$

If $p(w_i/N) = 0$ & $p(w_i/S) \neq 0$ or reverse, add 1 to all word counts

New email received with words w_3, w_5 & w_6 [in any order]

Naive ≡ [we also assume word occurrence to be independent events]

$$P(N/w_3 \cap w_5 \cap w_6) = \frac{P(w_3 \cap w_5 \cap w_6/N)\,P(N)}{P(w_3 \cap w_5 \cap w_6)} \quad \leftarrow \text{Prior probability}$$

$$= \frac{P(w_3/N)\,P(w_5/N)\,P(w_6/N)\,P(N)}{P(w_3 \cap w_5 \cap w_6)}$$

High Bias – ignores word order
Low variance ↳ works well

$$\propto 0.19 \times 0.08 \times 0.12 \times \frac{0.7}{0.5} = \frac{0.00127\,68}{0.000912}$$

$$P(S/w_3 \cap w_5 \cap w_6) = \frac{P(w_3/S)\,P(w_5/S)\,P(w_6/S)\,P(S)}{P(w_3 \cap w_5 \cap w_6)}$$

$$\propto 0.14 \times 0.21 \times 0.19 \times \frac{0.3}{0.5} = \frac{0.00167\,58}{0.002793}$$

[Take log since numbers get very small underflow]

Gaussian Naive Bayes

Estimating $P(x_i/c_j)$ using Gaussian Distribution

used when 'x_i' values can be any real number in a range.

Eg: Boy vs. Girl classification using height, weight & feet size.

fit a Gaussian using mean & SD.

Use mean & variance of available data to fit Gaussian.

Bernoulli Naive Bayes : use Bernoulli distribution

$x_i \in \{0, 1\}$

word present or absent

Categorical Naive Bayes

Multinomial Naive Bayes : $x_i \in \{0, 1, 2, \ldots, K\}$

word count taken into account.

Bayes optimal classifier

Model parameter dependencies leading to the most general classifier. Abstract notion & hard to implement in practice.

BAYESIAN PARAMETER ESTIMATION

Event A : k Heads in n tosses

p : probability of getting Head in one toss

Event A' : Head in $(n+1)$th toss

Since we do not know the value of 'p' & are trying to estimate it, we need to consider it as a random variable

Before event A, lets say we have no information about 'p'

So, assume $f(p) = \begin{cases} 1, & 0 \le p \le 1 \\ 0, & \text{otherwise} \end{cases}$

$$P(A/p) = p^k (1-p)^{n-k}$$

$$f(p/A) = \frac{P(A/p)\, f(p)}{P(A)} = \frac{P(A/p)\, f(p)}{\int_0^1 P(A/p) f(p)\, dp}$$

$$= p^k (1-p)^{n-k} \cdot \frac{(n+1)!}{(n-k)!\, k!} \qquad 0 \le p \le 1$$

$\sim \beta(\alpha, \beta)$ Beta-distributed $\alpha = n$, $\beta = k$ here.

$$P(A'/A) = \int_0^1 \underbrace{P(A'/p)}_{p} f(p/A)\, dp$$

$$= \frac{k+1}{n+2}$$

This is generally very complicated to estimate in most problems. Hence, we instead estimate the value of 'ρ' that best explains the observed data.

There are two ways of doing it.

MLE & MAP

MLE: Assume that you have no prior info about the parameters [uniform distribution]

MAP: Assume availability of prior info based on reasonable considerations. also leads to "regularisation" which helps in preventing overfitting.

(x, y) Training Data

$\tilde{x}$ new input data

$\tilde{y}$ predicted output data.

$$p(\tilde{y}/x, y, \tilde{x}) = \int \underbrace{p(\tilde{y}/\rho, \tilde{x})}_{\text{model}} p(\rho/x, y)\, d\rho$$

$$p(\rho/x, y) = \frac{p(y/x, \rho)\, p(\rho)}{p(y/x)}$$

$$= \frac{p(y/x, \rho)\, p(\rho)}{\int p(y/x, \rho)\, p(\rho)\, d\rho}$$

BAYESIAN PARAMETER ESTIMATION.

