17/9/18

Lec-12

* Parameter estimation:-

. classifiers were built on the assumption that we know the distro.

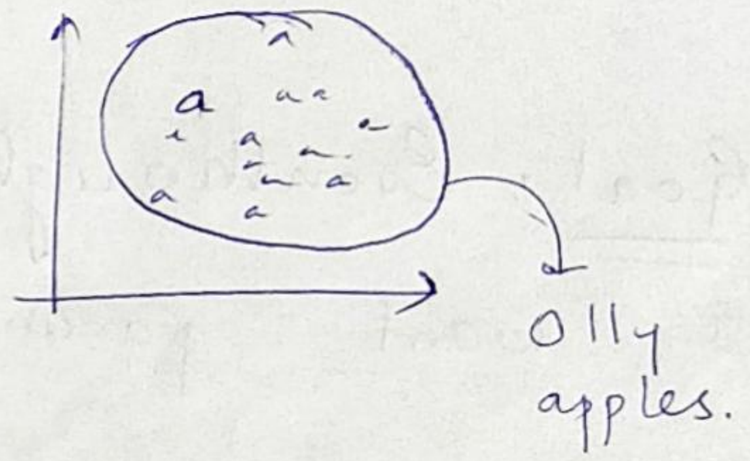
But irl, we don't exactly know the distro.

and

In ~~regressn~~ linear classifier / regression, we found how to update w (which is learning).

So, given a set of data points, how sure am I that the model works?

A Parameter Estimation.



. Before learning, we assume that the distro is normal with elliptical contours.

. If we want to fit a model to a data, we want to find parameters that will help me in predicting new points

Some terms

Population:

Every apple I could observe (if world = ∞)

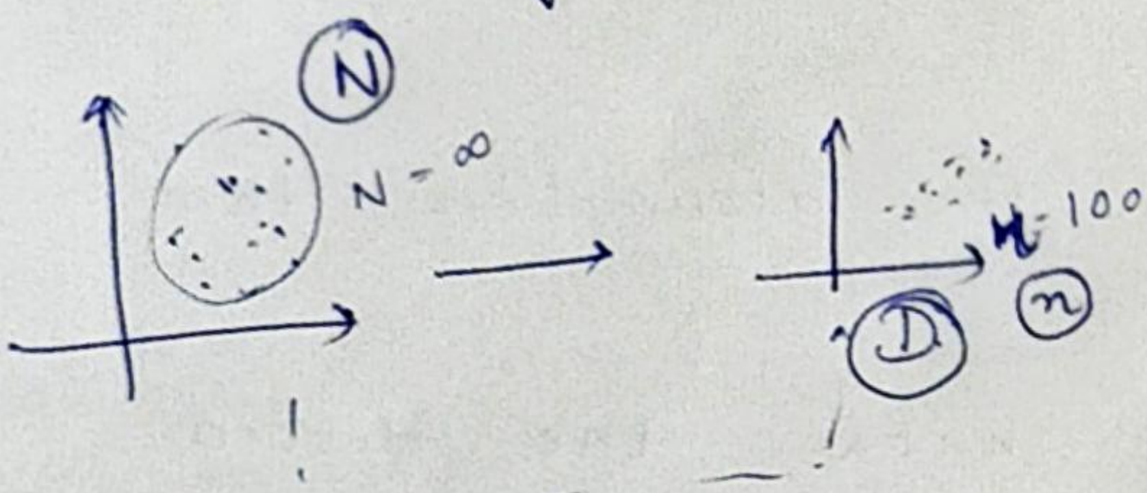
Every apple in Gachibowli (if world = Gachibowli)

Prob: We can never know the whole population.

Sampling:- Take a set of random points from the entire population. It is important, so that, we have no biases.

→ The population may have any distribution

but sampling is uniform. So, the sampled distro also follows similar distro as original one



So sample + Params → parameters

In normal distro:

Params: μ, σ + Sample ⟶ $\mu = \begin{bmatrix} a \\ b \end{bmatrix}$, $\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}$

Goal: Eventhough I am looking at a sample. I want parameters over the sample

---

Two approaches:-

1> Compute $\theta : [\theta_1, \theta_2, -- \theta_n]$

$\omega : [\omega_1, \omega_2 - \omega_n] \rightarrow$ line

Compute θ that explains D.

