①

Lecture-6a

Predicting Missing Links

Feb 20, 2024.

Reminder: Project proposal hard.
PSA

actual $\Rightarrow$ $G = (V, E)$

observed $\Rightarrow$ $G^* = (V, E^*)$, where $E^* \subseteq E$ $\left(E^* = f(E) \right)$

unconnected pairs $Y = V \times V - E^* \sim \Theta(n^2)$ haystack

missing links $X = E - E^* \sim O(n)$ needles.

missingness function.

baseline accuracy by guessing $\Theta(\frac{1}{n})$

[so hard!]

Reasons we care:

- most 图 are incomplete! / ex: -observing ecosystem -gene-gene interaction - NSA / CIA
- use correlations to marshall resources.
- compare methods or models. (cross-validation)

Facebook as eg.
1. recency
2. friends of friends.

Idea of a predictor is to optimize a score function.

$$\text{score}(i, j) : Y \to \mathbb{R}$$

Key message:

We can define a "spurious link prob", but it will (by changing the score function) be easier than link prediction. (why?)

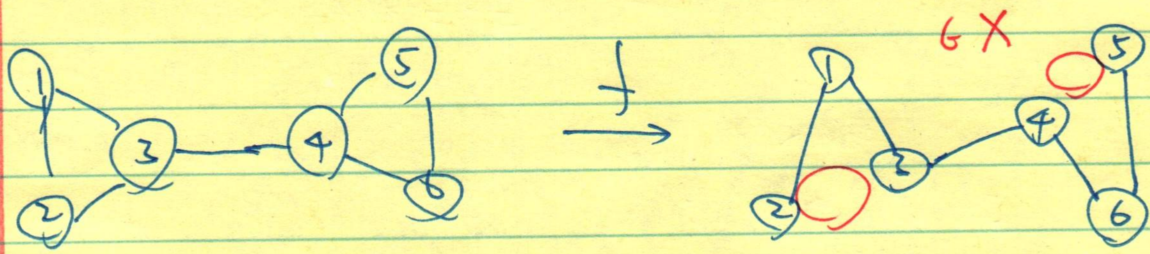
$\Theta(n^2)$ to explore v.s. $\Theta(n)$ to explore.

A baseline algm.

if $i, j \in Y$, $score(i, j) = Uniform(0, 1)$

↓

independent of G^* !



score table

$\in Y$

i	j	score(i, j)
1	5	r = Uniform(0, 1)
1	4	r
1	6	r
⋮		r
2	3	r
⋮		r
4	5	⋮

Topological predictors and local smoothness.

the idea

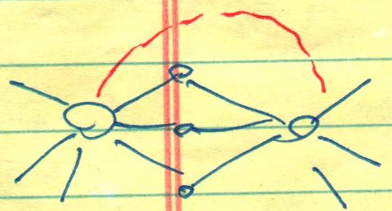
predict a missing link to occur where it would cluster with other (observed) edges

define this as a predictor

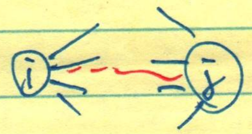
Two ways to operationalize this idea :

1) Jaccard coefficient.

$$\mathrm{Jaccard}(i,j) = \frac{|nb(i) \cap nb(j)|}{|nb(i) \cup nb(j)|}$$



2) degree product $= k_i k_j$



$$\mathrm{score}(i,j) = k_i k_j + \mathrm{Uniform}(0,\epsilon) \quad \epsilon \ll 1$$

$$\mathrm{scores}(i,j) = \mathrm{Jaccard}(i,j) + \mathrm{Uniform}(0,\epsilon)$$

break ties randomly

$$\epsilon \ll \frac{1}{n-2}$$

④

an example Jaccard :

DP :

another example :

⑤

ask ?!

Measuring performance: the AUC

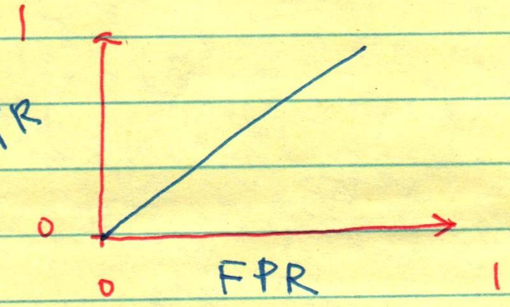
predicting missing links is a binary classifier. (req. only TP / FP).

we can use Area Under the Curve (AUC)

→ Receiver operating characteristics (ROC) curve.