MLE & MAP : $p(\tilde{y}/x, y, \tilde{x}) = p(\tilde{y}/\rho_E, \tilde{x})$

$$\rho_E = \arg\max_{\rho} p(\rho/x, y)$$

MLE & Regression

Let $y_i = g(x_i, \theta) + \epsilon_i$ → curve fitting

$$\theta_{MLE} = \underset{\theta}{\text{argmax}} \sum_i \log p(y_i | x_i, \theta)$$

$$= \underset{\theta}{\text{argmax}} \sum_i \log p(\epsilon_i)$$

$$= \underset{\theta}{\text{argmax}} \sum_i \log \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\epsilon_i^2/2\sigma^2}$$: Assuming Gaussian Error/Noise

$$= \underset{\theta}{\text{argmax}} \left[\sum_i \underbrace{\log \frac{1}{\sqrt{2\pi\sigma^2}}}_{\text{constant}} - \frac{1}{2\sigma^2} \sum_i \epsilon_i^2 \right]$$

$$= \underset{\theta}{\text{argmax}}\ \underset{\theta}{\text{argmin}} \sum_i \epsilon_i^2$$

$$= \underset{\theta}{\text{argmin}} \sum_i \left[y_i - g(x_i, \theta) \right]^2$$

Least Squares Error Cost function.

MLE & Classification

Predict output $y \in \{0, 1\}$ → Disease / Healthy; Bank default / pay; Fake news / factual

using a probabilistic function

$g(x, \theta) = P[y(x) = 1]$ ↳ could be Logistic Regression or ANN, etc.

$$\theta_{MLE} = \underset{\theta}{\text{argmax}}\ p(y_i | x_i, \theta)$$

$$= \underset{\theta}{\text{argmax}} \prod_i g_i^{y_i} (1 - g_i)^{1 - y_i}$$

$$= \underset{\theta}{\text{argmax}} \sum_i \left[y_i \log g_i + (1 - y_i) \log (1 - g_i) \right]$$

used in Logistic Regression

$$= \underset{\theta}{\text{argmin}} \left(- \right) \sum_i y_i \log g_i + (1 - y_i) \log (1 - g_i)$$

Binary Cross-Entropy

MLE ~~MAP~~ Maximum Likelihood Estimate

Naive Bayes assumes independence of parameter which is generally not true in real world.
So, need to build models & estimate parameter.

$$f_{\phi}(\theta|x,y) = \frac{p(y|x,\theta) f_{\phi}(\theta)}{p(y|x)}$$

$p(y|x,\theta)$ → Likelihood; $f_{\phi}(\theta)$ → prior, $p(\theta|x) = p(\theta)$

$f_{\phi}(\theta|x,y)$ — Posterior estimate

MLE

Assume $f_{\phi}(\theta)$ to have uniform distribution
i.e. no prior knowledge of parameters available.

$$\therefore \theta_{MLE} = \underset{\theta}{\text{argmax}}\, f_{\phi}(\theta|x,y)$$

$$= \underset{\theta}{\text{argmax}}\, p(y|x,\theta) \xrightarrow{} L$$

$$= \underset{\theta}{\text{argmax}} \prod_i p(y_i|x_i,\theta)$$

$$= \underset{\theta}{\text{argmax}} \sum_i \log p(y_i|x_i,\theta) \rightarrow \text{Log likelihood}$$

Coin toss experiment

ρ = prob. of heads (1) tails = 0

$$p(Y_i = y_i|\rho) = \rho^{y_i}(1-\rho)^{1-y_i} \qquad y_i \in \{1,0\}$$

↳ Bernoulli dist.

$$LL = \log p(y|\rho)$$

$$= \sum_{i=1}^{N} \log p(y_i|\rho) = \sum_{i=1}^{N} \log \rho^{y_i}(1-\rho)^{1-y_i}$$

$$= \log\rho \sum y_i + \log(1-\rho) \sum (1-y_i)$$

$$= n^{(1)} \log\rho + n^{(0)} \log(1-\rho)$$

$$\frac{\partial LL}{\partial \rho} = 0 \Rightarrow \frac{n^{(1)}}{\rho} - \frac{n^{(0)}}{1-\rho} = 0 \Rightarrow \frac{1}{\rho} - 1 = \frac{n^{(0)}}{n^{(1)}} \Rightarrow \rho_{MLE} = \frac{n^{(1)}}{n^{(1)} + n^{(0)}}$$

[Intuitive Estimate is MLE]

MAP Estimation - Maximum a posteriori

$$f_p(\theta/y,x) = \frac{p(y/x,\theta)\, f_p(\theta/x)}{p(y/x)} \qquad \leftarrow f_p(\theta)$$

$$\theta_{MAP} = \operatorname{argmax} f_p(\theta/y,x)$$

$$= \operatorname{argmax}\ p(y/x,\theta)\, f_p(\theta)$$

$$= \operatorname{argmax}\left[\sum_i \log p(y_i/x_i,\theta) + f_p(\theta)\right]$$

Usually clear info of $p(\theta)$ is not available.
So use $f_p(\theta)$ to enforce certain constraints.
[regularisation]

Eg: choose $f_p(\theta)$ such that there is low probability for θ taking high values.