Maximum likelihood estimate: (MLE)

Find θ so, we observe max of D.

→ we can never find θ so $\theta \sim \theta'$

$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\ p(D|\theta)$

MLE

If I predict some, μ and Σ, how good is the normal distro.

2> Bayes' style: [Maximum Posteriori Estimate]

Use Bayes theorem to find θ

$$p(D|\theta) = \frac{P(\theta|D) \times P(D)}{p(\theta)}$$

$p(\theta) \rightarrow$ Continous

$\Rightarrow \hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}} [P(\theta|D) \times P(D)]$

- If $p(\theta)$ = constant ⇒ No prior info about θ ⇒ MLE == MAP
- Bayesian gives nothing with certainity

So $\hat{\theta}$ is also a density fn, with some distro

MLE :-

$$\underset{\theta}{\text{argmax}}\ P(D|\theta) = \prod_{k=1}^{n} p(x_k|\theta)$$

If samples are independent

Taking log won't affect

$$\Rightarrow \hat{\theta} = \underset{\theta}{\text{argmax}}\ \ln\left(\prod_{k=1}^{n} p(x_k|\theta)\right)$$

$$\hat{\theta} = \underset{\theta}{\text{argmax}} \sum_{k=1}^{n} \ln\{p(x_k|\theta)\}$$

To maximise, find $\nabla_\theta l = 0$

$$\nabla_\theta l = \sum_{k=1}^{n} \nabla_\theta \ln\{p(x_k|\theta)\} = 0$$

So each comp has to be zero.

Case1 : $N(\mu, \Sigma)$: Σ is known

Likelihood : $\ln p(x_k|\mu) = \ln\left(\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}\right) = (2\pi)^{-d/2}|\Sigma|^{-1/2}$ constant

$$= \ln(\text{stuff}) + -\frac{1}{2}(x_k-\mu)^T\Sigma^{-1}(x_k-\mu)$$

$\Sigma \rightarrow$ Covariance

$$\nabla_\mu \ln p(x_k|\mu) = \frac{\partial}{\partial \mu}\left(-\frac{1}{2}(x_k-\mu)^T\Sigma^{-1}(x_k-\mu)\right)$$

$$= -\frac{1}{2} \times 2\Sigma^{-1}(x_k-\mu) \times -1$$

↳ substitue $\mu = \hat{\mu}$ as we know the best estimate = $\hat{\mu}$

satisfies $\frac{\partial}{\partial \mu}\ln = 0$

$$\Rightarrow \Sigma^{-1}\left(\sum_{k=1}^{n}(x_k - \hat{\mu})\right) = 0$$ ↳ Removing Σ

$\hat{\mu}_{MLE}$

$$\sum_{k=1}^{n} x_k - \sum_{k=1}^{n}\hat{\mu} = 0$$

$$\hat{\mu}_{MLE} = \frac{\sum_{k=1}^{n} x_k}{n}$$

Average of samples, is the best estimate of μ

write as

Step 2

$N(\mu, \sigma^2)$ ▩ $N(\theta_1, \theta_2)$

$\mu \equiv \theta_1, \quad \sigma^2 \equiv \theta_2$

While ∇_μ, $\Rightarrow$ Same as Step 1: Assume const $\sigma^2 \Rightarrow \hat{\mu} =$

