Fairness, Explainability, and Accountability for ML

Combining Human and Machine Decisions

Team:
Martin Blapp
Doruk Çetin
Bernhard Kratzwald

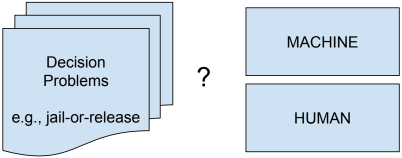
Spring 2019

Why combine human and machine decisions?

- Decision problems that need human involvement:
  - Jail-or-release
  - Stop-and-frisk
  - Accept-or-reject
- Human decision makers can profit from machine decisions:
  - Reduce the workload
  - Increase accuracy
  - Enlarge fairness

Human and machine decisions

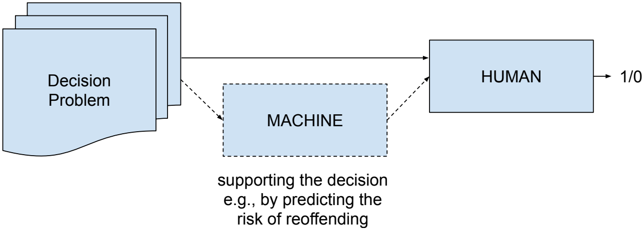
- How do we combine human and machine decisions?
- How are humans influenced by machine predictions?



Outline

1. Algorithm-in-the-loop analysis of fairness [Green and Chen (2019)]
2. Matching decision problems to humans [Valera et al. (2018)]
3. Learning to defer [Madras et al. (2018)]

Algorithmic risk assessment



+ Human makes the final decision
+ Clear responsibility

– What about fairness?
– How is the human influenced by the machine prediction?

How are humans influenced by algorithmic risk assessments?

- Experiment on 500 pre-trial cases with known ground truth y
- Treatment group w/ algorithmic assessment ($N = 6250$)
- Control group w/o algorithmic assessment ($N = 7600$)

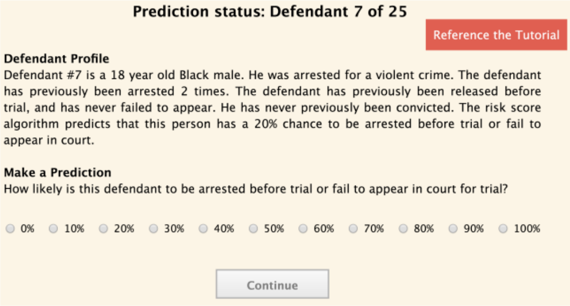


Figure: Amazon Turk experiment by Green and Chen (2019)

Performance under risk assessment

	Control	Treatment
Average reward	0,756	0,786
False positive rate	17.7%	14.8%

- Participants in the treatment group earned a 4.0% larger average reward and a 16.4% lower false positive rate than participants in the control group (both with $p < 10^{-5}$)

Performance under risk assessment

	Control	**Treatment**	**Risk assessment**
Average reward	0,756	0,786	0.807
False positive rate	17.7%	14.8%	10.1%

- Despite being presented with the risk assessment's predictions, the treatment group achieved a 2.6% lower average reward and a 46.5% higher false positive rate than the risk assessment (both with $p < 10^{-8}$)
- Only 23.7% of participants in the treatment group earned a higher average reward than the risk assessment over the course of their trial

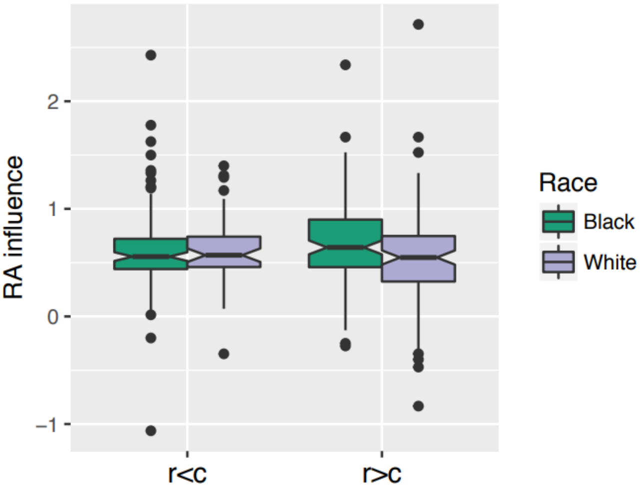
Self evaluation of participants

- The more confidence participants expressed in their predictions, the less well they actually performed ($p = 0.0186$)
- No significant relationship between the participant's evaluation of the risk assessments accuracy and actual performance
- No significant relationship between actual and perceived fairness
- Participants could generally discern how strongly they were influenced by the risk assessment

