Tom Mitchell

Naive (simple) Bayes Classifier

Types of Classifier → Discriminative
→ Generative

Linear Non Linear

Discriminative → All you need to know is to which side of line this datapoint belong to. (Discriminates)

Generative → You need to know what is the underlying process which leads us to the data we are seeing. (eg → Bayesian)
Your task is to learn distribution that gave rise to this data.

Instead of knowing if this point lie here or there, we tend to know which part of feature space belongs to which class.

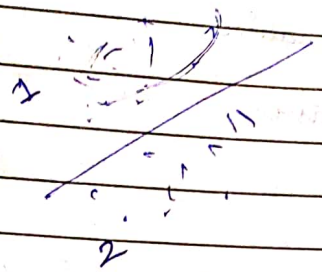
(In discriminative model you ~~know~~ predict label from features).

But in Generative → you see what is the probability of getting that feature vector for a class and then we deduce the probability of ~~getting~~ label given the feature vector.

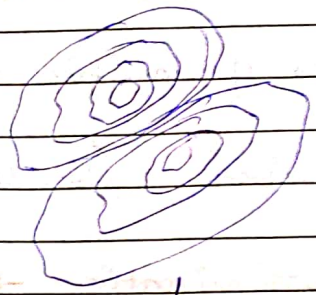
So, basically it is the probability of feature vector given that the label is something and then you find probability of label given feature vector $P(y/x)$.

$P(x/y)$ — x: feature vector, y: label

by → Dis



Generative



↓
Learning distribution
Two and their parameters

input is
$\langle x_1, x_2 \ldots x_d \rangle$
d-dimensional vector

→ [Naive Bayes] → Output is Joint Distribution $P(x_1, x_2 \ldots x_d, \text{label})$

Joint Distribution

Joint Probability → $P(A \cap B)$ where A and B are two events & joint Prob is the probability of occurrence of both events simultaneously. $(P(x,y))$

Joint ~~Probability~~ Distribution → when x and y are discrete random variables and joint distribution is $P(x,y)$ for every pair of values (x,y).

Marginal Distribution →

Marginalization refers to the process of removing the influence of one or more events from a probability. $g(x) = \sum_y f(x,y)$

Marginal of x ; ↳ Joint Pro. ⇒ y removed!

So, lets say we have 2 random variables x_1 & x_2.

↳ For each value of x_1, x_2 will have some value

x_1	x_2	$P(x_1, x_2)$
x_{11}	x_{21}	
	x_{22}	
x_{12}	x_{23}	
x_{13}	x_{21}	
x_{14}	x_{22}	
	x_{23}	

Why Joint Distribution?

→ We can ~~gate~~ get Conditionals
→ ——— " ——— Marginals
→ ——— " ——— every damn thing!

So, we know everything about data. Anything cannot be better than this.
But not possible to do it everytime. (Hard to implement)

How can you get Marginal from Joint?

Marginal ⇒ $\sum_{x_2} P(x_1, x_2) = P(x_1)$ [Summation of all over all x_2 from table]

Conditional ⇒ $P(x_1 / x_2 = \beta)$

↳ We freeze value of x_2 & see how the distribution of x_1 looks like.
i.e, subset of Joint Distribution table!

We can ans quest" like → (From ~~Good~~ Joint Distribut" table)

$P(x_1 < 5 \mid \text{Label} = 1) = ?$

$P(x_1 / x_2, \text{label}) = ?$

$P(\text{Label} / x_1, x_2) = ?$ → Only Quest" that a discriminative model can ans!.

Probability Axioms →

3 imp things → (that define Probability space)
- → Sample Space
- → Event Space
- → Probability function

① $0 \leq P(x) \leq 1$

② Sum of (all probabilities) Sample Space $P(S) = 1$.

