Learning to predict vs. learning to decide

Floyd Kretschmar

December 9, 2019

1 Decision making as prediction

Many decision making task are being supplemented or fully automated by data-driven predictive models. The most prolific approach to integrate machine learning into such decision making tasks is to split decision making into two seperate tasks: prediction and decision making based on the prediction. Within this framework data can then be utilized to train models that make prediction. Based on these predictions decisions are made according to a decision policy. There are many different decision problems that can be viewed under this lense, but this summary will focus on the case of binary classification with binary sensitive attributes as described in [1].

Therefore let $\mathcal{X} \subseteq \mathbb{R}^d$ be the feature domain, $\mathcal{S} = 0, 1$ the range of sensitive attributes and $\mathcal{Y} = 0, 1$ the set of ground truth labels. It is assumed that for each individual $\boldsymbol{x} \in \mathcal{X}$, $s \in \mathcal{S}$ and $y \in \mathcal{Y}$ are given by the joint probability distribution $P(\boldsymbol{x}, s, y) = P(y \mid \boldsymbol{x}, s)P(\boldsymbol{x}, s)$ as defined by [1]. Within this framework a **prediction task** is defined as training a model $Q(y \mid x, s; \theta)$ in a supervised mannor, which means solving the optimization problem

$$\begin{aligned} \underset{\theta}{\operatorname{argmin}} \quad & \mathcal{L}(\theta) \\ s.t. \quad & \mathcal{F}(y, x, s) \end{aligned}$$

where $\mathcal{L}$ is the loss function and $\mathcal{F}$ is a function measuring the fairness. This fairness function can be chosen in a multitude of ways and the core of this work will be about exploring different formulations of $\mathcal{F}$, their mathematical properties with regards to optimization as well as their comparative performance. The resulting model $Q(y \mid x, s; \theta)$ is a probability distribution defining the probability of an individual being a member of specific class given both their features $\boldsymbol{x}$ as well as their sensitive attribute s.

The second part of a decision making task is actually **making a decision based on a decision policy** and the predictions of $Q(y \mid x, s; \theta)$. As seen in [1] such a policy can be defined as a mapping $\pi : \mathcal{X} \times \mathcal{S} \rightarrow \mathcal{P}(0,1)$ that maps an individuals feature vector and sensitive attribute to a probability distribution over binary decisions $d \in 0, 1$. That means $\pi(d \mid x, s)$ is the probability distribution over the possible decisions, which in this case are binary, given the features and sensitive attributes of an individual. The simplest choice for a decision policy is a deterministic threshold policiy of the form

$$\pi_Q(d = 1 \mid \boldsymbol{x}, s) = \mathbf{1}[Q(y = 1 \mid \boldsymbol{x}, s) \geq c]$$

where the policy π_Q makes a positive decision $d = 1$ for all individuals for which the trained model Q exceedes a certain cofidence threshold c that the individual is part of class $y = 1$. In this scenario we assume that $Q(y = 1 \mid \boldsymbol{x}, s) \approx P(y = 1 \mid \boldsymbol{x}, s) - \delta_s$, meaning that Q is approximatly equal to the true ground truth probability distribution $P(y \mid \boldsymbol{x}, s)$. This approach directly incorporates fairness constraints by allowing group specific δ_s and therefore indirectly allowing for different thresholds c w.r.t. group affiliation.

As discussed by [1] this approach has been shown by [2] to often lead to better performance than post-processing a potentially unfair predictor. But the same paper argues that due to the fact that $Q(y = 1 \mid \boldsymbol{x}, s)$ and $P(y = 1 \mid \boldsymbol{x}, s) - \delta_s$ are not exactly equal the resulting policy π_Q will usually be suboptimal.

2 Decision making under imperfect data

The methods described so far in section 1 were built assuming i.i.d. samples to optimize the parameters θ of a given model Q. The authors of [1] explain that in many real-world applications this assumption is incorrect. In these applications the original data has not been sampeled from the true ground truth distribution $P(\boldsymbol{x}, s, y)$ but instead from a weighted distribution

$$P_{\pi_{Q_0}} = P(y \mid \boldsymbol{x}, s)\pi_0(d = 1 \mid \boldsymbol{x}, s)P(\boldsymbol{x}, s)$$

where π_0 is some intial decision policy that is employed while collecting the data. In this scenario the decision whether or not y $P(y \mid \boldsymbol{x}, s)$ comes into existence is based on the decision generated by π_0. An example of this is a loan decision scenario. If a bank wants to train a machine learning model to make predicitions on whether or not a potential client will default on their loan, they can only use historic data of past loans that they have given out. But this data is weighted by the historical decision rules based on which the bank has made their decisions so far, since they only have data for the scenario where they have granted a loan in the first place.

The authors of [1] argue that in such a scenario “for error based learning algorithms under no fairness constraints, learning within detmerninistic threshold policies is guaranteed to fail.”. They instead propose to directly learn a probabilistic decision policy π_θ that maximizes the utility, which can be formuated as the following optimization problem:

$$\begin{aligned} \underset{\theta}{\text{argmax}} \quad & u(\pi_\theta) \\ s.t. \quad & \mathcal{F}(y, x, s) \end{aligned}$$

Where $\mathcal{F}$ is again the same fairness measurement function described in 1 and u is the function that measures the utility of the learned decision policy that is parameterized by θ. The authors propse a utility function of the form $u_P(\pi) = \mathbb{E}_{x,y,s\ P(\boldsymbol{x},s,y),d\ \pi(d|\boldsymbol{x},s)}[yd - cd] = \mathbb{E}_{x,y,s\ P(\boldsymbol{x},s,y),d\ \pi(d|\boldsymbol{x},s)}[d(y - c)] = \mathbb{E}_{x,y,s\ P(\boldsymbol{x},s)}[\pi(d = 1 \mid \boldsymbol{x}, s)(P(y = 1 \mid \boldsymbol{x}, s) - c)]$

References

[1] N. Kilbertus, M. Gomez-Rodriguez, B. Schölkopf, K. Muandet, and I. Valera, “Improving consequential decision making under imperfect predictions,” *CoRR*, vol. abs/1902.02979, 2019.

[2] B. E. Woodworth, S. Gunasekar, M. I. Ohannessian, and N. Srebro, “Learning non-discriminatory predictors,” *CoRR*, vol. abs/1702.06081, 2017.