∇_{σ^2}

$$\nabla_\theta \ln(x_k/\theta) = \begin{bmatrix} \nabla\theta_1 \\ \nabla\theta_2 \end{bmatrix} \ln(x_k/\theta) = \begin{bmatrix} \text{same as step 1} \\ \hline \\ \downarrow \\ \text{need to find} \end{bmatrix}$$

$$\nabla_{\theta_2} = \nabla\left(\ln\left[\frac{1}{\sqrt{2\pi\theta_2}}\right] - \ln\left[\frac{(x_k-\theta_1)^2}{2\theta_2}\right]\right)$$

$$= \nabla\left[-\frac{1}{2}\ln\theta_2 - \frac{(x_k-\theta_1)^2}{2\theta_2}\right]$$

$$\nabla\theta_2 = \frac{-1}{2\theta_2} + \frac{(x_k-\theta_1)^2}{2\theta_2^2}$$

$$\Rightarrow \sum_{k=1}^{\hat{n}} \frac{-1}{2\theta_2} + \frac{(x_k-\theta_1)^2}{2\theta_2^2} = \Big|_{\text{at } \theta_2 = \hat{\theta}} = 0$$

$$\Rightarrow \sum_{k=1}^{n} \frac{-1}{\hat{2\theta_2}} + \frac{(x_k-\theta_1)^2}{\hat{2\theta_2^2}} = 0$$

$$\frac{-n}{2\theta_2} + \sum_{k=1}^{\hat{n}} \frac{(x_k-\theta_1)^2}{2\theta_2^2} = 0$$

$$\Rightarrow \frac{1}{2\hat{\theta}_2} \left\{ \sum_{k=1}^{n} (x_k-\theta_1)^2 \right\} = \frac{n}{2\theta}$$

$$\Rightarrow \sum_{k=1}^{n} (x_k-\theta_1)^2 = \frac{n}{x} \times x\hat{\theta}_2$$

$$\hat{\theta}_2 = \sum_{k=1}^{n} \frac{(x_k-\theta_1)^2}{n} \Rightarrow \sigma_2^2 = \sum_{k=1}^{N} \frac{(x_k-\theta_1)^2}{n}$$

Hence, we proved $\left\{ \begin{matrix} \mu \\ \sigma^2 \end{matrix} \right\}$ of popul for a normal distro $\left\{ \begin{matrix} \mu \\ \sigma^2 \end{matrix} \right.$ of sample

11ly for a multivariate quassian ⇒ $\bar{\mu}$, Σ

As sampling changes ⇒ $\hat{\mu}$ changes.

What will happen if many $\hat{\mu}$ are averaged.

Will $\text{mean}(\hat{\mu}) = \mu_{global}$?

i.e will $\boxed{\mu_{global} = E(\hat{\mu})}$ → unbiased

If small sample size ⇒ $\hat{\mu}$ jumps a lot.

If large ⇒ μ remains same

and $\text{Var}(\hat{\mu}) \propto \frac{1}{n}$

as $n \to \infty$, Var → 0, so it peaks

Will $\sigma^2_{global} = E(\hat{\sigma}^2)$?

No, $\frac{n}{H}$ is not

Infact $\boxed{E(\hat{\sigma}^2) = \frac{n-1}{n}\sigma^2_{glob}}$ → Biased ⇒ $E(\hat{\sigma}^2) < \sigma^2_{glo}$ always

So MLE $\hat{\sigma}^2$ is a biased estimate, [as we estimate smaller]

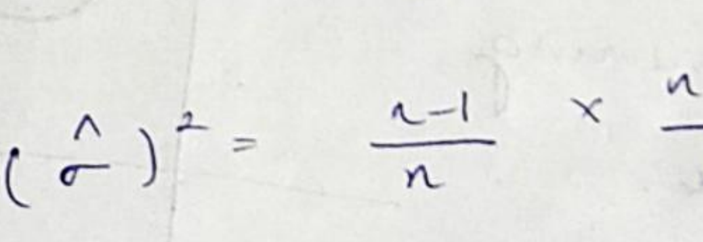
$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{k=1}(x_k - \mu)^2$$