③ $P(A \cup B) = P(A) + P(B) - P(A \cup B)$
for disjoint events → $P(A \cup B) = P(A) + P(B)$

$P(A) = P(A \cap B) + P(A \cap \bar{B})$

Conditional Probability →

$$P(A/B) = \frac{P(A \cap B)}{P(B)} \quad [P(B) \neq 0]$$

Chain Rule →

~~P(A/B) = P(A)~~

$P(A1, A2, A3) = P(A_1) \cdot P(A_2/A_1) \cdot P(A3/A1A2)$

Bayes Rule →

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

Given ~~a~~ the outcome ~~compute~~ determine the feature.

Application in health care, company for clinical trials →

They have → People with Disease = True Group A
→ —— " —— = false Group B

So you will apply test on A & know what cond" are (features) giving outcome true & you will select them.

~~So, P(A True / B false) ✗ This you want.~~
~~P(A True / B True)~~

conditional P(Predicted true / have disease) ✓
→ P(Pred true / dont have disease) ✓ You will know these kind of P first

Hence, for new people you will know.

Bayes → P(do this Person have disease / Pred true)

Bayes Rule - Interpretation

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

- $P(B|A)$ → Likelihood
- $P(A)$ → Prior
- $P(B)$ → Normalization factor
- $P(A|B)$ → Posterior Probability

Trickiest thing here → $P(A)$ prior probability before you run a test

So, Prior * Evidence from test = Posterior Prob.

Eg → Suppose 1% people have funny disease
∴ $P(f) = 0.01$ i.e prior.

→ Now Posterior = ① $P(f / \text{test is } \overset{+ve}{})$ = $P(f) * P(\overset{+ve}{} \text{ test} / f)$

- $P(f / \text{test is } +ve)$: knowing if there is funny disease or not given test results are +ve
- $P(+ve \text{ test} / f)$: Prob of having test results +ve given disease is there.
- $P(f)$: Prior that 1% have it.

② $P(!f / \text{test is } +ve) = P(!f) * P(+ve \text{ test} / !f)$

Is it easy to estimate Joint Distribution?

eg → if we have d-dimensional features.

$X = \langle x_1, x_2 \ldots x_d \rangle$

Each x_i takes k distinct values (of feature) for datapoints

and we have c distinct classes.

So how many enteries in Joint D Table?

k^d

and c classes ⇒ for each class k^d values are replicated. ⇒ $\boxed{c \cdot k^d}$

if $d = 100$, $k = 2$ & $c = 2$ ⇒ $2\,2^{100}$ i.e Huge

So here comes Naive Bayesian Assumption →

"The joint distribution of features conditioned on label becomes independent"

Once they become independent, we can write joint as the product of individual ones.

$$\therefore P(x_1 x_2 x_3 \ldots x_d \mid y) = \prod_i P(x_i \mid y)$$

↓

Product of Prob of each feature conditioned on label.

i.e, Product of individual conditional probability

∴ Now table size ⇒ ?

So, now to make table we only need $P(x_1) * P(x_2)$ i.e no. of values in x_1 * no. of values in x_2.

$\therefore$ kdc = 400. just that.

So,

$$P(y=y_k \mid x_1 \dots x_n) = \frac{P(y=y_k)\, P(x_1 \dots x_n / y=y_k)}{\sum_j P(y=y_j)\, P(x_1 \dots x_n / y=y_j)}$$

$P(x_1 \dots x_n / y=y_k)$ → $\pi_i P(x_i \mid y=y_k)$

$P(x_1 \dots x_n / y=y_j)$ → $\pi_j P(x_j / y=y_j)$

because →

So you will pick up that y for which numerator is highest ☺.

out of all possible combinations of x_i.

$$y^{new} = \arg\max_{y_k} P(y=y_k) \prod_i P(x_i^{new} / y=y_k)$$

→

ML Problem

Training Phase

You learn prob of occurrence of each of the classes & prob that for a given class a feature vector will take on a certain value.

Inference (Test) Phase

You will use these values of features & compute all y's and take on y which is maximum.

Lol! Knuckle?