nice properties

1) scale invariant (only cares about the relative ordering)
2) threshold invariant (general measure of distinguishability)



Mathematically

$$AUC = \Pr\left(\text{score}(TP) > \text{score}(TN) \right)$$

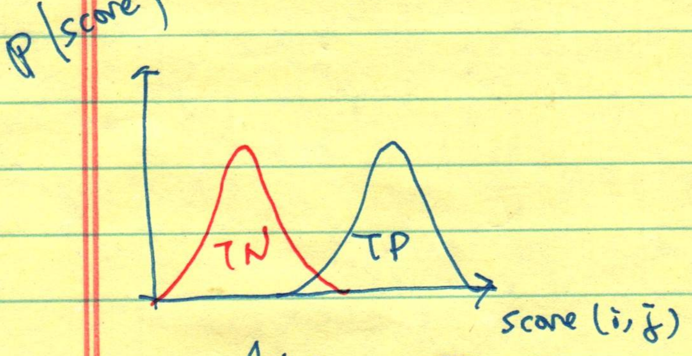
$AUC = 1 \Rightarrow$ perfect distinguishability

$= \frac{1}{2} \Rightarrow$ if I cannot tell a TP from TN.

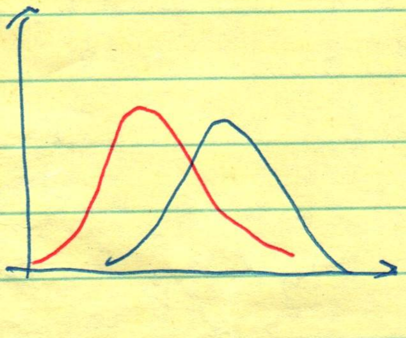
In link prediction

TN: $\underset{|Y-X|}{\Theta(n^2)}$

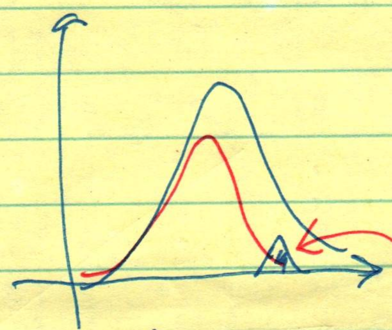
TP: $\underset{|X|}{O(n)}$



AUC = 0.99



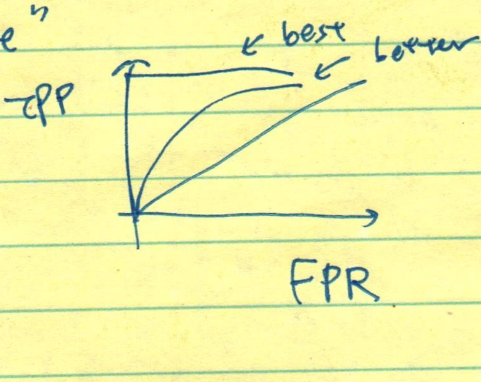
AUC = 0.76



AUC = 0.55

AUC ⇒ useful for comparing predictors (agnostically) actually ...

if we want to see how much better, we need to look at "the curve"



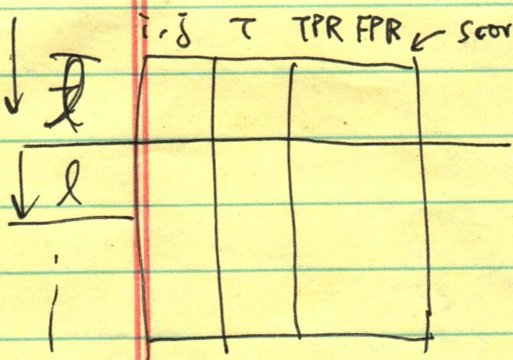
1

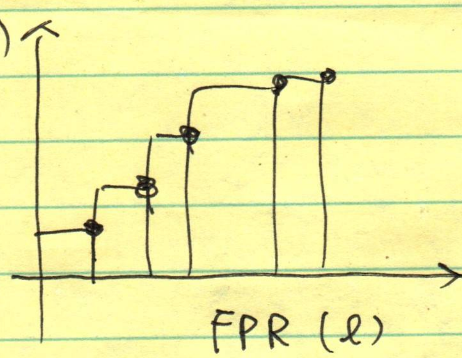
Wk-6b. Feb 22, 2024.

