

Recap: Unsupervised + Reinforcement Learning

Other topics: Bayesism methods

Decision trees

Boosting + bagging

graphical models

Online learning

Semi-supervised learning

* Fairness in ML

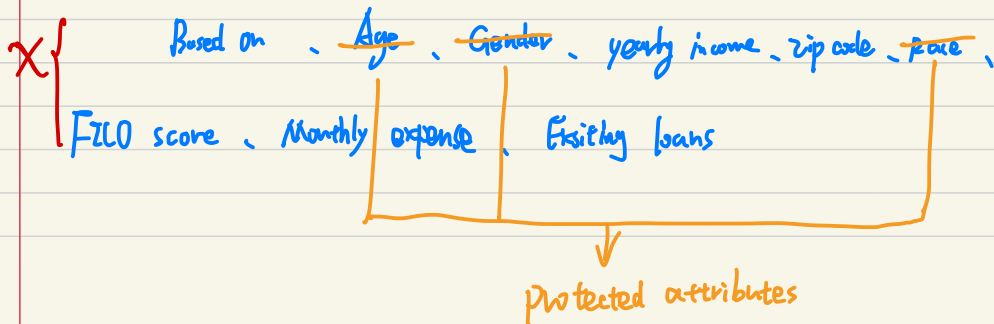
* Security in ML

* Ethics in ML

伦理

Fairness

Example: Decide who to give loans } y



→ classifier can be biased

Attempt: Exclude protected features

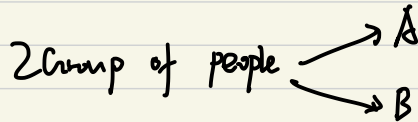
"Fairness through exclusion"

or "Fairness through unawareness"

Issue: Features can be correlated

∴ available features can be heavily linked to protected attributes.

Statistical parity



$$\Pr(\text{Loan} = \text{true} \mid \text{person} \in A)$$

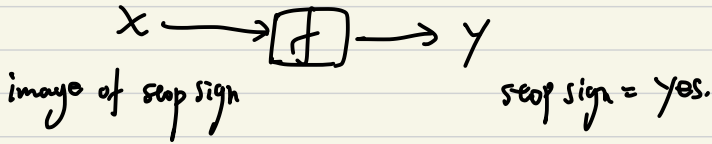
$$= \Pr(\text{Loan} = \text{true} \mid \text{person} \in B)$$

$$\Rightarrow \text{Loan} \perp \text{Group}$$

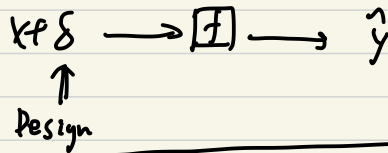
↑
"statistically independent"

other fairness objectives → Equal odds, Counterfactual parity, ...

Security & Robustness of ML



Adversarial Settings 对抗式的



Standard setting: $\frac{1}{2} \sum_{i=1}^n \text{crossentropy}(y_i, \langle w, x_i \rangle) = L(w)$
 $\min L(w)$

$$\max_{\delta} \text{crossentropy}(y, \langle w, x + \delta \rangle)$$

$$\|\delta\|_{\infty} \leq \epsilon$$

"adversarial learning"

like adding an
irrelevant image to data-
set?

