☆Recap : Unsupervised + reinforcement learning

☆other topic : 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

X { Based on , ~~Age~~ , ~~Gender~~ , yearly income , zip code , ~~Race~~ ,

FICO score , Monthly expense , Existing loans

→ Protected attributes

→ 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

$$\begin{tikzcd} & A \\ \text{2 Group of people} \arrow[ur] \arrow[dr] & \\ & B \end{tikzcd}$$

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

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

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

" statistically independent "

other fairness objectives → Equal odds, counter factual parity, ...

Security & Robustness of ML

$$x \longrightarrow \boxed{f} \longrightarrow y$$

image of stop sign　　　　stop sign = yes.

Adversarial settings

对抗式的

$$x+\delta \longrightarrow \boxed{f} \longrightarrow \hat{y}$$

$\uparrow$

Design

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

$$\min L(w)$$

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

$$\|\delta\|_\infty \leq \varepsilon$$

" adversarial learning "

like adding an irrelavant image to data-set ?