Then $E(\hat{\sigma})^2 = \frac{n-1}{n} \times \frac{n-1}{n} \times \sigma$

MAP estimate: → There is D, H, S

$P(\theta)$ is a distro and so, has μ_0, σ_0

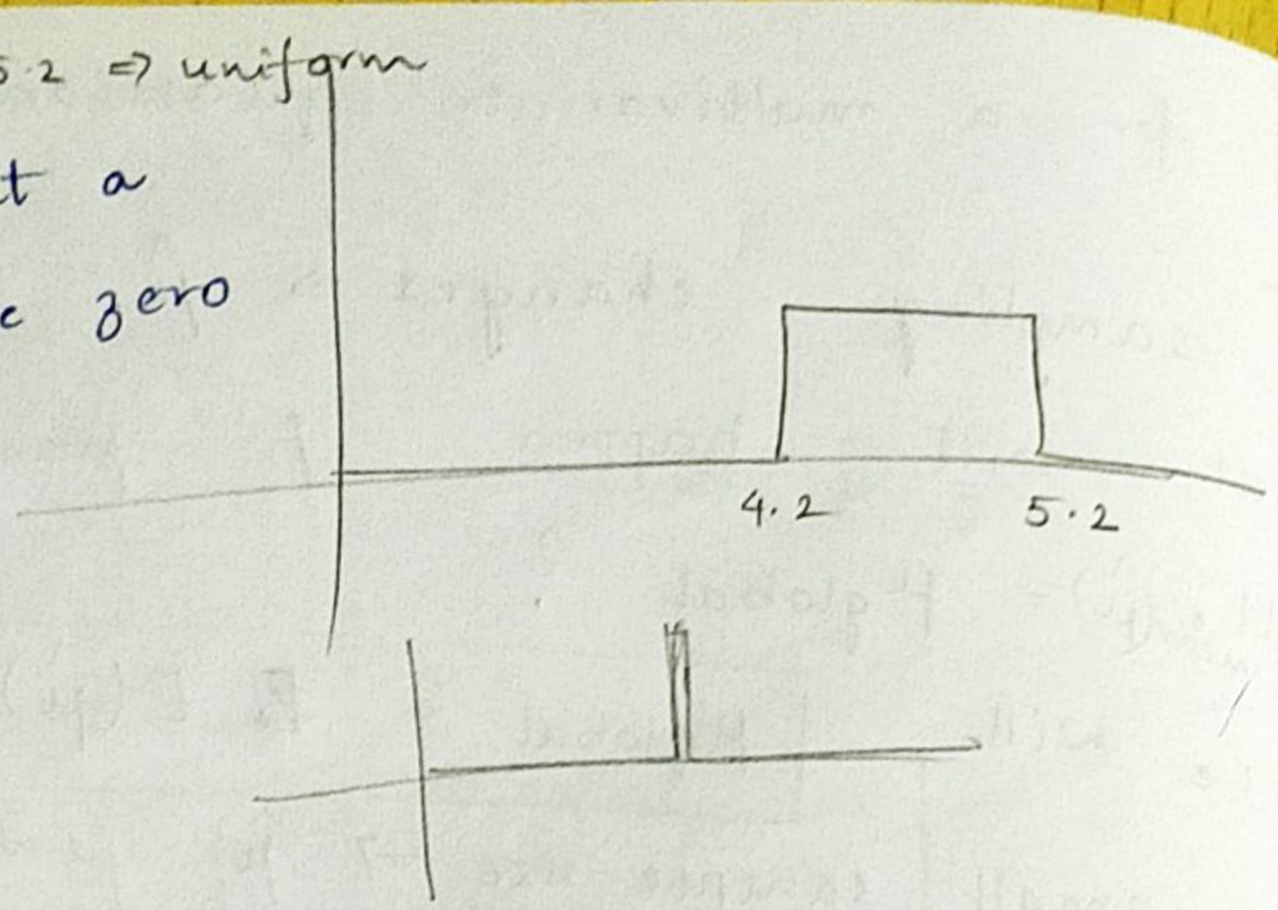
Only one θ, ⇒ μ



Bayesian learning: Process of updating beliefs

What if $P(\theta) \rightarrow 4.2$ to $5.2 \Rightarrow$ uniform

If prior density is zero at a point, it will never be zero posterior



20/9/18

Lec - 13

- Bayesian Parameter Estimation:-

$$p(\omega_i|x) = \frac{p(x|\omega_i)P(\omega_i)}{p(x)}$$ ← Bayes rule

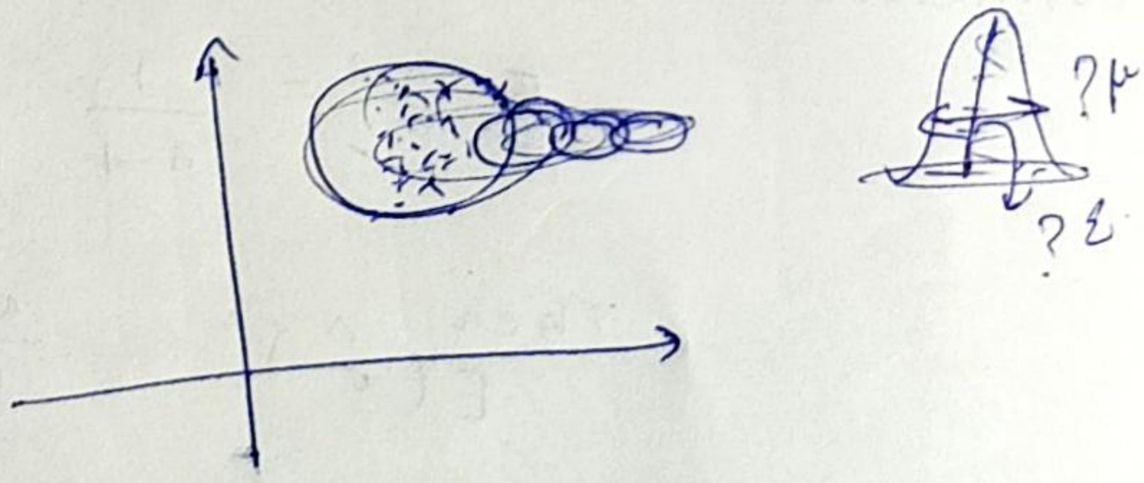
Generative model

- Now, we need to know $P(x|\omega_i)$.

- If no assumptions are made on the probility distro, we can't use it.

- Many density estimate functions, need to know the form of the distro

- Here we are assuming the distro is Normal distro.



[So, at first we needed ∞ params, (no idea of density at all) to normal distro μ Σ μ_1 μ_2 a_{11} $a_{12}=a_{21}$ a_{22} 5 params

