Value alignment and control

- King Midas, paperclips and trolleys
- **Measure what matters and manage tradeoff**
- Prediction versus optimization and control
- Maintaining human oversight

Measuring impact holistically

We have established numerous dimensions along which performance must be measured. Some will be quantitative (numeric score), whereas other quantitative (low/med/high). These include:

- Fairness (quantitative): e.g., demographic disparity, difference in odds, or in opportunity
- Privacy (quantitative/qualitative): e.g., the k in k-anonymity, or randomization in randomized response
- Explainability (low/med/high): to be discussed next week
- Safety and security (low/med/high): to be discussed in the last week

We must then also consider costs “external” to the AI:

- User welfare (might be in the form of a misclassification cost, or opportunity cost)
- Environmental impact
- Social welfare

Pareto fronts

We can view accuracy as a loss function:

$$\underset{\theta}{\operatorname{argmin}}\, \mathbb{E}[L(y;\hat{y})], \text{ where } \hat{y} = f_\theta(X)$$

Or even more generally as a special cases of optimizing a loss L(θ), observed with noise. Then ROC curve analysis involves optimising two objectives at the same time:

$$g_1(\theta) = \text{Sens}(y,\hat{y}) \qquad g_2(\theta) = \text{Sens}(y,\hat{y})$$

Pareto fronts

We can view accuracy as a loss function:

$$\underset{\theta}{\operatorname{argmin}}\, \mathbb{E}[L(y;\hat{y})], \text{ where } \hat{y} = f_\theta(X)$$

Or even more generally as a special cases of optimizing a loss L(θ), observed with noise. Then ROC curve analysis involves optimising two objectives at the same time:

$$g_1(\theta) = \text{Sens}(y,\hat{y}) \qquad g_2(\theta) = \text{Sens}(y,\hat{y})$$

Imperial College London

ROC curves are a special case of multi-objective optimization problems (MOOPs):

MOOP: $\underset{\theta \in \Theta}{\operatorname{argmin}} (g_1(\theta), \ldots, g_m(\theta)$

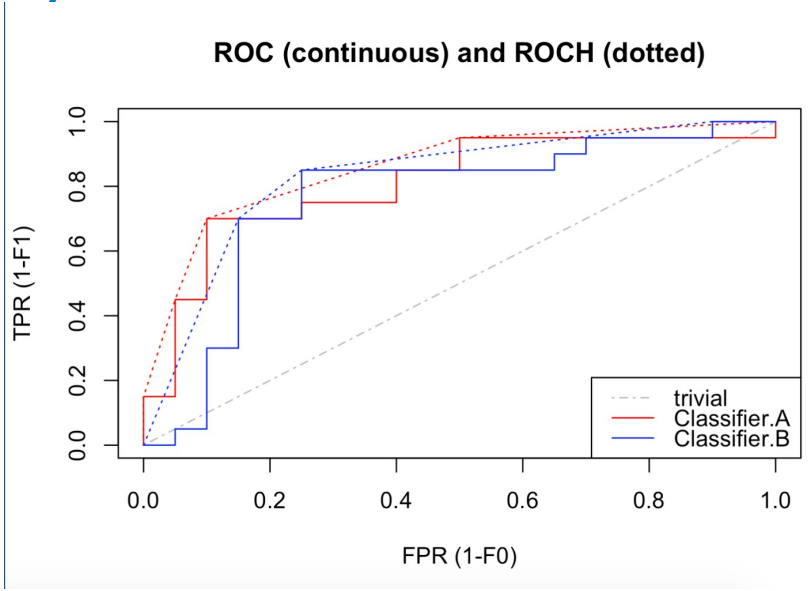
Objective vector:

$$z^\star = (g_1(\theta^\star), \ldots, g_m(\theta^\star)$$

θ *dominates* ψ if:

$g_i(\theta) \leq g_i(\psi)$, for all $i = 1, \ldots, k$ and

$g_i(\theta) < g_i(\psi)$, for at least one i

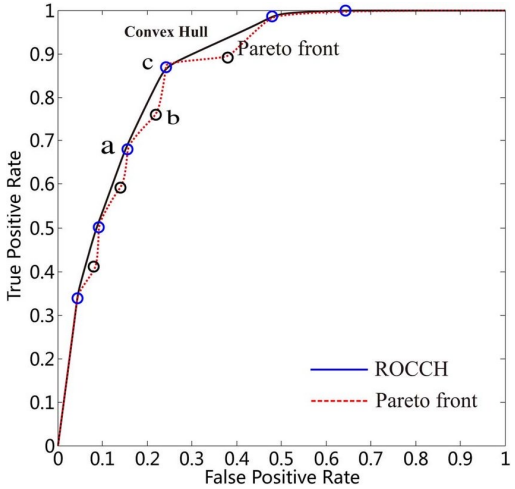


Imperial College London

ROC curves are a special case of multi-objective optimization problems (MOOPs):

