Why Probability in ML?

→ Bayesian setup

→ Entire Premise of ML is based on the notion of data coming from some distribution

→ classifiers etc. which Bayes's rule

Infinite stream of data

$x_1, x_2, \ldots, x, \ldots$

↓

Choose of sample of Data ⟸ Training Data ↷ Some probability

↓

Train your Model on this data

Apply Model to make predictions about unseen data

Test data

Most material in these slides/notes is taken with permission form Prof. Mukesh Tiwari , DA-IICT

What?

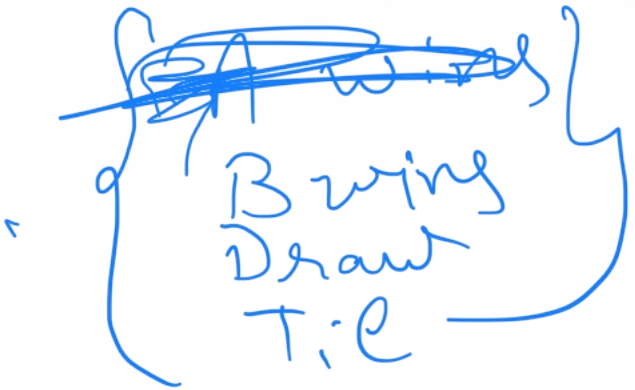
Quantifies the likelihood of the occurance of an event.

Why?

- Model real world uncertainty in terms of simplified random processes.
- Make sense of data from real world: Family of models for many complex systems.

Sample Space and Events

- **Sample space** (Ω): set of all possible outcomes of a random experiment.
  - flipping of a coin $\Omega = \{H, T\}$
  - rolling of a dice $\Omega = \{1, 2, 3, 4, 5, 6\}$
- **Sample point** ω: Each element of Ω, e.g. $\Omega = \{H, T\}$, sample point H or T (one realization of the random experiment)
- **Event:** any subset E of Ω, e.g. $\Omega = \{H, T\}$, event: $\{H\}, \{T\}, \{H, T\}, \phi$
- **Probability**: For each subset $E \in \Omega$, there is a probability $P(E)$ (probability is defined on events).



Axioms of Probability

Properties of probability measure:

- $P(\Omega) = 1$
- $0 \leq P(E) \leq 1.$
- For any sequence of mutually exclusive events A, B $(A \cap B = \phi)$

$$P(A \cup B) = P(A) + P(B)$$

Negation / Impossibility

$P(A) = 1 \Rightarrow$ Certain

$A \cap B = \phi$

$\cup$ – oring

$\cap$ – anding

A wins

B wins

$P(A \cap B) = 0$

$P(A \cup B) = P(A) + P(B)$

A large number of important mathematical results follow from these axioms:

- Since $E \cup E^c = \Omega$, and $E \cap E^c = \phi$, $P(E^c) = 1 - P(E)$
- $P(\phi) = 0$
- $P(E \cup F) = P(E) + P(F) - P(E \cap F)$
- If $E \subseteq F$, then $P(E) \leq P(F)$

$P[E^c) = 1 - P(E)$

E: P (Students > 50 marks)

E: P (Student > 60 marks)

It rains

It doesnot

P (It rains) = 0.7

P (It doesnot rain) = 0.3

Conditional Probability

E = Roads are wet
F = There is Dew
F has already occured

$$P(E|F) = \frac{P(E \cap F)}{P(F)}$$

$$P(E \cap F) = P(E|F)P(F)$$

Law of total probability

F1 Either you have B-Tech with 8.5 OA
F2 BA in literature

Let $F_1 \cdots F_n$ be disjoint events that form a partition of the sample space. Then for any event E

E = I call you for an interview

$$P(E) = \sum_{i=1}^{n} P(F_i)P(E|F_i)$$ **(Total Probability Theorem)**

$$P(E) = \sum_{i=1}^{n} P(F_i) P(E|F_i)$$

Baye's Formula

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

B, B^c
B, B'