- A rule of thumb is you need to know atleast 10 m points for m params.

⇒ 2D guassian

↳ Use atleast 50 points

- Classifier doesnot work properly
  1) fn - form is not valid
  2) fn - form is good, but samples are so-less that improper params

- So when, we say Bayesian estimation is best, we skip the above problems.

Let training data = $\{x_1, x_2, \dots x_n\}$

$p(x) \sim N(\mu, \sigma^2)$

1. $D/\theta = ?/\theta$?
2. Training data

$$p(\theta) = \underset{\theta}{\operatorname{argmax}}\ p(D/\theta)$$

$$= \underset{\theta}{\operatorname{argmax}} \prod_{k=1}^{n} p(x_k|\theta)$$

$$\hat{\mu}_{MLE} = \frac{1}{n}\sum_{k=1}^{n} x_k \quad , \quad \hat{\sigma}_{ME} = \frac{1}{n}\sum (x_k - \hat{\mu})^2$$

MAP: estimate

$$\hat{\theta}_{map} = \underset{\theta}{\operatorname{argmax}}\ p(D|\theta) \times p(\theta)$$ → Is a density now.

$p(\theta|D) =$

$p_\theta(x|\omega_i) \sim N(\mu, \sigma^2)$

↓

$p(\mu)$ is a density fn.

↳ μ is a random variable.