Influence of risk scores on defendants

- When risk score was lower than the average prediction in control group ($r < c$):
  - Risk assessment's influence similar regardless of the race
- When risk score was higher than the average prediction in control group ($r > c$):
  - 25.9% stronger average influence on predictions about black defendants than on predictions about white defendants
  - Risk assessment leads to larger increase in risk for black defendants

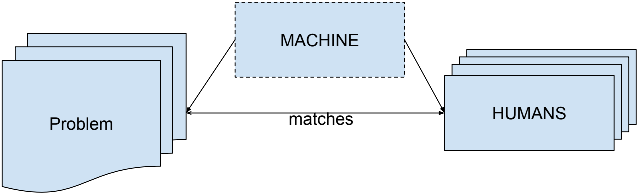
Influence of risk scores on defendants



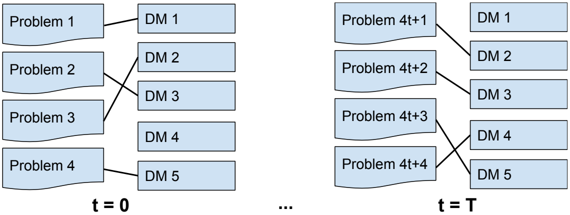
Limitations and Discussion

- Amazon Mechanical Turk and no actual judges
- Only textual description, no face-to-face
- Maybe accuracy is not the metric we should optimize?
- Maybe risk assessment is not the optimal pattern?

Alternative pattern: Problem matching



Problem Definition



each decision:

$$d_j(X_i, S_i) \to Y,$$

where $\mathrm{X}_i \in \mathbb{R}^d, S_i \in \{0,1\}, Y_i \in \{0,1\}$

we assume $\mathrm{P}(\mathrm{Y}_i | X_i, S_i)$ known to decision makers, but each decision maker has own thresholds $\theta_{j,s}$

$$d_j(X_i, S_i) = \begin{cases} 1, & \text{if } P(Y_i = 1 | X_i, S_i) \geq \theta_{j,S_i} \\ 0, & \text{otherwise} \end{cases}$$

Metrics

We measure

- Utility

$$u(d, c) = \sum_{i \in \{decisions\}} Y_i d(X_i, S_i) - c d(X_i, S_i)$$

- Fairness Constraints
  $\Rightarrow$ Disparate Impact

$$b_s = \mathbb{E}[1 - d(X, S = s)]$$

$$DI = |b_{s=1} - b_{s=0}| \leq \alpha$$

See also [Corbett-Davies et al. (2017)]

Matching with known tresholds

- Simple Case: Assume we know how humans decide, $\theta_{j,s}$ are known $\Rightarrow$ maximum weighted bipartite matching
$$w_{ji} = \begin{cases} P(Y_i = 1|X_i, S_i) - c, & \text{if } P(Y_i = 1|X_i, S_i) \geq \theta_{j,S_i} \\ 0, & \text{otherwise} \end{cases}$$

Matching with unknown thresholds

- If human thresholds are unknown
  - initialize with prior $\theta_j(0) \sim Beta(\alpha, \beta)$
  - for each new round $\theta_j(t+1) \sim p(\theta_j(t)|D(t))$
  - maximum weighted bipartite matching
- Regret: $R(T) = u^*(d, c) - u(d, c)$
  - expected regret shrinks in $O(\sqrt{T})$

Fairness Constraints

When enforcing fairness constraints

$$DI = |b_{s=1} - b_{s=0}| \leq \alpha$$

matching must satisfy for each s:

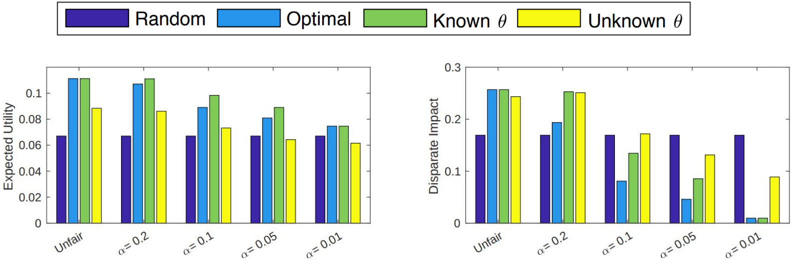
$$m_{S=s}(b^*_{S=s} - \alpha) \leq \sum_{\forall (i,j), S=s} \mathbb{1}(w_{ji} = 0)$$

$$\sum_{\forall (i,j), S=s} \mathbb{1}(w_{ji} = 0) \leq m_{S=s}(b^*_{S=s} + \alpha)$$

where $j \in \{humans\}, i \in \{problems\}, m_{S=s} :=$ #decisions with sensitive attribute s

bounded color matching problem $\Rightarrow$ bi-criteria algorithm with $\frac{1}{2}$-approximation guarantee.

Results

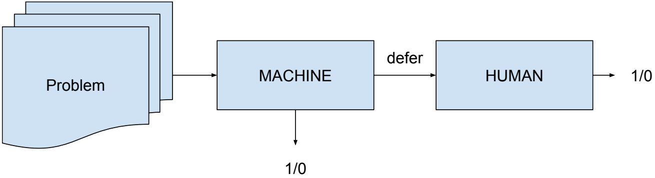


- wide range of θ_j beneficial

Limitations

- Assuming $P(Y|X, S)$ known to each DM
- 1-Human to 1-Problem matching

An alternative pattern: PASS option



Rejection learning

$$\mathcal{L}_{reject}(Y, \hat{Y}_M, \hat{Y}_D, s) = -\sum_i [(1 - s_i)\ell(Y_i, \hat{Y}_{M,i}) + s_i \gamma_{reject}]$$

- $\hat{Y}_M$: decision of the machine learning model
- $\hat{Y}_D$: decision of the external decision maker
- s: gating variable (1 for rejections, 0 otherwise)

Find more details here: [Cortes et al. (2016)]

Learning to defer (Madras et al., 2018)

$$\mathcal{L}_{reject}(Y, \hat{Y}_M, \hat{Y}_D, s) = -\sum_i [(1 - s_i)\ell(Y_i, \hat{Y}_{M,i}) + s_i\gamma_{reject}] \quad (1)$$

$$\mathcal{L}_{defer}(Y, \hat{Y}_M, \hat{Y}_D, s) = -\sum_i [(1 - s_i)\ell(Y_i, \hat{Y}_{M,i}) + s_i\ell(Y_i, \hat{Y}_{D,i}) + s_i\gamma_{defer}] \quad (2)$$

Setup

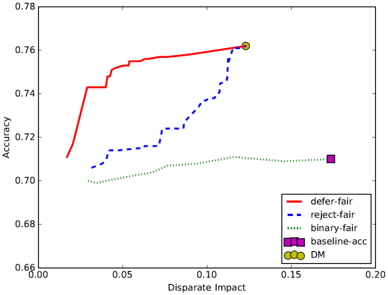
- COMPAS and Heritage Health datasets
- Equalized odds as fairness metric (Disparate impact as regularizer)
- “Semi-synthetic data”: simulated DMs on real data

Scenarios

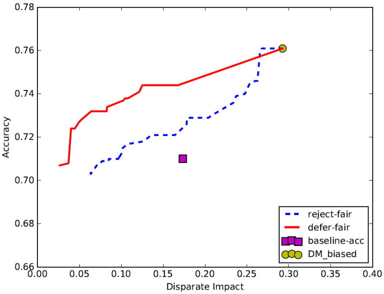
A) High-accuracy DM: ignores fairness
B) Highly-biased DM: strongly unfair
C) Inconsistent DM: ignores fairness (noisy)

In each scenario DM receives extra information (one feature) in training.

Results: Scenarios A and B

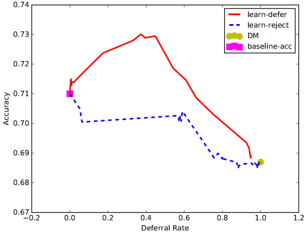


(a) High-accuracy DM

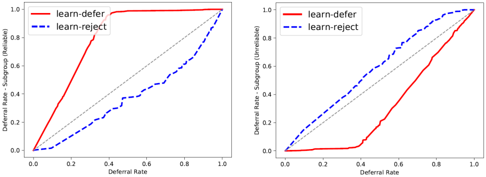


(b) Highly-biased DM

Results: Scenario C



(a) Inconsistent DM



(b) Deferral Rates

Why learning to defer?

