Lecture 2 ML & Opt

- Recap of some probability
- Parametrized machine learning models
- Maximum likelihood estimation

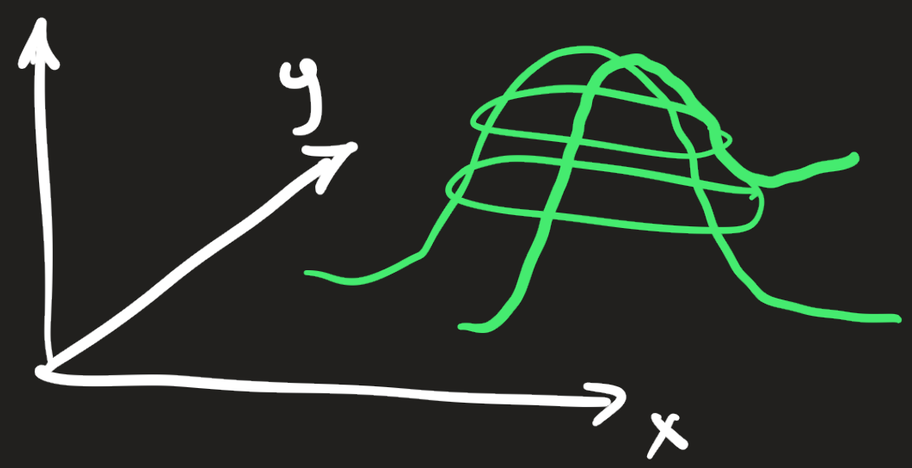
Last lecture we saw that we model our ML problem as learning

$$f : x \longmapsto y$$

y → targets

x ↑ predictor vars/ features

by modeling $(x, y) \sim \mathbb{P}$ and minimizing the average loss of f w.r.t. $\mathbb{P}$



Distributions

$X \sim \mathbb{P}$ means $\mathbb{P}(X \in A)$ is the probability that X takes on values in the set A

Two cases

X is discrete:

$$\mathbb{P}(X \in A) = \sum_{x \in A} p_X(x)$$ where p_X is the probability mass function. (pmf)

We have $\sum_x p_X(x) = 1$ and $p_X \geq 0$ everywhere

X is continuous:

$$\mathbb{P}(X \in A) = \int_A p_X(x)\,dx$$ where p_X is the probability density function (pdf)

We have $\int_{\mathbb{R}} p_X(x)\,dx = 1$ and $p_X \geq 0$ everywhere

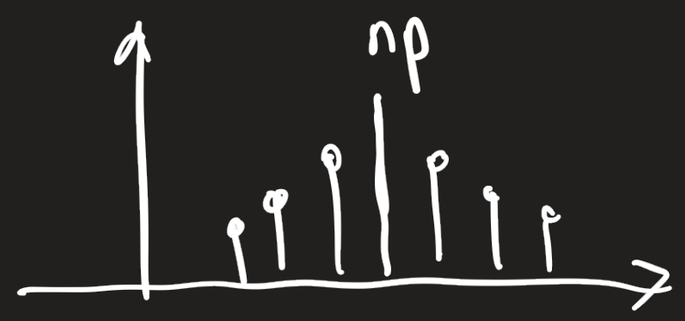
Examples

$X \sim \text{Bern}(p)$ — Bernoulli distribution w/ parameter p
$p \in [0,1]$

means $X = \begin{cases} 1 & \text{with prob } p \\ 0 & \text{w.p. } 1-p \end{cases}$

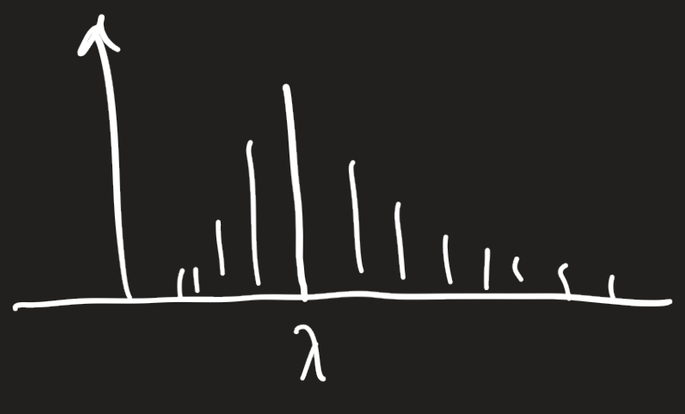
$X \sim \text{Binom}(n,p)$ — Binomial dist. w/ params n & p

means $\mathbb{P}(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$ for $k \in \{0,1,\dots,n\}$