$\phi(\mu) \sim N(\mu_0, \sigma_0^2)$

$$p(\mu|D) = \frac{p(D|\mu) \times p(\mu)}{p(D)}$$

$p(\mu) \sim N(\mu_0, \sigma_0^2)$ → Uncertainty in guess

↳ Our best guess

An example

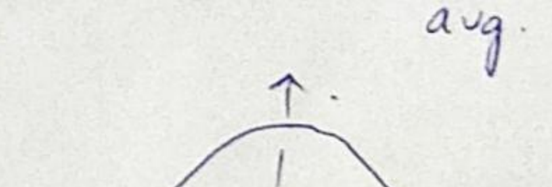
$p_\theta(x|\omega_i) \sim N(\mu, \sigma^2)$

$250, \sigma = 50$

$p(\mu) \sim N(\mu_0, \sigma_0^2)$

↓

Uncertainty in avg. wt



↓

$p(\mu|D) = p(D|\mu) \times p(\mu)$

μ_0, σ_0^2 won't affect, as we increase samples : $P(\mu)$ will go to

correct μ

Case 1 $p(\mu|D)$. $p(x|\mu) \sim N(\mu, \sigma^2)$

$p(\mu) \sim N(\mu_0, \sigma_0^2)$

$$p(\mu|D) = \frac{p(D|\mu) \cdot p(\mu)}{\int p(D|\mu) \times p(\mu)\, d\mu} \longrightarrow \text{Const} = \alpha$$

$$p(\mu|D) = \alpha \prod_{k=1}^{n} p(x_k|\mu) \times p(\mu)$$

$p(x_k|\mu)$: A normal here

$p(\mu)$: A normal here

Posterior is also a normal density

- Reproducing Density: If prior x distro a ⟶ posterior distro a

  ↳ Both are same

⇒ The distro a is

$$p(\mu|D) = \alpha \prod_{k=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(x_k - \mu/\sigma)^2} \times \frac{1}{\sqrt{2\pi}\sigma_0} e^{-\frac{1}{2}\left(\frac{\mu-\mu_0}{\sigma_0}\right)^2}$$

$$= \alpha \prod_{k=1}^{n} \frac{1}{\sqrt{2\pi}\sqrt{\sigma\sigma_0}} \times e^{-\frac{1}{2}\left\{\left(\frac{x_k-\mu}{\sigma}\right)^2 - \left(\frac{\mu-\mu_0}{\sigma_0}\right)^2\right\}}$$

$$p(\mu|D) = \frac{1}{\sqrt{2\pi}\sigma_n} e^{-\frac{1}{2}\left(\frac{\mu-\mu_n}{\sigma_n}\right)^2}$$

So, before observing we had $p(\mu) \sim N(\mu, \sigma^2)$; now

after (posteriori) $p(\mu|D) \sim N(\mu_n, \sigma_n^2)$

where $\mu_n = \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2} \hat{\mu}_n + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2} \mu_0$

$\hat{\mu}_n$: Average of samples we observed [MLE estimate]

μ_0 → Before observing

$\hat{\mu}_n$ → Only observing

Bayes says : use weighted μ_0 and $\hat{\mu}_n$

μ_n as $n \to \infty$, $\mu_n = \hat{\mu}_n$

and if higher $\sigma \Rightarrow$ higher n is needed

meaning, if varied data $\Rightarrow$ more samples are required

$$\sigma_n^2 = \frac{\sigma_0^2 \sigma^2}{n\sigma_0^2 + \sigma^2} \quad \text{as } n \to \infty, \ \sigma_n^2 = 0.$$

. So as n changes from $\mu_n \to \hat{\mu}_n$ and $\hat{\sigma}_n \to 0$.