- Adaptive rejection
- Considering the model impact
- Predicting responsibly

Conclusion

- Great potential lies in the cooperation
  - Polson and Scott (2018)
- Many nuances and pitfalls
  - Hamilton (2015)
- Increasing importance

References

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. (2017). Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, pages 797–806, New York, NY, USA. ACM.

Cortes, C., DeSalvo, G., and Mohri, M. (2016). Learning with rejection. In *International Conference on Algorithmic Learning Theory (ALT 2016)*.

Green, B. and Chen, Y. (2019). Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 90–99. ACM.

Hamilton, M. (2015). Adventures in risk: predicting violent and sexual recidivism in sentencing law. *Ariz. St. LJ*, 47:1.

Madras, D., Pitassi, T., and Zemel, R. (2018). Predict responsibly: Improving fairness and accuracy by learning to defer. In *Advances in Neural Information Processing Systems*, pages 6147–6157.

Polson, N. and Scott, J. (2018). *AIQ: How artificial intelligence works and how we can harness its power for a better world*. Random House.

Valera, I., Singla, A., and Rodriguez, M. G. (2018). Enhancing the accuracy and fairness of human decision making. In *Advances in Neural Information Processing Systems*, pages 1769–1778.

Hypothesis

- **H1 (Performance):** Participants presented with a risk assessment will make predictions that are less accurate than the risk assessment's.
- **H2 (Evaluation):** Participants will be unable to accurately evaluate their own and the algorithm's performance.
- **H3 (Bias):** As they interact with the risk assessment, participants will be disproportionately likely to increase risk predictions about black defendants and to decrease risk predictions about white defendants.

Evaluation

- Reward: $r = [1 - (\text{prediction} - \text{outcome})^2]$
- Risk-score influence on defendant j:

$$I_j = \frac{t_j - c_j}{r_j - c_j}$$

- Influence of risk assessment on participant k is:

$$I^k = \frac{1}{25} \sum_{i=1}^{25} \frac{p_i^k - c_i}{r_i - c_i}$$

r_j ... prediction made by risk assessment
t_j ... avg. prediction of treatment group on defendant j
c_j ... avg. prediction of control group on defendant j
p_i^k ... prediction of participant k on defendant j

Learning to defer loss in detail

Final loss for learning to defer:

$$\begin{aligned}\mathcal{L}_{defer}(Y, \hat{Y}_M, \hat{Y}_D, \pi; \theta) &= \mathbb{E}_{s \sim Ber(\pi)} \mathcal{L}(Y, \hat{Y}_M, \hat{Y}_D, s; \theta) \\ &= \sum_i \mathbb{E}_{s \sim Ber(\pi)} [(1 - s_i)\ell(Y_i, \hat{Y}_{M,i}; \theta) + s_i \ell(Y_i, \hat{Y}_{D,i})]\end{aligned}$$

Loss function with fairness regularization:

$$\mathcal{L}_{defer}(Y, \hat{Y}_M, \hat{Y}_D, \pi; \theta) = \mathbb{E}_{s \sim Ber(\pi)} \mathcal{L}(Y, \hat{Y}_M, \hat{Y}_D, s; \theta) + \alpha_{fair} \mathcal{R}(Y, \hat{Y}_M, \hat{Y}_D, s)$$

The regularization term is a continuous relaxation of disparate impact (DI) as

$$\mathcal{R}(Y, \hat{Y}_M, \hat{Y}_D, s) = \frac{1}{2}(DI_{Y=0}(Y, A, \hat{Y}) + DI_{Y=1}(Y, A, \hat{Y}))$$

where

$$DI_{Y=i}(Y, A, \hat{Y}) = |\mathbb{E}_{\hat{Y} \sim Ber(p)}(\hat{Y} = 1 - Y | A = 0, Y = i) - \mathbb{E}_{\hat{Y} \sim Ber(p)}(\hat{Y} = 1 - Y | A = 1, Y = i)|$$