Now Lets say we wish to make a prediction for new data point

$$p(\tilde{y}/\tilde{x},y,x) = p(\tilde{y}/\tilde{x},\theta) \rightarrow \theta_{MAP} \text{ or } \theta_{ML}$$

Usually,

$$p(\tilde{y}/\tilde{x},y,x) = \int p(\tilde{y}/\tilde{x},\theta)\, f_p(\theta/y,x)\, d\theta$$

But actually, $p(\tilde{y}/\tilde{x},y,x)$ =

→ usually we estimate this just using θ_{ML} or θ_{MAP} which is not same as RHS.
But estimating RHS is very hard since it requires $p(x)$ too.

MAP for Coin Toss

$$\theta_{MAP} = \operatorname{argmax} \sum_{i=1}^{N} \log p$$

MAP for Coin Toss

$$\theta_{MAP} = \arg\max \sum_{i=1}^{N} \{\log p(y_i/\theta) + \log f_\theta(\theta)\}$$

$$= \arg\max \left[n^{(1)} \log \theta + n^{(0)} \log(1-\theta) + \log \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)} \right]$$

Beta-distribution
$\alpha, \beta > 0$

$\theta = p(y_i = 1)$

$n^{(1)}$ = no. of 1s

$n^{(0)}$ = no. of 0s

$n^{(1)} + n^{(0)} = n$

$$= \arg\max \left[n^{(1)} \log \theta + n^{(0)} \log(1-\theta) + (\alpha-1)\log\theta + (\beta-1)\log(1-\theta) \right]$$

$$= \arg\max \left[(n^{(1)} + \alpha - 1)\log\theta + (n^{(0)} + \beta - 1)\log(1-\theta) \right] = I$$

$$\frac{\partial I}{\partial \theta} = 0 \Rightarrow \frac{n^{(1)} + \alpha - 1}{\theta} - \frac{n^{(0)} + \beta - 1}{1-\theta} = 0$$

$$\Rightarrow \frac{1}{\theta} - 1 = \frac{n^{(0)} + \beta - 1}{n^{(1)} + \alpha - 1}$$

$$\Rightarrow \theta_{MAP} = \frac{n^{(1)} + \alpha - 1}{n + \alpha + \beta - 2}$$

$\alpha = 1, \beta = 1$

$\Rightarrow$

$\theta_{MAP} = \theta_{MLE}$

$\theta = \frac{n^{(1)}}{n}$

$(\alpha - 1) - \alpha + 2$

Bayesian Estimation

$$p(\tilde{y}/y) = \int_0^1 p(\tilde{y}/\theta)\, p(\theta/y)\, d\theta$$

$$p(\theta/y) = \frac{p(y/\theta)\, p(\theta)}{p(y)} = \frac{p(y/\theta)\, p(\theta)}{\int_0^1 p(y/\theta)\, p(\theta)\, d\theta}$$

$$= \frac{\left[\prod_i p(y_i/\theta)\right] p(\theta)}{\int_0^1 \left[\prod_i p(y_i/\theta)\right] p(\theta)\, d\theta}$$

$$= \frac{\theta^{n^{(1)}} (1-\theta)^{n^{(0)}} \theta^{\alpha-1} (1-\theta)^{\beta-1}}{B(n^{(1)} + \alpha,\ n^{(0)} + \beta)}$$

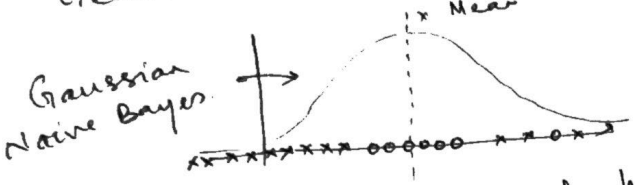
↳ Assume Beta distribution

GENERATIVE vs. DISCRIMINATIVE MODELS

Bayesian Reasoning (BPE)