Notion of Posterior

θ - Model
x - data

$$p(\theta|x) = \frac{p(x|\theta)}{p(x)} p(\theta)$$

Likelihood
Prior we used to generate data

$P(\theta|x) = P(\text{Data}|\text{Model}) \cdot P(\text{Model})$

observed data

Posterior

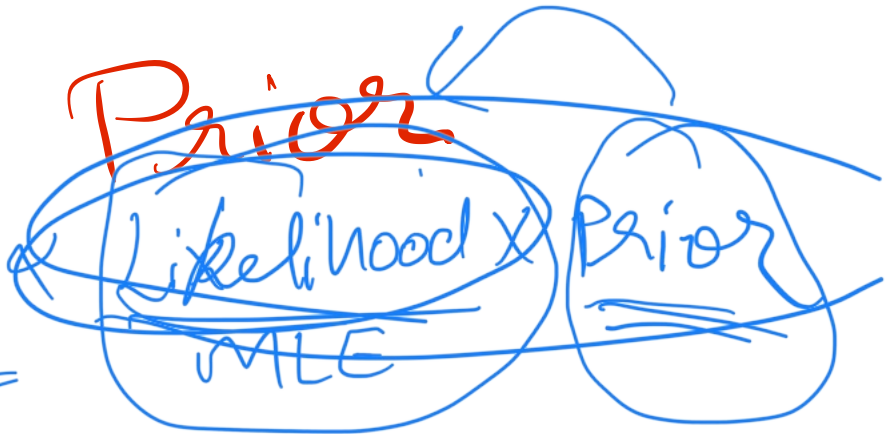
Overfitting

Posterior

MAP

Likelihood

Prior



Coin whether it is a fair or coin not

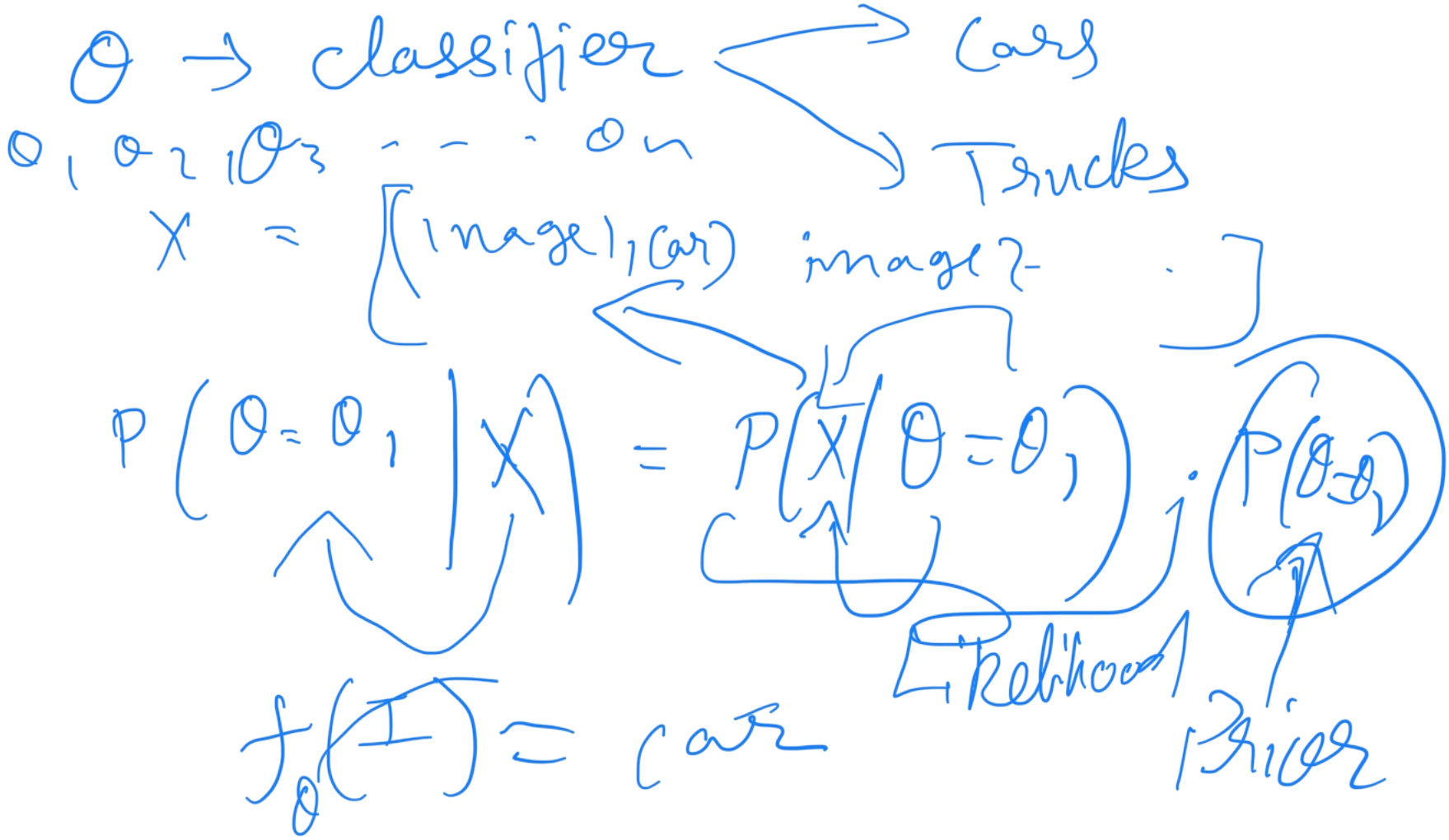
$P(H) = P(T) = 0.5$

$\theta = P(H)$ $P(H) = 0.5$

$x = [T\,H\,H\,T\,H\,T\,H\,H\,T\,T\,T\,H]$

$P(\theta|x) \propto P(x|\theta) \cdot P(\theta)$

$P(H) = 0.65$ Fair coin


θ → classifier
Cars
Trucks
θ1 θ2 θ3 - - - θn
X = [image1 (car) image2 -]
P(θ=θ1 | X) = P(X | θ=θ1) · P(θ=θ1)
Likelihood
Prior
fθ(I) = car

Posterior ∝ Likelihood × Prior

EM Algorithms

Gradient Descent

Prior

$P(\theta) = 0.5$

$Pr(H) = 0.5$

[H H T T T T H H H H]

$P(x|\theta_1)$

0.6

P = $P(0.5)$ 0.5

$P(H) = 0.5$

0.37

$P(H = 0.65)$

Random Variable

- It frequently occurs that in performing an experiment we are interested in some function of the outcome rather than the outcome itself. These real valued functions defined on the sample space are known as *random variables*.
  - Discrete random variable: takes either finite or a countably infinite number of values.
  - Continuous random variable: continuum of possible values.

Example

- Outcome of tossing two coins ($\Omega = \{T, T\}, \{T, H\}, \{H, T\}, \{H, H\}$).
  - Define X as the number of heads: $X = 0, 1, 2$
  - Probability $P\{X = x\}$: $P\{X = 0\} = 1/4$, $P\{X = 1\} = 1/2$, $P\{X = 2\} = 1/4$
- Probability of the events are denoted as $P(\cdot)$
- Probability of the random variables are denoted as $P\{\cdot\}$

Convention: X, Y, Z - random variables, x, y, z - real numbers.

X = # of Heads in coin toss

X can takes {0, 1, ...

Expectation of a random variable

The expectation or the average value (mean, weighted average, center of mass) of a random variable is:

$$E[X] = \sum_x x p_X(x) \qquad \text{(discrete)}$$

$$E[X] = \int_{-\infty}^{\infty} dx \; x \; f_X(x) \qquad \text{(continuous)}$$

Variance

$$\mathrm{Var}[X] = \sigma_X^2 = E[(X - E[X])^2]$$

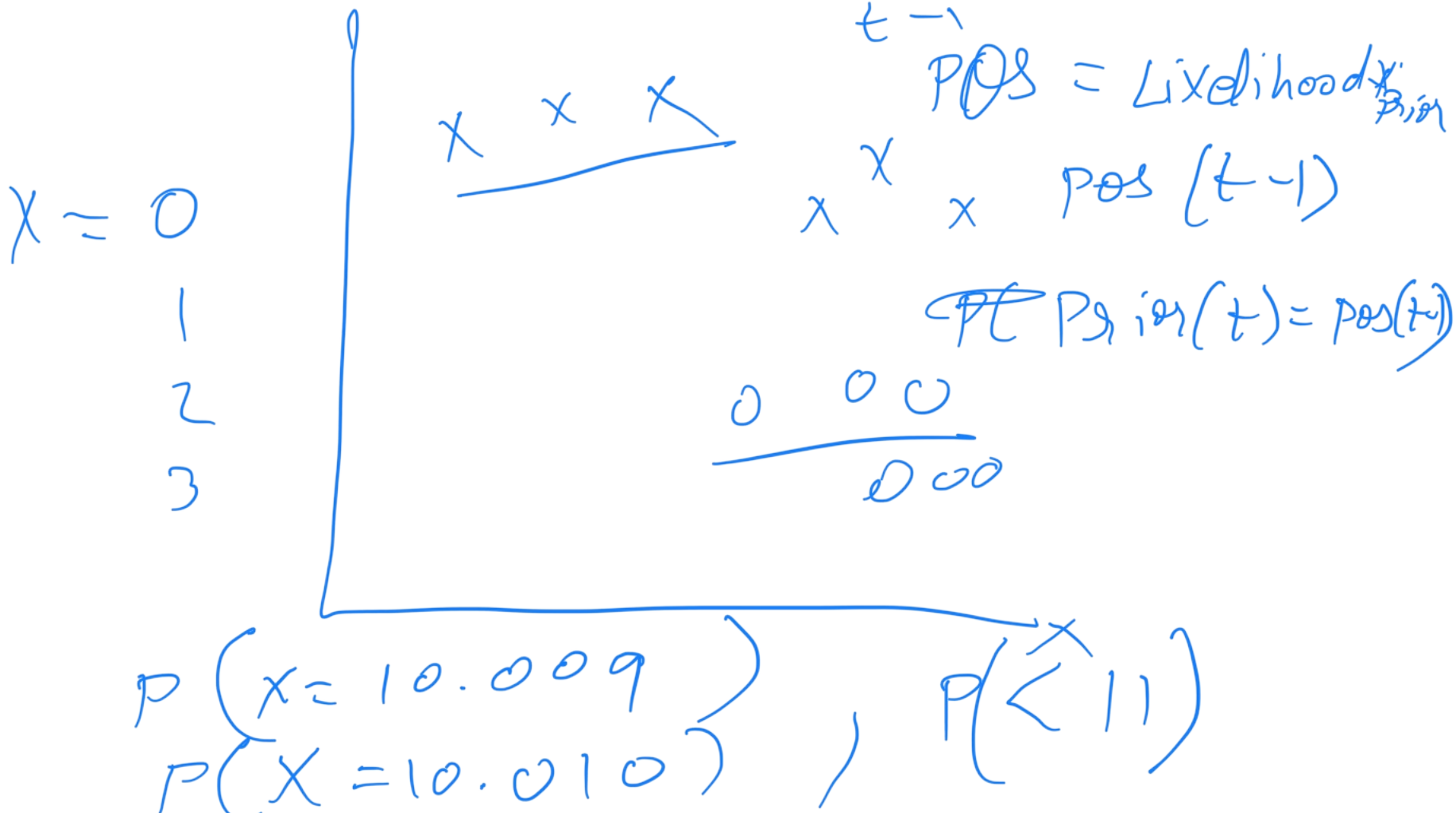
- Variance calculates how spread out a distribution is
- It puts special emphasis on the outliers
- standard deviation $\sigma = \sqrt{\mathrm{Var}[X]}$
- Other (relevant) statistical quantities: Mode (most likely value), Median (middle of the distribution)

$$E[X] = \sum_{x} x\, p(x)$$

$$\frac{1}{3}\left((4-5)^2 + (3.5-5)^2 + (6-5)^2\right)$$

$$= 1\cdot\left(\frac{1}{6}\right) + 2\left(\frac{1}{6}\right) + 3\left(\frac{1}{6}\right) + \cdots + 6\left(\frac{1}{6}\right)$$

$$= 4$$


t →
POS = Likelihood
pos (t-1)
Position(t) = pos(t-1)
0 0 0
0 0 0
X = 0
1
2
3
X
P(X = 10.009)
P(X = 10.010)
P(< 11)

PMF (Probability Mass Function)

$$p_X(x) = P(X = x)$$

Discrete Random Variables

PMF

Probability Density Function / Cumulative Density Function

$$f_X(x) = \frac{d}{dx} F_X(x).$$

$$F_X(x) = \mathrm{P}(X \le x)$$

CDF

Cumulative Distribution

$$\mathrm{P}(a < X \le b) = F_X(b) - F_X(a)$$

$x = \{1, 2, 3, 11, 20\}$

$P(X=1) = \frac{20}{100} = 0.2$

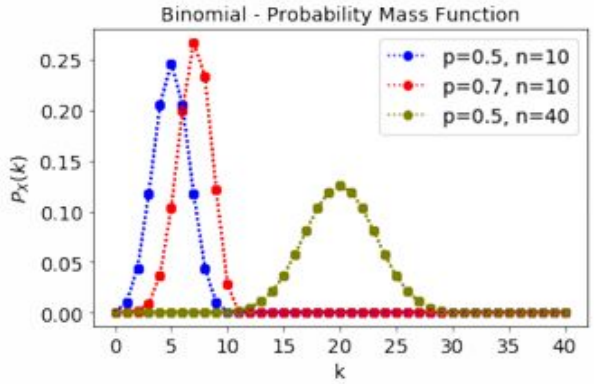
Bernoulli Distribution

Experiment where the outcome is either a *success* ($X = 1$) with probability p or a *failure* ($X = 0$) probability $1 - p$,