$\Rightarrow$ Bayesian learning $\uparrow$ $\Rightarrow$

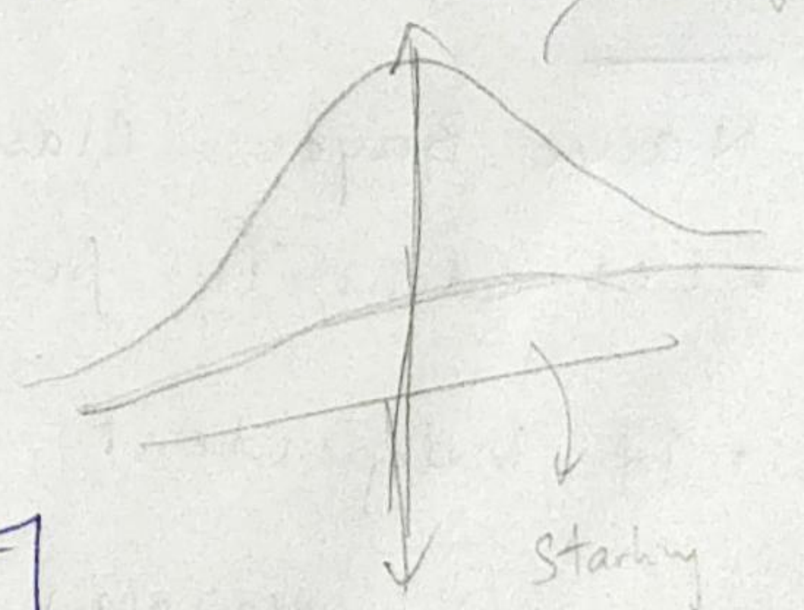
. Recursively: we can do the same, everytime we get a new sample.

$$p(D^n|\theta) = p(x_n|\theta) \times p(D^{n-1}|\theta)$$

$$p(x|D) = \int p(x|\mu) \times p(\mu|D)\, d\mu \quad \text{can be checked by integration}$$

$$\sim N(\mu_n, \sigma^2 + \sigma_n^2)$$

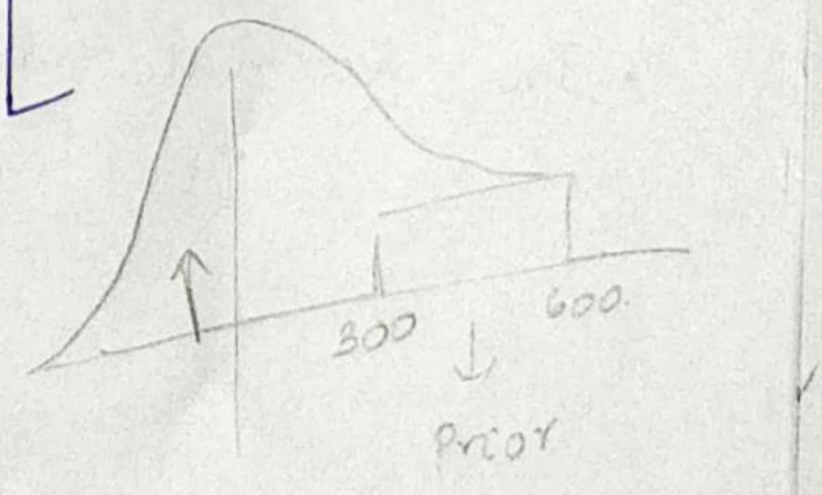
as $n \to \infty$, $p(x|D) \sim N(\mu_n, \sigma^2)$.