$$TPR(\ell) = \frac{1}{\underset{\#TP}{T}} \sum_{k=1}^{\ell} \tau_k \; ; \quad \tau_k = \begin{cases} 1 & \text{if it's missing a edge} \\ 0 & \text{otherwise} \end{cases}$$

$$FPR(\ell) = \frac{1}{\underset{\#FP}{F}} \sum_{k=1}^{\ell} (1 - \tau_k) \; ; \qquad FPR(0) \equiv 0.$$

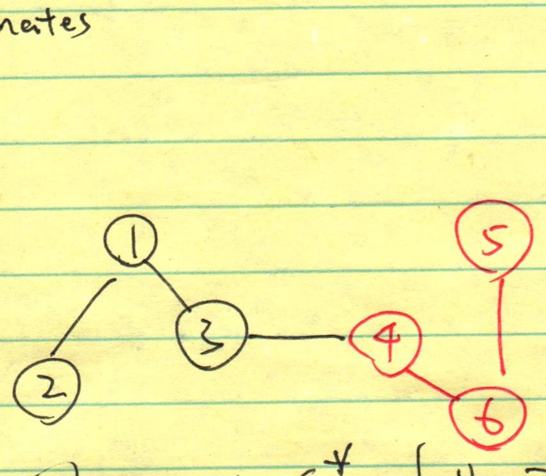
$$AUC = \sum_{\ell=1}^{|T|} \underbrace{TPR(\ell)}_{y} \times \underbrace{\left(FPR(\ell) - FPR(\ell-1)\right)}_{\Delta x}$$





ROC coordinates

Example (See Clauset's notes)



Observed $G^* = (V, E^*)$

Wk 6-6. Feb 22, 2024.

The world of missing link predictors.

Three main classes.

1) topological: score(i,j) use only G^* (local measure of structure around i,j)

Jaccard, deg prod, $f(v_i, v_j)$
$\uparrow$
node-level statistic

(implicitly assumes an edge generating mechanism)

2) model-based

$$\Pr(G|\theta) = \prod_{i<j} \Pr(i \to j \,|\, \theta)$$ edges are

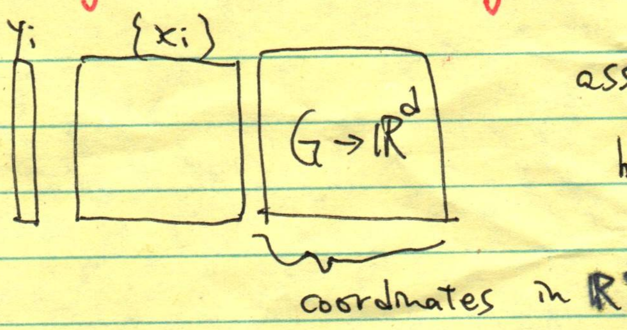
elaborate random graph models $\rightarrow$ conditionally indep

(assume some explicit edge generating mechanism)

Stochastic block model (SBM)

3) embeddings

$$d(i,j)^{-1} \propto \text{score}(i,j)$$



assumes a spatial homophily mechanism for edge generating.

③

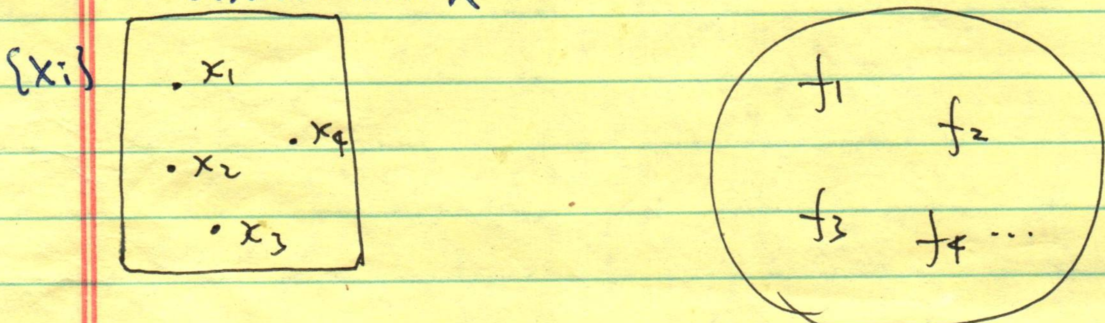
But even more exciting are $\boxed{\text{meta-learning}}$ solutions
(aka. ensemble learning)

→ Key assumption: edges are the $\boxed{\text{observations}}$ and the $\boxed{\text{predictors}}$

brief review

model selection via cross validation

observed data X



candidate models. F.
$\{f_n\}$
random forest, neural network, regression, etc.

