link prediction

link inference

SYSM 6302

CLASS 19

Edge Prediction

→ Almost all real networks are samples of a real "system"

does Facebook have all of your friendships?

$G = (V, E)$ — True Network

sampling →

$\tilde{G} = (\tilde{V}, \tilde{E})$ — Sampled Network

→ How do we infer the edges that were missed?

↳ most methods use a generalized notion of similarity

Classifier

Suppose $z_{u,v} = \begin{bmatrix} 1 \\ z_1 \\ z_2 \\ \vdots \\ z_m \end{bmatrix}$ ← # common neighbors; ← # of common letters in their name

} Can include network structure and non-network node attributes

is the similarity vector between nodes u & v

⇒ Use a threshold to decide when to add an edge

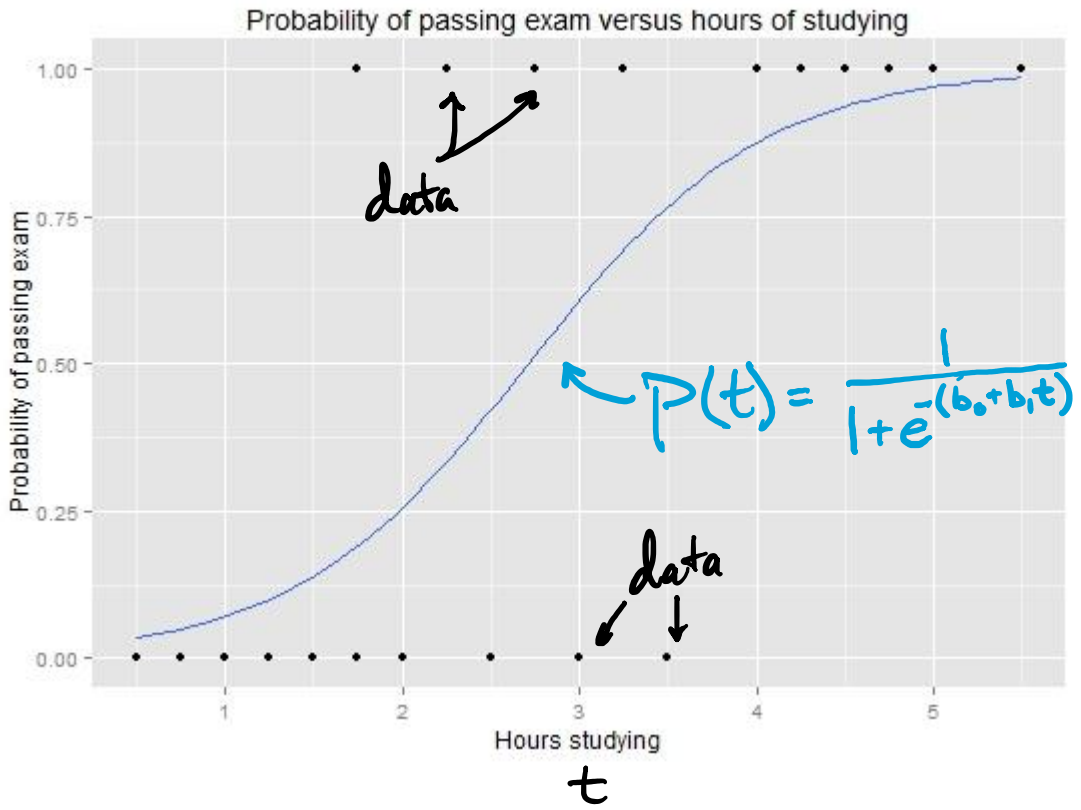
Logistic Regression

$$P((u,v) \in E \mid z_{u,v} = z) = \frac{1}{1 + e^{-\beta z}}$$

↑ probability that nodes u & v have an edge

β: parameter vector learned from sampled data

regression



Link Inference

→ If there is no "real" network that we are trying to recover

→ Movie/Artist similarity

→ protein gene expression

→ Use similarity measure $\sigma_{ij} \neq 0$ to infer an edge

recall various definitions

needs to be statistically different from zero

To do this properly

1. Test statistic
2. Null model
3. Deal with the multiple testing problem

https://xkcd.com/882/