Whatever we think is impossible in prior, $\Rightarrow$ it will never

. If uniform density, μ_{act} is outside, you can never reach. μ_{act}

So, use some distro with range $(-\infty, \infty)$



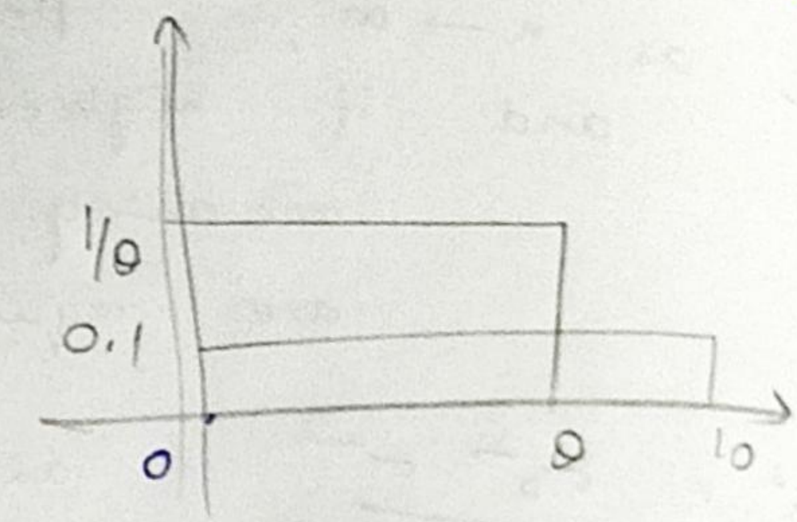
TB example

$D = \{4, 7, 2, 8\}$

If $\theta = 5$

$\Rightarrow P(D) = 0.$

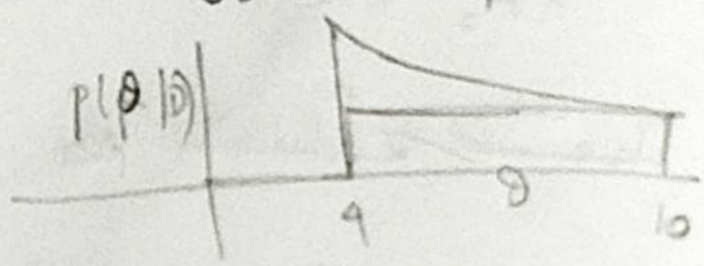
$\theta \rightarrow$ max $\rightarrow 8.$



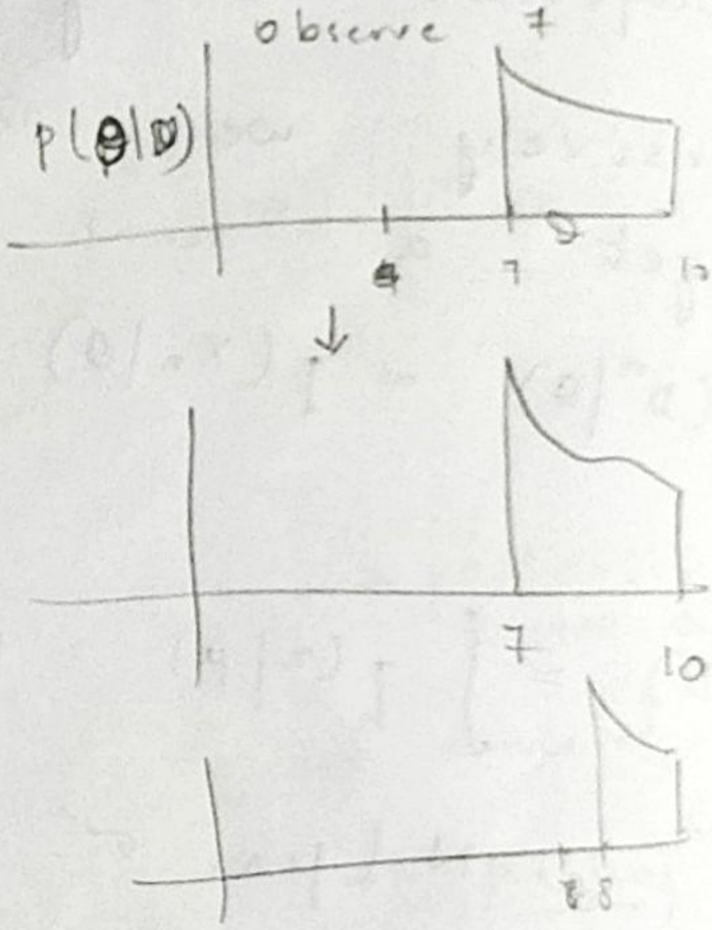
If $\theta = 10,$

$P(\theta \leq 4) = 0.$

observe 4



observe 7



Naive Bayes Classifiers

- Let density fn be normal.
- If independent features//

$$p(X) = p(x_1)p(x_2) \cdots p(x_d) \quad \mu = \begin{bmatrix} \\ \\ \end{bmatrix}_{d \times 1}, \quad \Sigma = \begin{bmatrix} & & 0 \\ & \ddots & \\ 0 & & \end{bmatrix}_{d \times d}$$

- In normal distro:
  It implies off diagonal