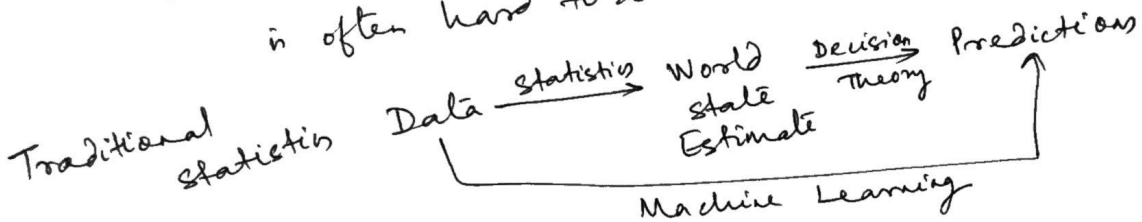
Function Approximation (MLE & MAP) ML

Generative models can go wrong when there are outliers.



SVM, for example, can handle this with ease.

Generative may work here if we remove outliers or make better choice of distribution, but that is often hard to do.



Generative	Discriminative
Goal: To explain data (concept learning)	Goal: To make predictions (shortcut coaching)
More accurate when they use correct models	More robust against outliers & bad models
Explainable	Blackbox as we use more complex models
Generate realistic data samples (artwork, super-resolution, colorization, etc.)	
simulation & planning for reinforcement learning	

Generative Adversarial Networks (GANs)

GANs

Generators vs. Discriminators

Generate new data (fake) to fool the discriminator into believing that samples ~~data~~ generated is real

Identify fake (new generated) samples & force the generator to improve.

Applications: Art AI
Imaginary fashion models
Model dark matter distribution & predict Gravitational lensing

DeepFakes!!

MNIST handwritten digits

Bayesian Networks - Help in identifying how causes ~~for~~ generate ~~the~~ outcomes/events.
[work both ways bottom up & top down]

PROBABILISTIC GRAPH MODELS
[Bayesian Networks & Markov Network]

Probability Theory + Graph Theory

Directed Acyclic Graphs — Bayesian Networks
↓
AI & Stats

Undirected Graphs — Markov ~~Net~~ Networks or MRF
[could be cyclic]
↳ Physics & Vision

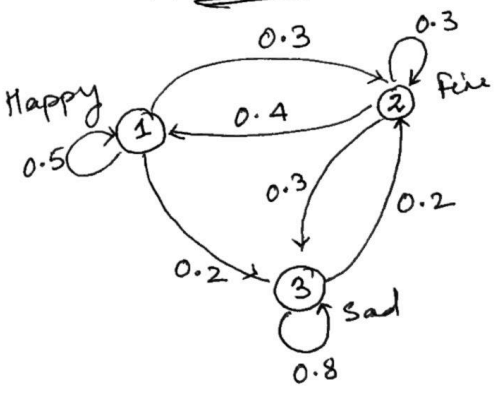
Chain Graph — both directed & undirected paths.

DAGs — Encode causal relationships
Easier to Learn from data
But cannot represent cyclic dependencies

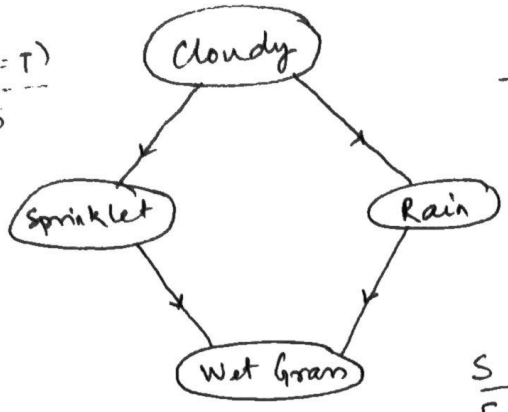
PGM ≠ Markov chain
↳ Nodes are elements of state space & edges denote transition probabilities
↓
Nodes are random variables & edges denote conditional dependencies.

Markov: Future is independent of the past given the present.

