Recap lecture

No assumptions.
non-parametric

hardest
Regression

find F — simplest

Pattern recognition — next simplest

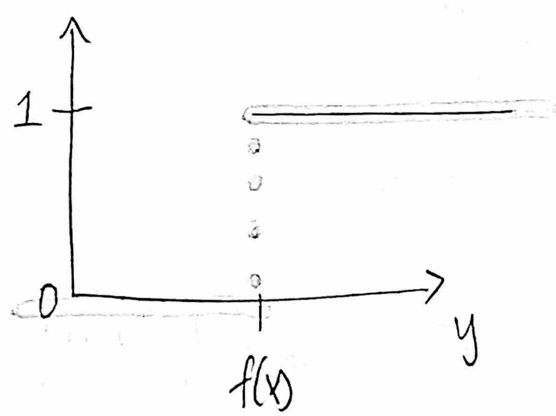
Find f:

$\boxed{Y = f(X)}$ exact relationship.

$$\mathcal{F} := \{ F_{Y|X}(y|x) = \mathbb{1}_{y \geq f(x)}(y)$$

distribution func

for a given f(x)

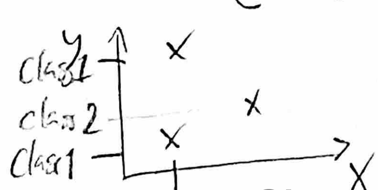


Pattern Recognition

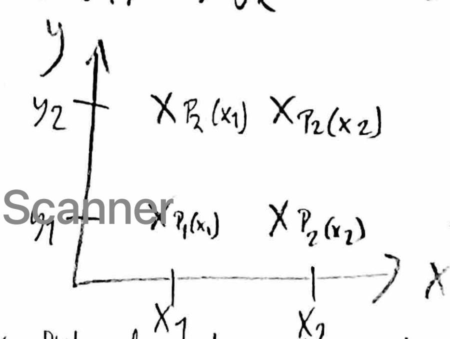
it has a # data points belonging to classes.

all X belong to the same classes

$$\mathcal{F} := \{ F(y|x) \text{ is discrete, } y_1, \dots, y_k \text{ are the classes}$$



this would not work. The X's need to belong to the same classes



this would work. Need data that looks like this.

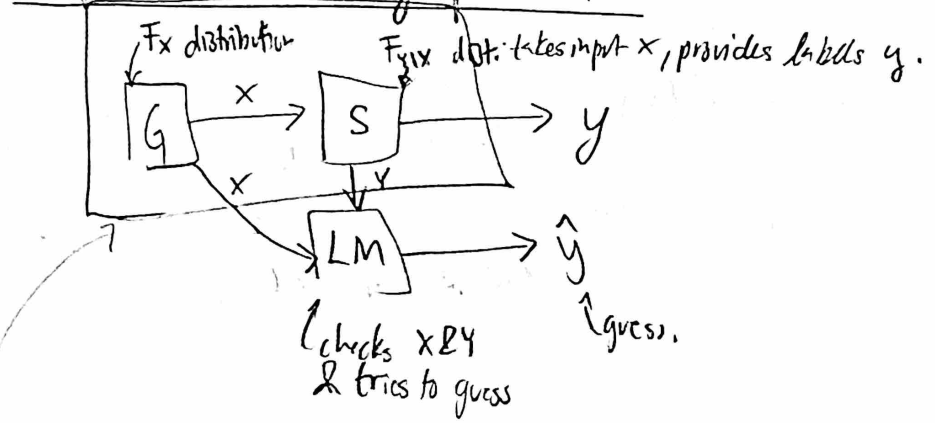
The prob. depend on the x's.

$$\mathbb{P}(Y = y_i \mid X = x) = P_i(x)$$

[These are non-parametric because the function $F(x)$ could be any function. And we make no assumptions on how the Xs goes.
So two ways that determine if it is non-parametric.]

Regression: see lecture notes.
Assume regression & y-variable has finite 2nd moment.

Supervised learning problem

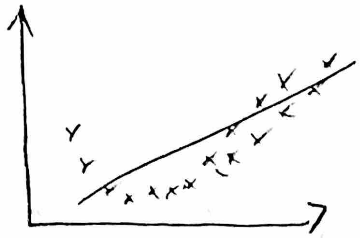


$\underbrace{\text{Statistical model}}_{F_{XY} \in \mathcal{F}}$ $\qquad$ $\underbrace{\text{Model space}}_{LM}$ $\mathcal{M}$ (Where LM searches)

Risk (loss) (how LM searches)

Statistical model $\neq$ Model space.
We are trying to find a model space with our approximation. We want to find a good approx to reality. But that doesn't mean we have found the correct model space we have just found an approximative model space.

The approx linear model space F
the actual model space.



Tentaplugg
- individuella
- Group assignments
- lab itds
- problem solving sessions.