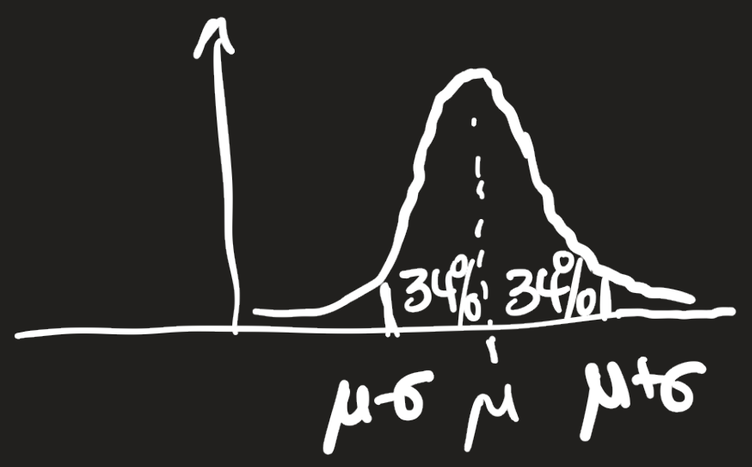
$X \sim \text{Poisson}(\lambda)$ – Poisson dist w/ rate parameter λ

$$\mathbb{P}(X=k) = e^{-\lambda}\frac{\lambda^k}{k!} \quad \text{for } k=0,1,2,\ldots$$



$X \sim \mathcal{N}(\mu, \sigma^2)$ – Normal or Gaussian r.v. w/ mean μ and variance σ^2

$$p_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$



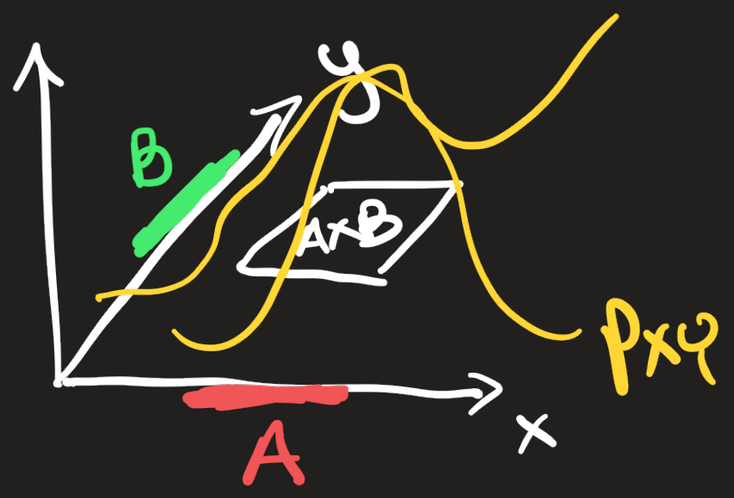
To model dependencies b/w random variables we use joint distributions

$\mathbb{P}(X \in A, Y \in B)$ written as $(X, Y) \sim \mathbb{P}$

Analogously we have joint pdfs and joint pmfs.

We assume there exists a joint pdf

$$\mathbb{P}(X \in A, Y \in B) = \int_{A \times B} p_{XY}(x, y)\, dx\, dy$$



where $p_{XY} \geq 0$ and

$$\int_{\mathbb{R} \times \mathbb{R}} p_{XY}\, dx\, dy = 1$$

Marginals

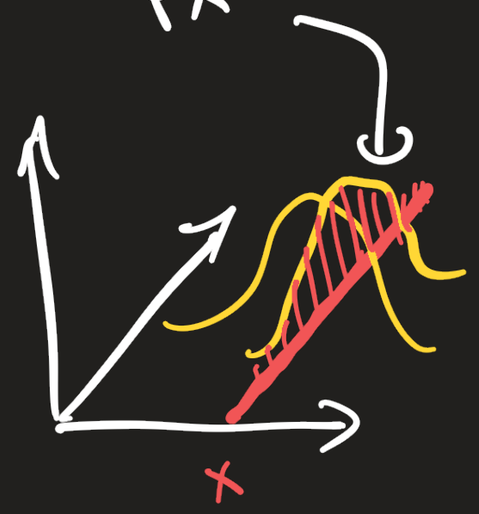
Given p_{XY} we can extract the probability dist of X

$$\mathbb{P}(X \in A) = \mathbb{P}(X \in A, Y \in \mathbb{R}) = \int_{A \times \mathbb{R}} p_{XY}(x,y)\, dx\, dy$$

$$= \int_A \int_{\mathbb{R}} p_{XY}(x,y)\, dy\, dx$$

marginal pdf of x

$$:= \int_A p_X(x)\, dx \quad \text{where} \quad \boxed{p_X(x) = \int_{\mathbb{R}} p_{XY}(x,y)\, dy}$$