* Which f_k provides the best out-of-sample prediction accuracy on not-yet-observed x_i ?

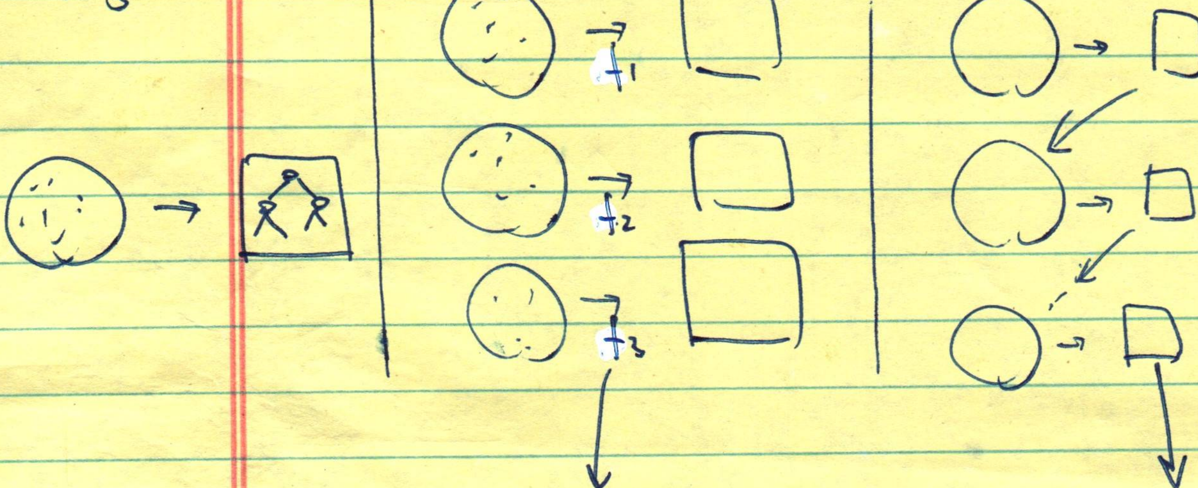
Solution:
① partition X in test and train data, randomly $X_{train} \cup X_{test}$
② learn apply each f_k to $X_{train} \mapsto \theta_k$
③ measure accuracy of each θ_k on $X_{test} \longrightarrow \varphi_k$
④ select $\underset{k}{\operatorname{argmax}}\ \varphi_k \to$ best model
"winner take all approach"

④

"learning w/ ensembles"

bagging (parallel) | boosting (sequential)

single model



bootstrap aggregating (1994)
(special case of model averaging)

simulate noise in data

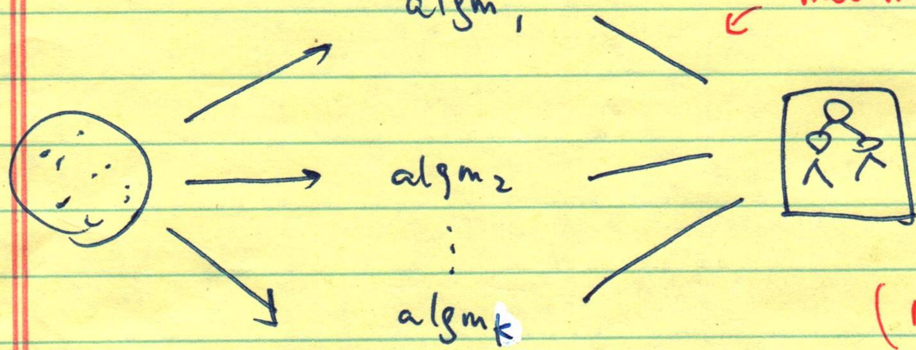
(1990)

seq. produce "weak predictors"
[best single feature model]
upweight hardest remaining X_i
linearly combine weak predictors
AdaBoost
(main algm people use)

⑤

Stacking (1992)

metamodel



(non-Bayesian)

train many strong or weak predictors

learn a model of which predictor is best for a given type of input.

Delphi method
→ constructing a predictive distribution

not model selection or model averaging, but model combination.

35 min — start the slides.

predicted label \ true label	1	2
1	TN	FN
2	FP	TP

2 → positive.

$$\text{recall} = \frac{TP}{TP + FN}$$

$$\text{precision} = \frac{TP}{TP + FP}$$

>> 203 features v. 25-30 features.

⑦

on real data :

1) inherently harder.

2) overfit — driven by algm.

80 % social

remaining (b/w / info)

lease (others)

model class / (v.s.) all data (No-free lunch)