Markov chain



P(C=F)	P(C=T)
0.5	0.5



C	P(S=F)	P(S=T)
F	0.5	0.5
T	0.9	0.1

C	P(R=F)	P(R=T)
F	0.8	0.2
T	0.2	0.8

S	R	P(W=F)	P(W=T)
F	F	1.0	0.0
T	F	0.1	0.9
F	T	0.1	0.9
T	T	0.01	0.99

~~P(S,~~

$$P(C,S,R,W) = P(C)P(S|C)P(R|C,S)P(W|C,R,S)$$

$$= P(C)\,P(S|C)P(R|C)\,P(W|R,S)$$

[simplification of conditional probabilities]

Most common task : Probabilistic Inference.

$$P(S=T|W=T) = \frac{P(S=T \,\&\, W=T)}{P(W=T)}$$

Bottom Up →

$$= \frac{\sum_{R,C \in T/F} P(S=T \,\&\, W=T \,\&\, R=T/F \,\&\, C=T/F)}{P(W=T)}$$

$$P(W=T) \rightarrow \sum_{R,S,C \in T/F} P(S=T/F \,\&\, W=T \,\&\, R=T/F \,\&\, C=T/F)$$

Top Down

↳ ~~P(C=T)~~

$P(W=T|C=T)$

$$= \frac{0.2781}{0.6471} = 0.430$$

∴ If grass is wet, it is more likely to be due to rain!

"Explaining away"

$$P(R=T|W=T) = \frac{0.4581}{0.6471} = 0.708$$

$P(S=T|W=T \,\&\, R=T) = 0.1945$ Berkson's Paradox

S & R actually are independent

Also called "Selection Bias"

EXPECTATION MAXIMIZATION

Bayesian Networks

Structure	Observability	Method
Known	Full	MLE
Known	Partial	Expectation Maximization (or Gradient Descent)
Unknown	Full	Search through Model space [NP-Hard]
Unknown	Partial	EM + Search Model space

Two coins with biases θ_A & θ_B

Pick a coin at random & toss it 10 times

Repeat this 5 times

Record $X = (x_1, x_2, \cdots, x_5)$ & $(z_1, z_2, z_3, z_4, z_5) = Z$

$x_i \in \{0, 1, 2, \cdots, 10\}$ = No. of heads observed

$z_i \in \{A, B\}$ = Coin chosen

Estimating θ_A & θ_B is easy if this full info is available

But what if z_i are <u>hidden variables</u>??

<u>Simple Approach</u> -

Assume some initial values for $\theta_A^{(0)}, \theta_B^{(0)}$

Then estimate $\{z_i\}$ based on which coin is more likely to generate the corresponding $\{x_i\}$

Estimate $\theta_A^{(1)}, \theta_B^{(1)}$ using this completed info

Repeat the above till convergence reached.

<u>EM Approach</u> - Compute probability of each possible completion of table.

Created weighted training set of all these possibilities

Then use a modified form of MLE.

$X = \{5, 9, 8, 4, 7\}$ $\quad z = ?$

Simple Approach [MLE]

Assume $\theta_A^{(0)} = 0.60$ $\quad \theta_B^{(0)} = 0.5$

$P(x_1/A) = {}^{10}C_5 0.6^5 0.4^5 = 0.2$ $\quad P(x_1/B) = 0.25$

$P(x_2/A) = {}^{10}C_9 0.6^9 0.4^1 = 0.04$ $\quad P(x_2/B) = 0.0098$

$P(x_3/A) = {}^{10}C_8 0.6^8 0.4^2 = 0.12$ $\quad P(x_3/B) = 0.044$

$P(x_4/A) = {}^{10}C_4 0.6^4 0.4^6 = 0.11$ $\quad P(x_4/B) = 0.205$

$P(x_5/A) = {}^{10}C_7 0.6^7 0.4^3 = 0.22$ $\quad P(x_5/B) = 0.117$

$$\theta_A^{(1)} = \frac{9+8+7}{30} = 0.8 \qquad \theta_B^{(1)} = \frac{5+4}{20} = 0.45$$

Expectation Maximization

another way to estimate MLE/MAP

→ E-step

x_i	$P(x_i/A)$	$P(x_i/B)$	ω_{Ai}	ω_{Bi}	A	B
5	0.2	0.25	0.45	0.55		
9	0.04	0.0098	0.80	0.20		
8	0.12	0.044	0.73	0.27		
4	0.11	0.205	0.35	0.65		
7	0.22	0.117	0.65	0.35		