$$P\{X = 1\} = p, \; P\{X = 0\} = 1 - p$$

Binomial Distributione $b(n, p)$

- X: number of successes in n (fixed) independent Bernoulli trials.
- $p_X(k) = \binom{n}{k} p^k (1-p)^{n-k}, \; k = \; 0, \; 1, \; 2, ..., \; n$



Uniform Distribution

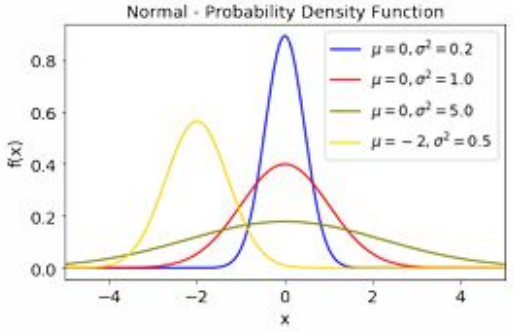
A random variable is said to be uniformly distributed over the interval (α, β) if its pdf is given by

$$f_X(x) = \begin{cases} \frac{1}{\beta - \alpha}, & \alpha < x < \beta \\ 0, & \text{otherwise} \end{cases}$$

Normal Distribution

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad -\infty < x < \infty$$

$$X \sim \mathcal{N}(\mu, \sigma^2)$$



Jointly distributed random variables

If we have to deal with two or more random variables (let us say X and Y)

- Joint **cumulative** distribution of X and Y

$$F_{XY}(x, y) = P\{X \leq x, Y \leq y\} = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{XY}(x, y) dx dy$$

- Distribution (**marginal** distribution) of X ($F_X(x)$)

$$F_X(a) = P\{X \leq a\} = P\{X \leq a, Y < \infty\} = F(a, \infty)$$

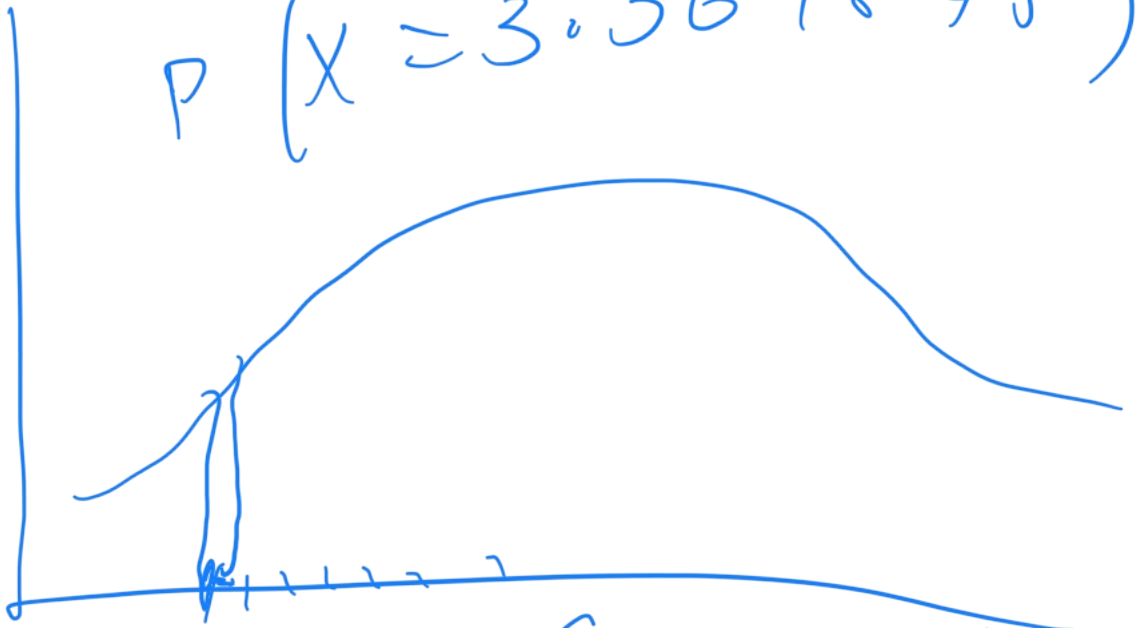
- Joint probability **mass** function (convenient to have for discrete distribution)