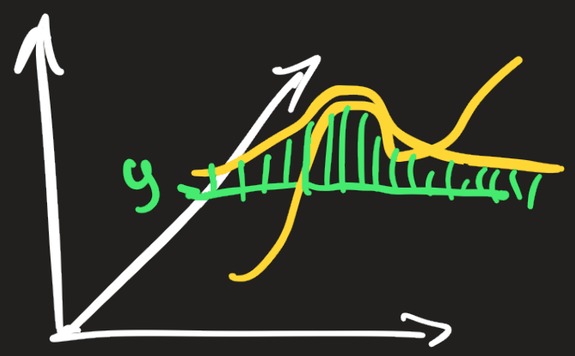


and $p_X(x) \geq 0$ and $\int_{\mathbb{R}} p_X(x)\, dx = 1$

Similarly we get p_Y by marginalizing over x

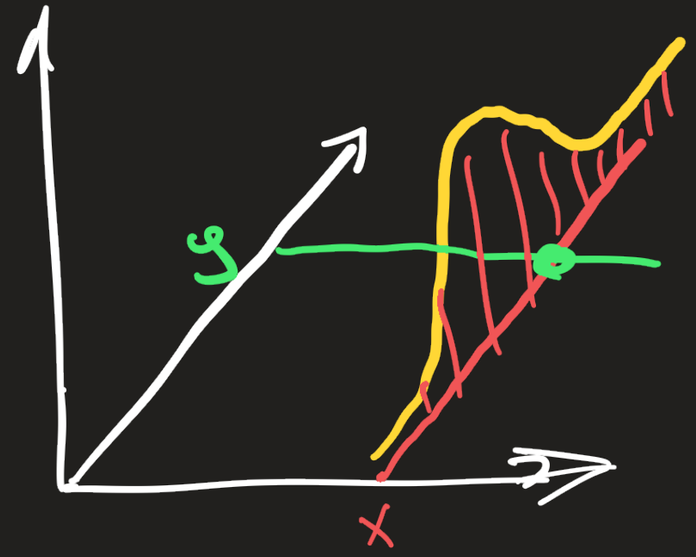
$$p_Y(y) = \int_{\mathbb{R}} p_{XY}(x, y)\, dx$$

↑ marginal pdf of Y



In the best case, knowledge of y conditional on x is captured by the <u>conditional</u> distribution of y given x

$$P_{Y|X=x}(y) = P_Y(y|X=x) = \frac{P_{XY}(x,y)}{P_X(x)}$$



Expectation

Average value of a r.v. X

$$\mathbb{E}X = \int_{\mathbb{R}} x\, p(x)\, dx$$

Ex

$X \sim \text{Bern}(p) \Rightarrow \mathbb{E}X = 0 \cdot (1-p) + 1 \cdot p = p$

$X \sim \mathcal{N}(\mu, \sigma^2) \Rightarrow \mathbb{E}X = \mu$

Expectation of random vectors

If $\begin{bmatrix} X \\ Y \end{bmatrix} \sim \mathbb{P}$ then $\mathbb{E}\begin{bmatrix} X \\ Y \end{bmatrix} = \int_{\mathbb{R}\times\mathbb{R}} \begin{bmatrix} x \\ y \end{bmatrix} p_{X,Y}(x,y)\, dx\,dy$

$$= \int \begin{bmatrix} x\, p_{X,Y}(x,y) \\ y\, p_{X,Y}(x,y) \end{bmatrix} dx\,dy$$

