Explainable ML

Yu Xueqing

2022 年 8 月 1 日

1 why we need explainable ML?

在特定情境下模型需要有可解释性，决策的理由，不能够是黑箱

2 goal of explainable ML

- local explanation-"why do you think this image is a cat?", explain the decision
  - object x, components$\{x_1, x_2, ..., x_N\}$, 哪个 component 重要
  - $\{x_1, x_2, ...x_N\}$->$\{x_1, .., x_n + \Delta x, ...x_N\}$
    $e- > e + \Delta e$
    $\frac{\Delta e}{\Delta x}$-saliency map, 各 component 对决策的重要性
  - how a network processes the input data?
    visualization
    probing
- global explanation-"what does a 'cat' look like?"
  - CNN: 想要知道每个 filter 关注的 pattern 是什么，$X^* = argmax(x) \sum \sum a_{ij}$, a_{ij} 是 filter 的输出。X 表现 filter 侦测的特征
  - 要让 X 接近真实的输入–对 X 加 constraint
    constraint from generator