$\omega_{Ai} = \dfrac{P(x_i/A)}{P(x_i/A) + P(x_i/B)}$ $\qquad \omega_{Bi} = 1 - \omega_A$

M-step

$$\theta_A^{(1)} = \frac{\sum x_i \omega_{Ai}}{10 \sum \omega_{Ai}} = 0.712$$

$$\theta_B^{(1)} = \frac{\sum x_i \omega_{Bi}}{10 \sum \omega_{Bi}} = \frac{\sum x_i (1 - \omega_{Ai})}{10 \sum (1 - \omega_{Ai})} = \frac{\sum x_i - \sum x_i \omega_{Ai}}{10(5 - \sum \omega_{Ai})}$$

$= 0.583$

After 10 steps converges to $\theta_A^{(10)} \approx 0.80$

$\theta_B^{(10)} \approx 0.52$

May converge to local maxima for non-concave functions.

but convergence is guaranteed.

GAUSSIAN MIXTURE MODELS

Feature Extraction from speech data

Multi-object tracking

Essentially, used when data is multi-modal, i.e. there are many "peaks" in the distribution.

$$P(\vec{x}) = \sum_{i=1}^{K} \phi_i \mathcal{N}(\vec{x} \mid \vec{\mu}_i, \Sigma_i)$$

($\vec{\mu}_i$: Mean vector; Σ_i: Covariance Matrix)

Multi-variate Gaussian Distribution.

$$\mathcal{N}(\vec{x} \mid \vec{\mu}_i, \Sigma_i) = \frac{1}{\sqrt{(2\pi)^K |\Sigma_i|}} \exp\left[-\frac{1}{2} (\vec{x}_i - \vec{\mu}_i)^T \Sigma_i^{-1} (\vec{x}_i - \vec{\mu}_i) \right]$$

$\sum \phi_i = 1$ → weights [ratio (fraction) of points in each category]

Parameters estimated using Expectation Maximization since analytic solution is usually not possible.

EM is guaranteed to converge.

EM Algo.

E-step : using $\phi_i, \vec{\mu}_i$ & Σ_i

estimate $p(C_i | x_j)$ for each data point

↳ K classes/components

M-step : using assigned C_i for each x_j, [weighted]

update $\phi_i, \vec{\mu}_i$ & Σ_i

See next page

Initialisation :

Randomly pick $\vec{\mu}_i$ from the given unlabelled data, x_j

Set variance estimates to variance of whole data

set $\phi_i = \frac{1}{K}$ $\forall i$

OR, use K-means to find initial parameter values (speeds up considerably)

MNIST - K = 10 components

Each image of size N×N is a vector of size N^2 of Bernoulli distributions (one per pixel) - PRML book

E-step

$$\gamma_{ji} = \frac{\phi_i \mathcal{N}(\vec{x}_j / \vec{\mu}_i, \Sigma_i)}{\sum_{i=1}^{K} \phi_i \mathcal{N}(\vec{x}_j / \vec{\mu}_i, \Sigma_i)} = P(C_i / \vec{x}_j)$$

M-step

$$\vec{\mu}_i = \frac{1}{N_i} \sum_{j=1}^{N} \gamma_{ji} \vec{x}_j$$

$$\Sigma_i = \frac{1}{N_i} \sum_{j=1}^{N} \gamma_{ji} (x_j - \vec{\mu}_j)(\vec{x}_j - \vec{\mu}_j)^T \rightarrow \text{updated value}$$

$$\phi_i = \frac{N_i}{N}, \qquad N_i = \sum_{j=1}^{N} \gamma_{ji} : \text{Effective no. of points in category 'i'}$$

Log-Likelihood

$$\ln p(X / \mu, \Sigma, \phi) = \sum_{j=1}^{N} \ln \left\{ \sum_{i=1}^{K} \phi_i \mathcal{N}(\vec{x}_j / \vec{\mu}_i, \Sigma_i) \right\}$$

Repeat the procen till LL or parameters converge.

EM required since analytical solution of parameters which maximise LL not available.

K-means	GMM
hard assignment of $\vec{x}_j$ to C_i, $\gamma_{ji} \in \{0, 1\}$	Soft assignment $0 \leq \gamma_{ji} \leq 1$
fast convergence even for high dimensional data	EM is slow
Assumes clusters to be spherical	Can handle ~~arbitrary~~ more complex shapes
Easier to interpret	Difficult to interpret