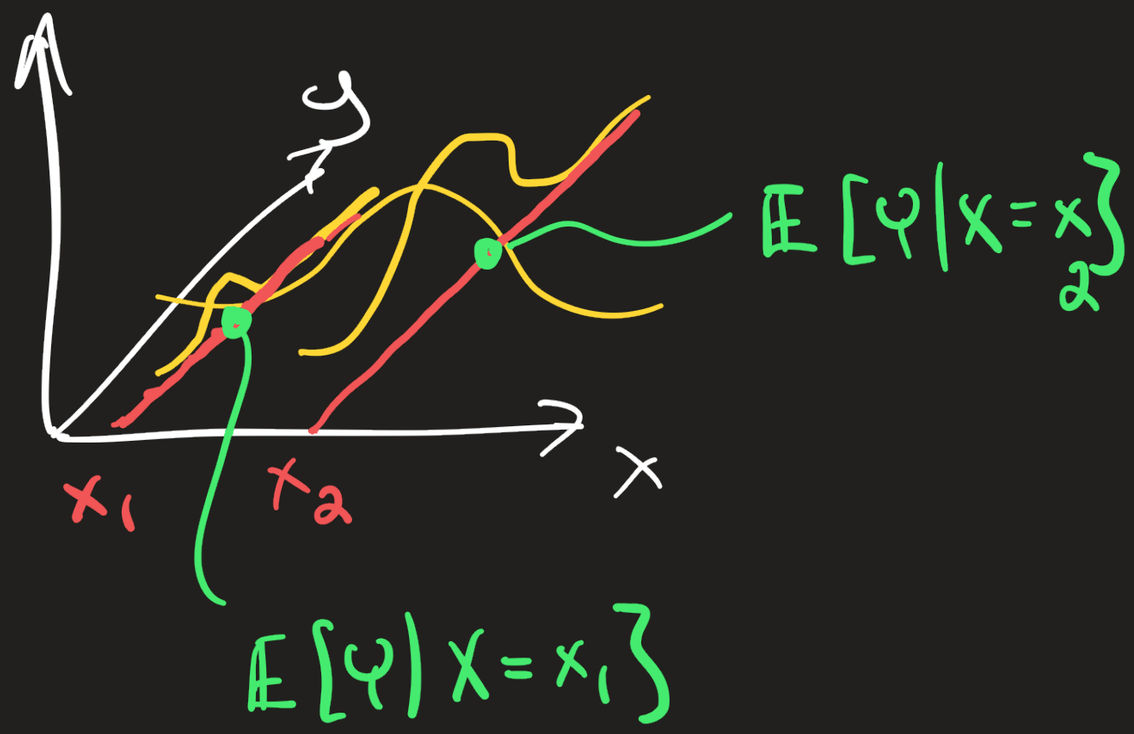
$$= \begin{bmatrix} \int x\, p_{X,Y}(x,y)\, dx\,dy \\ \int y\, p_{X,Y}(x,y)\, dx\,dy \end{bmatrix} \overset{\text{verify}}{=} \begin{bmatrix} \int x\, p_X(x)\, dx \\ \int y\, p_Y(y)\, dy \end{bmatrix}$$

$$= \begin{bmatrix} \mathbb{E}X \\ \mathbb{E}Y \end{bmatrix}$$

Conditional Expectation

$\mathbb{E}[Y|X]$ — average value of Y given knowledge of X

$$\mathbb{E}[Y|X=x] = \int_y y\, p_Y(y|X=x)\, dy$$



Variance

Variance measures how close a random variable stays to its mean, on average

$$\mathrm{Var}(X) = \mathbb{E}(X - \mathbb{E}X)^2$$

$$= \int_{\mathbb{R}} (x - \mathbb{E}X)^2 \rho_X(x)\, dx$$

$$= \int_{\mathbb{R}} \left[x^2 - 2x\,\mathbb{E}X + (\mathbb{E}X)^2 \right] \rho_X(x)\, dx$$

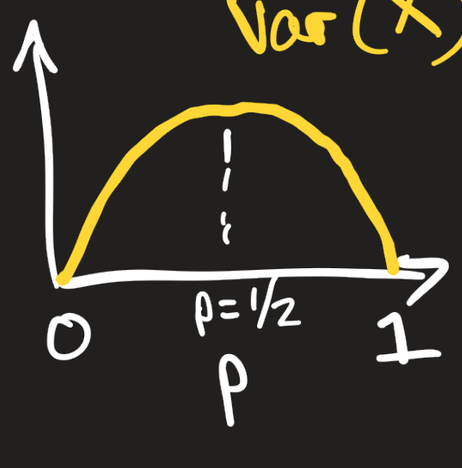
$$= \int_{\mathbb{R}} x^2 \rho_X(x)\, dx - 2\mathbb{E}X \int_{\mathbb{R}} x\, \rho_X(x) dx + (\mathbb{E}X)^2$$

$$= \mathbb{E}(X^2) - 2(\mathbb{E}X)^2 + (\mathbb{E}X)^2 = \boxed{\mathbb{E}X^2 - (\mathbb{E}X)^2}$$