Algo →

① First ~~Compare~~ $P(x/y)$ and $P(y)$

② Given new x :

$$P(y=1/x) = \frac{P(x/y=1)\,P(y=1)}{P(x)}$$

~~$P(x) = \sum_y P(x,y)$ • Marginal of x~~

and $P(y=0/x) = \frac{P(x/y=0)\,P(y=0)}{P(x)}$.

where $P(x) = \sum_y P(x,y) = P(x/y=1) \cdot P(y=1) + P(x/y=0) \cdot P(y=0)$

For Continuous Features → Gaussian Naive Bayes!

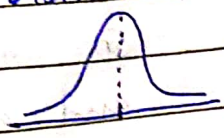
(Iris dataset)

We assume that they come from some distribution (Gaussian).

We will estimate mean & variance from that distribution.

eg → $\langle x_1, \ldots, x_n \rangle$ (let say $n=100$) data points coming from Gaussian

Suppose → IID → Independent & Identically distributed.



Now estimate μ & σ^2.

How will you do it?

Use fact IID to simplify joint.

Identical →
i.e these points are coming from a Gaussian with same μ & σ^2

Independent →
joint is the product of individual probabilities.

Joint $\Rightarrow \therefore P(x_1 \dots x_n / \text{class}) = P(x_1) * P(x_2) * P(x_3) * \dots P(x_n)$

$$P(x_1) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{\frac{-(x_1-\mu)^2}{2\sigma^2}}$$ [pdf of Gaussian]

$-\infty < x < \infty$ i.e continuous.

$$P(x_2) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{\frac{-(x_2-\mu)^2}{2\sigma^2}}$$

and so on!

$$\therefore P(x_1 \dots x_n / y) = \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^n e^{-\sum_{i=1}^{n}\left(\frac{(x_i-\mu)^2}{2\sigma^2}\right)}$$

↑ Likelihood function!

So this function we need to maximize
why?
Because we want Maximum likelihood.

So we will differentiate it wrt μ.
and equate it to zero.
This is how you will get best parameters for Gaussian.
⇒ we will get value of μ
⇒ Same mean is the maximum likelihood.
Same thing for Variance.

Continued → next Questions.

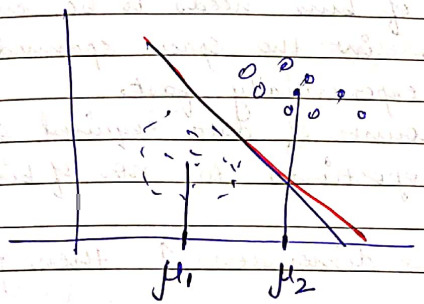
*) Features - Continuous → Gaussian NB.
*) " - Discrete → Multinomial NB
*) " - Binary (0/1) (Y/N) → Bernoulli NB.

What does a decision boundary looks like ?
(All Classes have →)

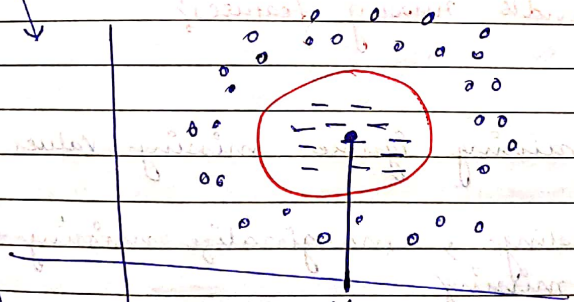
→ Different Means & Same Variance → Straight line / plane

→ Same mean & different Variance → Circle / ellipse

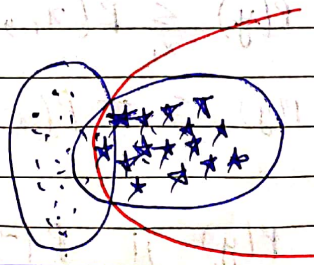
→ General Case → Parabolic Curve



But Same σ^2



μ Same μ But diff σ^2



General Case!

What is the impact of Dimensions on the classifier?