$$p(x, y) = P\{X = x, Y = y\}$$

- **Marginal PMFs**: Probability mass function of X ($p_X(x)$) or Y ($p_Y(y)$)

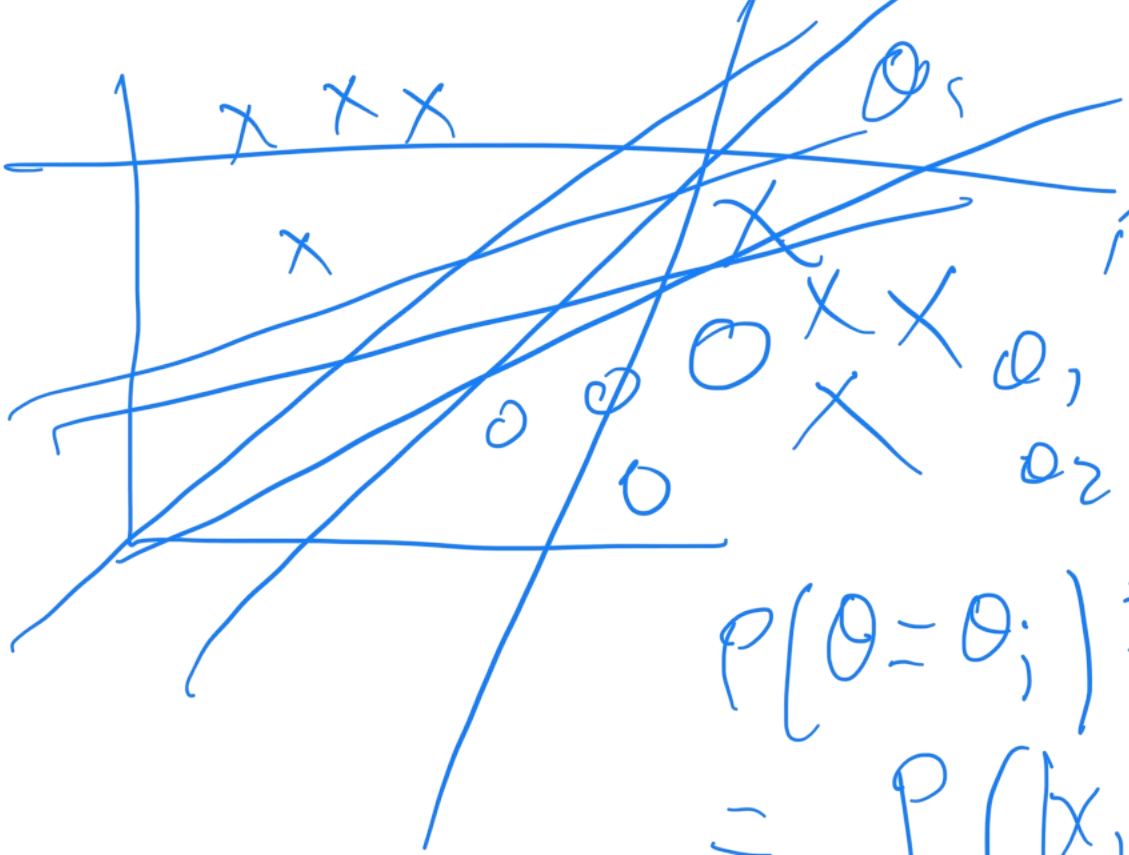
$$p_X(x) = \sum_{y: p(x,y) > 0} p(x, y)$$

$P(X = 3.369875) = 0$



$(x, x+\delta x)$

$P(X = \emptyset)$



θ_1

@ each line is 1 model

θ_1 is line 1

θ_2 - - line

$$P(\theta = \theta_i \mid x, y)$$

$$= P(x, y \mid \theta_i) \cdot P(\theta)$$

K L Divergence

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \log\left(\frac{P(x)}{Q(x)}\right).$$