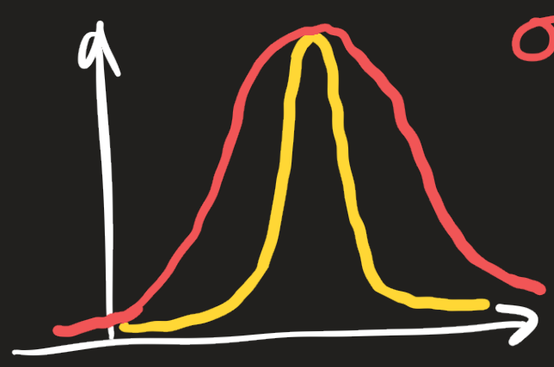
Ex of Variance

$X \sim Bern(p) \Rightarrow$

$$Var(X) = \mathbb{E}[X^2] - (\mathbb{E}X)^2$$
$$= \mathbb{E}X - (\mathbb{E}X)^2$$
$$= p - p^2$$



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



$\sigma^2 > \sigma^2$

Variance is a measure of uncertainty

Independence $(X \perp\!\!\!\perp Y)$

We say two random variables are independent if

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\,\mathbb{P}(Y \in B)$$

$$\Updownarrow$$

$$P_{XY}(x,y) = P_X(x)\,P_Y(y)$$

$$\Updownarrow$$

$$P_Y(y \mid X=x) = P_Y(y) \quad \& \quad P_X(x \mid Y=y) = P_X(x)$$

Conditional Independence

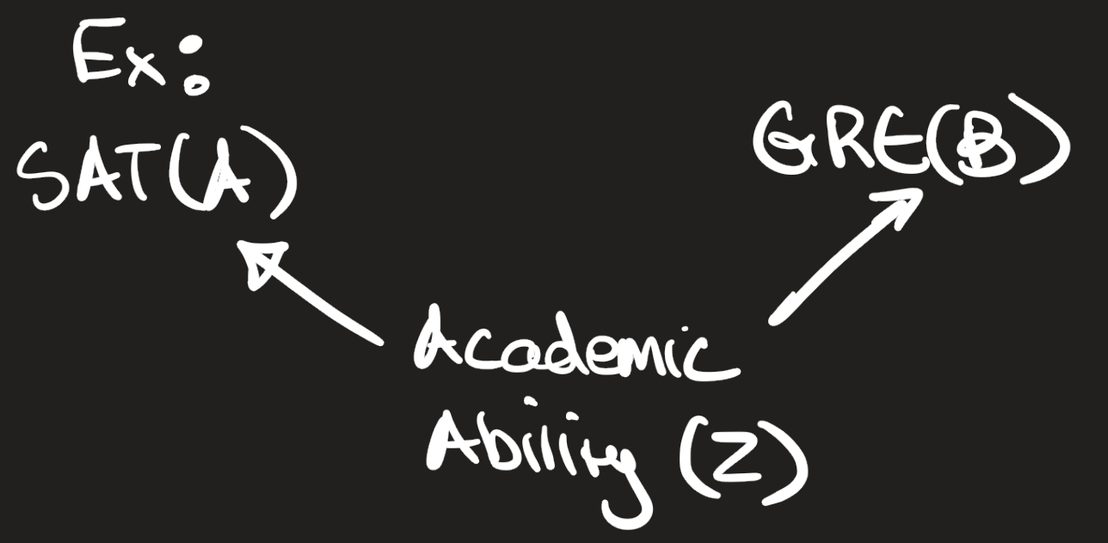
We say X & Y are conditionally independent given Z if $(X \perp\!\!\!\perp Y \mid Z)$

$$P_{X,Y|Z} = P_{X|Z} \cdot P_{Y|Z}$$

$$P_{X,Y|Z=z}(x,y) = \frac{P_{XYZ}(x,y,z)}{P_Z(z)} \quad \text{in general}$$

$$= P_{X|Z=z}(x)\, P_{Y|Z=z}(y) \quad \text{if } X \perp\!\!\!\perp Y \mid Z$$

Ex:



$A \perp\!\!\!\perp B \mid Z$

Parameterized ML models

$P_Y(y|x)$ – captures everything about values of Y given knowledge $X=x$

$\mathbb{E}[Y|X=x]$ – captures a point estimate of Y given x

Issues w/ using these in ML: either can be an arbitrarily complicated unknown function of x

To deal with this, we use <u>parameterized models</u>, i.e.

$$p(y|x) = f_\theta(x)$$ where $\theta \in \mathbb{R}^p$ is a parameter vector

$$= f(x;\theta)$$

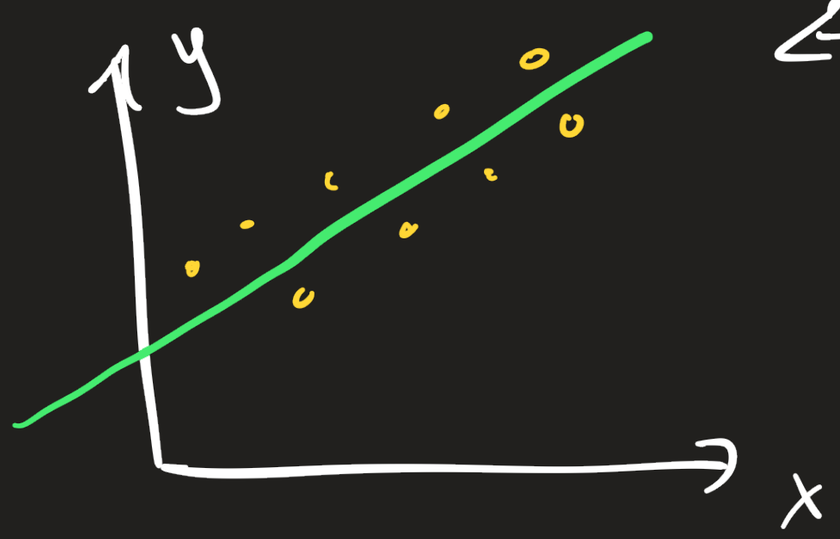
Ex

OLS: $y = \beta^T x + \varepsilon$ where $\varepsilon \sim \mathcal{N}(0, \frac{1}{n})$

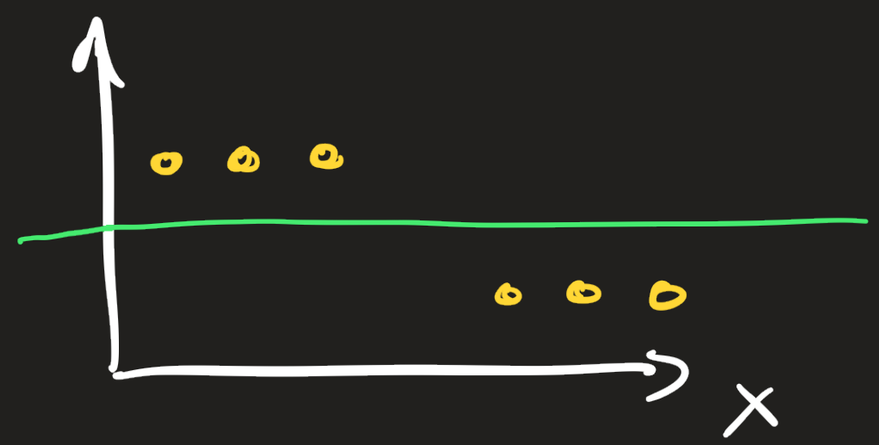
i.e. $y|x \sim \mathcal{N}(\beta^T x, \frac{1}{n})$

Assume data (x_i, y_i) came from this type of model for $y|x$, then we learn β that fits this data best

OLS is appropriate for data that looks like



and inappropriate for, e.g.



what about parameterized models for data that doesn't fit the OLS assumptions?

e.g.

- y is not a cont. r.v.
- y has a limited range
- $y - \mathbb{E}[y|x]$ is not indep of x

Ex

- Poisson / "shot noise" model

corresponds to Poisson(λ) being used, where λ is a function of x

$$y|x \sim \text{Poisson}(\lambda_\theta(x))$$

e.g. this gives

$$\mathbb{E}[y|x] = \lambda_\theta(x)$$

Notice since $\lambda_\theta(x)$ is a rate, it must satisfy $\lambda_\theta(x) > 0$. We impose this by taking

$$\lambda_\theta(x) = e^{f(x;\theta)}$$