· $d = 100$ & $d = 1000$?

for large n → $P(x_1, x_2 \dots x_n / \text{class})$ takes exponential amount of time.
as dimensions ↑, data becomes more sparse.
⇒ more no. of bins needs to be created.
i.e no. of bins in the space spanned by features grows exponentially fast.
and thus, amount of data required should be ~~α~~ proportional to no. of bins

∴ if data ↓ & dimensions ↑ then problem!

Can it handle missing features?

Yes!

→ While ~~training~~ ignore missing values.

→ while testing, marginalize missing ~~α~~ feature.
Suppose x_1 missing

$$P(y / x_2 \dots x_d) \propto P(y)\, P(x_2, \dots x_d / y)$$

$$= P(y) \sum_{x_1} P(x_1, x_2 \dots x_d / y)$$

$$= P(y) \sum_{x_1} \prod_i P(x_i / y)$$

$$= P(y) \prod_{i=2,3 \dots d} P(x_i / y)$$

Can it perform Model Updating ?

Suppose you trained your model, now I give you new data. How will you update your model? Will you train it once again all over again?

eg → $f_w = \frac{\text{frequency of word}}{\text{document length}}$

Let say your model categorize it & make 3 bins →

$0 \leq f_w < 0.001$ very rare

$0.001 \leq f_w < 0.01$ rare

$0.01 \leq f_w <= 1$ frequent

Thus feature value for word w must be within these 3 intervals.

As new example comes you can update your parameters. by →

maintaining → ① counts over these bins for all documents.

d_1 ← very rare, rare, frequent

d_2 ←

d_3 ←

Prob of bin in subsequent document = $\frac{\text{this count}}{\text{Sum of counts}}$ /

=

Continued calc of μ & σ^2

Finding μ

Likelihood $L = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{-\left(\frac{1}{2\sigma^2}\right)\sum_{i=1}^{n}(x_i-\mu)^2}$

When we have exponents then we take log

$\therefore$ log likelihood →

$$\log L = \underbrace{n \log\left(\frac{1}{\sqrt{2\pi}\sigma}\right)}_{\text{const}} + \underbrace{\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}_{\text{max.}}$$

equate to 0

$$\frac{d}{d\mu}(\log L) = 0 + \frac{-1}{2\sigma^2}\frac{d}{d\mu}\sum_{i=1}^{n}(x_i-\mu)^2$$

$$0 = \frac{d}{d\mu}\sum_{i=1}^{n}\left(x_i^2+\mu^2-2\mu x_i\right)$$

$$\Rightarrow \frac{-1}{2\sigma^2}\frac{d}{d\mu}\sum_{i=1}^{n}(x_i-\mu)^2$$

$$\Rightarrow \frac{-1}{2\sigma^2} * 2 \sum_{i=1}^{n}(x_i-\mu)$$

Now equate logarithmic Likelihood to zero.

$$\Rightarrow \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\hat{\mu}) = 0$$

$$\Rightarrow \sum_{i=1}^{n} x_i - n\hat{\mu}$$

$$\therefore \boxed{\hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n}}$$

$\frac{d}{dx} \log_e x = \frac{1}{x}$

take second order derivative →

$\Rightarrow \frac{-n}{\sigma^2}$ which is -ve

which shows it is Maximum likelihood estimation.

finding σ → Differentiate wrt σ^2

$$\frac{d}{d\sigma^2}(\log L) = \frac{d}{d\sigma}\left(\left(\frac{n}{\sqrt{2\pi}\sigma}\right)^{n}\right)$$

$$\frac{d}{d\sigma^2}(\log L) = \frac{d(n)\log\left(\frac{1}{\sqrt{2\pi}\sigma}\right)}{d\sigma^2} + \frac{-1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

$$= \frac{-n}{2} \cdot \frac{1}{\sqrt{2\pi}\sigma} * \frac{1}{2\pi\sigma^2}$$

$$= \frac{+n}{2} \frac{1}{d\sigma^2} \log$$

$$= \frac{-n}{2}\frac{d}{d\sigma^2}\left(\log\left(\frac{1}{2\pi\sigma^2}\right) + \frac{-1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right)$$