Pareto front is set of dominant solutions. Convex hull is a subset of Pareto front. These concepts appear in almost all disciplines (e.g., efficient frontier in finance)

Human decision makers can decide among Pareto optimal solutions *a priori* (by stating a preference for the relative importance of the competing objectives) or *a posteriori.*



Wang, Pu & Emmerich, Michael & Li, Rui & Tang, Ke & Bäck, Thomas & Yao, Xin. (2013). Convex Hull-Based Multi-objective Genetic Programming for Maximizing ROC Performance.

Imperial College London

A priori methods for MOOPs come in many flavours

Utility function. Sometimes the human decision maker can specify a utility function *u(g(.))* that captures their preferences, i.e., u(g(θ)) > u(g(ψ)) when θ ıs preferable to ψ.

In such a situation the problem becomes single-objective.

The MOOP framing is still useful as it helps validate and stress-test the utility function.

G. Chiandussi, M. Codegone, S. Ferrero, F.E. Varesio, Comparison of multi-objective optimization methodologies for engineering applications, Computers & Mathematics with Applications, Volume 63, Issue 5, 2012

Imperial College London

A priori methods for MOOPs come in many flavours

Linear scalarisation. Alternatively, a function can be introduced that helps the decision maker state their preferences in a way that respects Pareto optimality. Most commonly:

$$s(g_1(\theta), \ldots, g_m(\theta)) = \sum_{i=1}^{m} w_i g_i(\theta), \text{ where } w_i > 0 \text{ and } \sum_i w_i = 1$$

The total misclassification cost described last week is a scalarization technique over false positives and false negatives, with the relative misclassification costs as weights.

G. Chiandussi, M. Codegone, S. Ferrero, F.E. Varesio, Comparison of multi-objective optimization methodologies for engineering applications, Computers & Mathematics with Applications, Volume 63, Issue 5, 2012

Imperial College London

A priori methods for MOOPs come in many flavours

Constraint scalarization. Yet another very pragmatic solution is to only minimise one objective subject to the constraint that all other objectives stay below given thresholds:

$$\min g_1(\theta) \text{ subject to } g_i(\theta) < t_i \text{ for } i > 1$$

Commonly secondary and tertiary objectives will be expressed as constraints (e.g., keep equalized odds within ε in a fairness example), but they are still fundamentally a MOOP.

G. Chiandussi, M. Codegone, S. Ferrero, F.E. Varesio, Comparison of multi-objective optimization methodologies for engineering applications, Computers & Mathematics with Applications, Volume 63, Issue 5, 2012

Imperial College London

A posteriori and progressive methods give the chance to the decision maker to offer feedback on solutions

Preference articulation by the human decision maker can occur a priori, or a posteriori (with the Pareto optimal candidates at hand) or progressively, where the decision maker is stating a preference for an improvement in the local neighbourhood of a solution.

$$\begin{aligned} \text{A priori:} &\quad \text{Decide} \Rightarrow \text{Search} \\ \text{A posteriori:} &\quad \text{Search} \Rightarrow \text{Decide} \\ \text{Progressive:} &\quad \text{Search} \Leftrightarrow \text{Decide} \end{aligned}$$

G. Chiandussi, M. Codegone, S. Ferrero, F.E. Varesio, Comparison of multi-objective optimization methodologies for engineering applications, Computers & Mathematics with Applications, Volume 63, Issue 5, 2012

Imperial College London

Summary

- Multi-objective optimisation is a powerful framework to express attempts to optimise competing objectives, as is often the case in ethical AI.
- ROC curves are a special case of a MOOP formulation.
- Identifying Pareto optimal solutions can enable the human decision maker to focus only on the trade-offs, without ever losing an opportunity to improve on all fronts
- A priori preference articulation is one approach towards solving MOOPs. It can take the form of a single utility function, relative weights on different objectives enabling linear scalarization; or “minimum thresholds” on all but a primary objective.
- Alternatively, the decision maker can be consulted a posteriori once a set of candidate solutions have been identified, to avoid having an exhaustive preference articulation.

G. Chiandussi, M. Codegone, S. Ferrero, F.E. Varesio, Comparison of multi-objective optimization methodologies for engineering applications, Computers & Mathematics with Applications, Volume 63, Issue 5, 2012