Risk

Risk = expected loss. (L)

$\underbrace{L(g, x, y)}_{\text{"an error"}} \quad g \in \mathcal{M}$

$R(g) = E[L(g, \underbrace{\overbrace{x, y}^{\text{Random}}}_{\text{expectation to these random inputs/variables}})]$.

LM wants to find the g^* which is the function that minimizes the expected loss.

$g^* = \text{"arg min"}\ R(g) \quad g \in \mathcal{M}$.

$R(g^*) = \min(R(g))$

LM cannot find $R(g^*)$. It cannot do that as it doesn't have access to the true distribution. It only sees the $\underbrace{\text{data}}_{\text{observations}}$ so instead it minimizes := LM sees $(X_1, Y_1), \dots, (X_n, Y_n)$

$$\hat{R}_n(g) = \frac{1}{n} \sum_{i=1}^{n} L(g, x_i, y_i)$$

$\arg\min_{\mathcal{M}} \hat{R}_n(g) = \hat{g}_n^*$ hopefully $\hat{g}_n^* \sim g^*$ or at least

$\underbrace{R(\hat{g}_n^*)}_{\text{the true}}$ is small.

$n \gg 4$ = train-test

Estimate $R(\hat{g}_n^*)$ using testset.

Estimation

Examples:
empirical variance, mean etc.

Two concepts
* statistic : a function of the data.
* Estimator : = point estimator. = "a statistic that "estimates" some underlying value.

$$\mathcal{F} := \{ F(x;\lambda) \quad \lambda \in \Lambda \}$$

big lambda

$$\underbrace{\theta(\mathcal{F})}_{\text{parameter map}} \longrightarrow \mathbb{R}$$

We use the parameter map to map the non-parametric world.

Def. Parametric = finite number of parameters.

The estimator T

$$T(x_1, \dots, x_n) \sim \theta(F^*) = \theta^*$$

Ex 1:

$$\theta(F) = \overbrace{\int x \, dF(x)}^{\text{the expectation}}$$

Ex 2: $\theta(F) = \inf\limits_{g \in \mathcal{M}} R(g)$

Usually we use these two as estimators.

in Ex 1: what is a good estimator T?

$$T = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Ex 2: — " — $T = \min\limits_{g \in \mathcal{M}} \frac{1}{n} \sum_{i=1}^{n} L(g, X, Y)$

Ex 3: $\theta(F) = \operatorname{argmin}\limits_{g \in \mathcal{M}} R(g) = g^*$

$$(x_1, \dots, x_n) = \operatorname{argmin}_{g \in \mathcal{M}} \hat{R}_n(g) = \hat{g}_n^*$$

Propeties of estimation:

* bias (unbiased)
* standard error
* Mean squared error (MSE)
* asymptotically unbiased
* asymptotically consistent

Bias

$E[T] - \theta^*$: bias is diff expected value of the $\underbrace{\text{estimator}}_{T}$ − true value

negative bias:

Standard error (se): $\sqrt{V[T]}$

Measures the variability of your estimator T.

MSE

$E[(T-\theta^*)^2] = \text{bias}^2 + \text{se}^2$

{ just minimize se, bias might go up.
{ just minimize bias, se might go up.

if have Big data: use MSE.

Asymptotically unbiased

as # samples → inf, bias → 0.

Asymptotically consistent

as # samples increases, the estimator T converges to the true parameter.

$T \xrightarrow[n \to \infty]{P} \theta^*$

Use if MSE → 0.

Empirical dist. function

samples are $X_1, \dots, X_n$

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{X_i \le x}$$

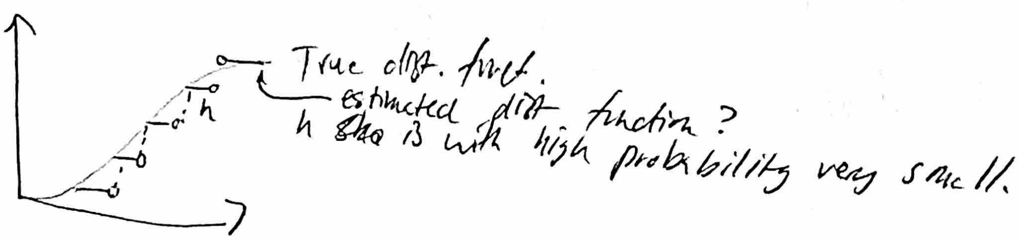
$$\underbrace{E[\hat{F}_n(x)]}_{T} = \frac{1}{n}\sum_{i=1}^{n} \underbrace{E[\mathbb{1}_{X_i \le x}]}_{\underbrace{\mathbb{P}(X_i \le x)}_{\text{dist. func } F(x)}} = F(x) \leftarrow \theta^*$$

This is our θ^*

DKW-inequality

= The thing that corresponds to hoeffding but for empirical dist. functions.

$$= \mathbb{P}(|\hat{F}_n(x) - F(x)| > \varepsilon) \le 2ne^{-2n\varepsilon^2}$$



This actually gives confidence intervals for the quantiles.