$$= \frac{-n}{2} \cdot \frac{1}{2\pi\sigma^2} * \frac{1}{2\pi} \cdot \frac{1}{\sigma^4} + \frac{1}{2\sigma^4}$$

$$= \frac{-n}{2\sigma^2} -$$

$$= \frac{d}{d\sigma^2}\left(-\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 \right)$$

$$= \frac{\partial}{\partial a}\left(-\frac{n}{2}\log(2\pi a) - \frac{1}{2a}\sum_{i=1}^{n}(x_i-\mu)^2 \right)$$

$$= -\frac{n}{2}\times\frac{1}{2\pi a}\times 2\pi - \frac{\partial}{\partial a}\left(\frac{1}{2a}\right)\left(\sum \quad\right)$$

$$= -\frac{n}{2}\times\frac{1}{a} - \frac{1}{2}\left(\frac{x^{-1}}{-a^2}\right)\left(\quad\right)$$

$$= -\frac{n}{2a} + \frac{1}{2a^2}\left(\sum_{i=1}^{n}(x_i-\mu)^2\right)$$

$$\Rightarrow -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2 = 0$$

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}$$

So best estimate is variance!

Eg of Time Series data →

① → Audio, Speech

② → Whether forecasting or 10 days history of weather.

We have set of states.

Data point is a time-ordered set of states

One day can take one state out of 3

{ KK, PC, CR }

KK → Clear, PC → Partly cloudy, CR → chance of Rain

Sequence of States → Ordering [for each day]

KK, PC, CR, PC, PC, C, PC, PC, KK, KK [we cannot change its order]

① Audio Waveform →

What is state? (Freq? No) Which lang?

State Space = Can be very large

eg → tech talk, Sports talk, Movie talk etc.

③ Text Data →

State Space = All possible sentences of a language.

④ Word data →

State Space = Syllabels. → {अ, आ, ऊ, उ, इ, क, ख etc}

Random Variable → function from sample space to R [Real values]

Random Vector → funct^n from sample space to R^n [n-dim vect]

Random process → A random process is a collection of indexed random variables defined over a probability space.

{Triplet of} → Sample Space, Event Space, Pdf.

Where index is time in most cases (time series)

Markov chain → It is a special type of random process where probability of the present state depends only on the previous state. (1)

It must be "memory-less".

P(Current State / All history) = P(Current S / Previous) (2)

(same only)

Hidden Markov Model → It is a markov model which models a Markov chain with unobservable (hidden) states.

Markov Model →

Set of states must be finite and discrete

$S = \{ s_1, s_2 \cdots s_N \}$

When you go from one state to other, it will generate sequence of states.

$s_{i1}, s_{i2}, s_{i3} \cdots s_{ik}$

Markovian Property →

$$P(s_{ik} / s_{i1}, s_{i2} \cdots s_{ik-1}) = P(s_{ik} / s_{ik-1})$$

① Transition Probability → $a_{ij} = P(s_j / s_i)$

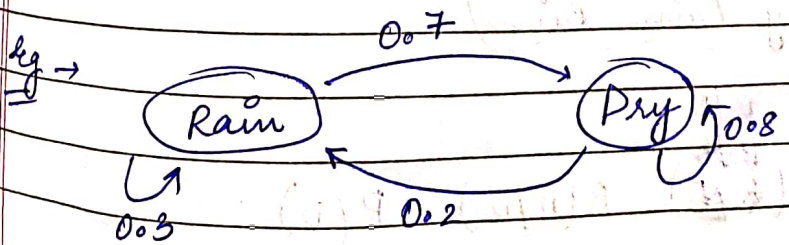
$i \rightarrow j$

It's a matrix of size N×N.

If entry = 0, we cannot directly reach there.

② Initial State / Vector / Probabilities →

$$\pi_i = P(s_i)$$



N = 2

	Rain	Dry
Rain	0.3	0.7
Dry	0.2	0.8

Transition Matrix.

But we should know initial Probabilities →

$P(\text{Rain}) = 0.4 \qquad P(\text{Dry}) = 0.6$

→ Joint Probability using Markovian Property →

$P(S_{i1}, S_{i2} \cdots S_{ik}) = P(S_{ik} / S_{i1}, S_{i2}, S_{i3} \cdots S_{ik-1}) \, P(S_{i1} \cdots S_{ik-1})$

$P(A \cap B) = P(A/B) * P(B)$ [Conditional Probability]

Now by using Markovian Property

$P(S_{i1} \cdots S_{ik}) = P(S_{ik} / S_{ik-1}) \cdot P(S_{i1} \cdots S_{ik-1})$

$= P(S_{ik} / S_{ik-1}) \cdot P(S_{ik-1} / S_{i1} \cdots S_{ik-2}) \cdot P(S_{i1} \cdots S_{ik-2})$

$\vdots$

Probability of Occurence of a Sequence

$= P(S_{ik} / S_{ik-1}) \cdot P(S_{ik-1} / S_{ik-2}) \cdot P(S_{ik-2} / S_{ik-3}) \cdots P(S_{i2} / S_{i1}) \, P(S_{i1})$

eg → P(Dry Dry Rain Rain)

$\Rightarrow P(R/R) \; P(R/D) \; P(D/D) \; P(D)$

$\Rightarrow 0.3 * 0.7 * 0.8 * 0.6$

Nothing hidden so far.

Hidden Markov Model →

States are not visible, but each state randomly generates one of M observations (or visible state)
True state are not known. $\{v_1, v_2 \dots v_N\}$
(IDK about temp but I can see you wearing sweater or not)

Eg → What gene is driving disease but we can see symptoms.
Text analysis S → 5 or s

Connectⁿ b/w visible states & sequence of states?

Visible & hidden states both are entirely different!

If im in S_{i1} state (hidden) what is the probability of going to v_{i1}, v_{i2} ...etc.

Eg →
If true thing was 'A' what is prob of writing 'S'?
Very low
" " " 'S' " " - - - 'S'?
Very high ↑ S_{ij} ↑ v_{ij}.

① Matrix of Transition Probabilities →

$A = (a_{ij})$, $a_{ij} = P(S_j / S_i)$

Going from i to j

② Matrix of Observation Probabilities →

$B = (b_i(v_m))$, $b_i(v_m) = P(v_m | s_i)$

Prob of getting observation v_m when you are in state s_i.

Size → (N × M) [N → true states (Hidden), M → visible observations]

③ Vector of Initial Probabilities →

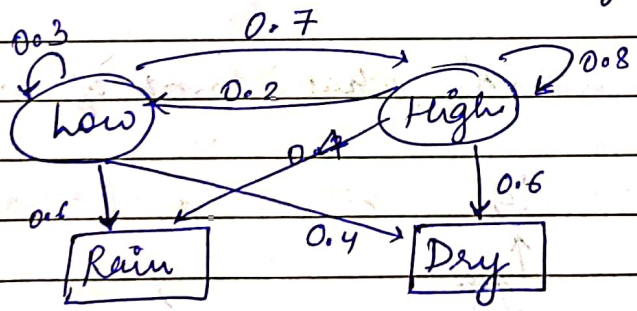
$\pi = (\pi_i)$, $\pi_i = P(s_i)$ [Prob of starting from s_i state]

Model → $M = (A, B, \pi)$ [Representation]

Eg →

Two observations : Rain & Dry.

Hidden : Low & High Atmospheric pressure!



Transitions are not b/w visible states!

A →

	L	H
L	0.3	0.7
H	0.2	0.8

B →

	Rain	Dry
L	0.6	0.4
H	0.4	0.6

π ⇒ P('low') = 0.4

P('High') = 0.6

Q → Sequence → Dry, Rain

D1 D2

Dry Rain

Low Low

$P(D|L) * P(R|L)$

$P(D,R) = P(\{D,R\},\{L,L\}) + P(\{D,R\},\{L,H\}) + P(\{D,R\},\{H,H\}) + P(\{D,R\},\{H,L\})$

$P(\{D,R\},\{L,L\}) = P(\{L,L\}) * P(\{D,R\}|\{L,L\})$

$= P(low)\, P(low|low) * P(D|L) * P(R|L)$

$= 0.4 * 0.3 * 0.4 * 0.6$

Do same for other 3 terms too.

Q → Computational Complexity of P(Observed Sequence) depends on?

If Sequence of states are T in length and 'c' are the no. of hidden states then Complexity = $O(c^2 T)$

~~There will be T no. of~~

Q→ What is more tolerable ?

① More Observable states
② more Hidden states
③ length of Observed state

Learning the HMM ?

We will need to learn A, B and π from data
Modeling limitations → first we need to fix no. of hidden states, Observed states etc. then find A, B & π

[Eg → in linear regression we fix model $y = \theta_0 + \theta_1 x + \theta_2 x^2$ then we find θ_0, θ_1 & θ_2.]

Maximize $P(O/M)$

↳ you'll take initial model and then model will give size to observation & you maximize it.

No closed form soln available

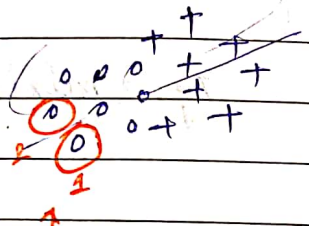
Baum - Welch Algo → Iterative method which gives you local maximum (& not global)
It is based on EM Algo.
↓
Expectation Maximization.

EM Algo → Given set of points knowing they are coming from 2 or more Gaussians.
Now you need to figure out parameters of those Gaussians.

EM Algo →

K means

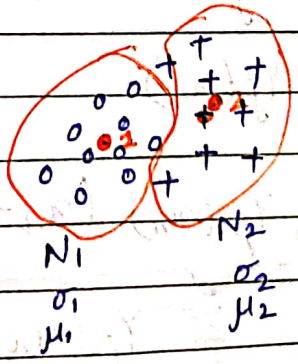
Given GMM → Gaussian Mixture Model (Two Gaussians)



1.) Select 2 representatives of 2 classes.

2.) K-Means → Representatives keeps on moving & getting refined.

At convergence,



- one centroid.

Each of the representatives will be assigned to one

3.) We are trying to estimate parameters through EM.
Now you got N_1 points of class 1 & N_2 points of class 2.
Now find μ & σ. that's it!
Through logarithmic likelihood.

E step

4) Now, Evaluate responsibilities using the current parameter values.

Assuming that these are parameters you'll compute what is the prob that the point was generated by the Gaussian.

Nothing but weightage. If point generated by (μ_k, Σ_k) gaussian then high else low weightage.

$\Sigma_k \rightarrow$ covariance
$\mu_k \rightarrow$ mean

$$\gamma_j(x) = \frac{\pi_k \, N(x/(\mu_k, \Sigma_k))}{\sum_{j=1}^{K} \pi_j^o \, N(x/\mu_j^o, \Sigma_j^o)}$$

Initial →

$\mu_j^o \rightarrow$ Mean $\quad \Sigma_j^o \rightarrow$ Covariance $\quad \pi_j^o \rightarrow$ Mixing coeff.

M step

5) Re-estimate parameters using current responsibilities.

$$\rightarrow \mu_j^o = \frac{\sum_{n=1}^{N} \gamma_j^o(x_n) \, x_n}{\sum_{n=1}^{N} \gamma_j^o(x_n)}$$

$$\rightarrow \Sigma_j^o = \frac{\sum_{n=1}^{N} \gamma_j^o(x_n) \, (x_n - \mu_j^o)(x_n - \mu_j^o)^T}{\sum_{n=1}^{N} \gamma_j^o(x_n)}$$

$$\rightarrow \pi_j^o = \frac{1}{N} \sum_{n=1}^{N} \gamma_j^o(x_n)$$

6) Evaluate log likelihood

$$\ln p(X \mid \mu, \Sigma, \pi) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \, N(x_n \mid \mu_k, \Sigma_k) \right\}$$

If no convergence, return to step 2.