Table of contents

- Real-world examples
- Error parity and confusion matrices
- Other fairness metrics
- **Pareto fronts**

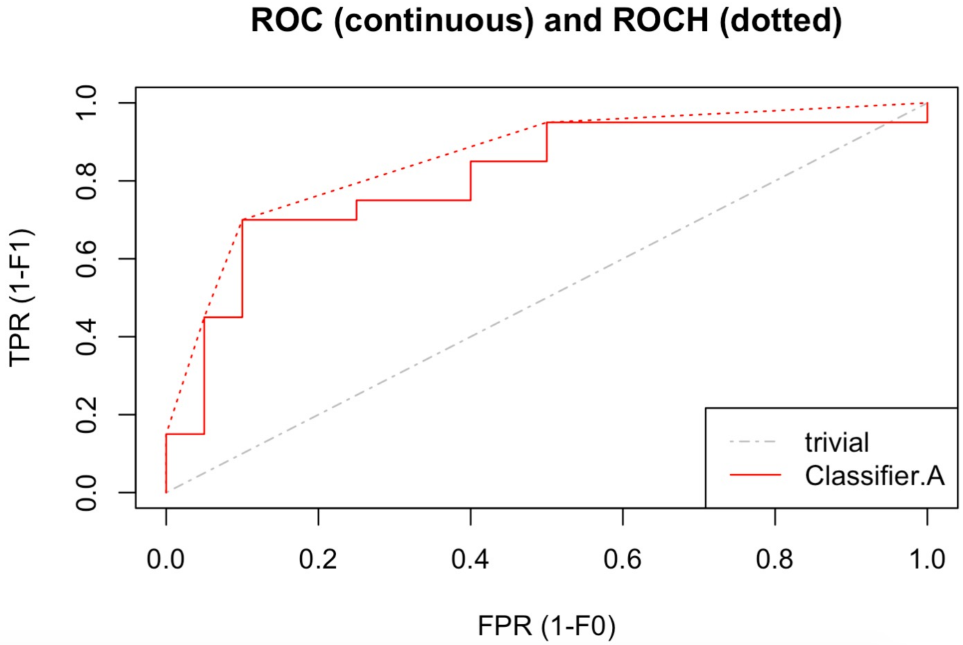
An explicit tradeoff is better than an implicit one

- We often face tradeoffs in supervised learning: a tradeoff between false negatives and false positives, and between different fairness metrics.
- A useful tool to help us navigate tradeoffs is the notion of Pareto fronts.
- In general, *we prefer making tradeoffs as explicit as possible*

ROC curves

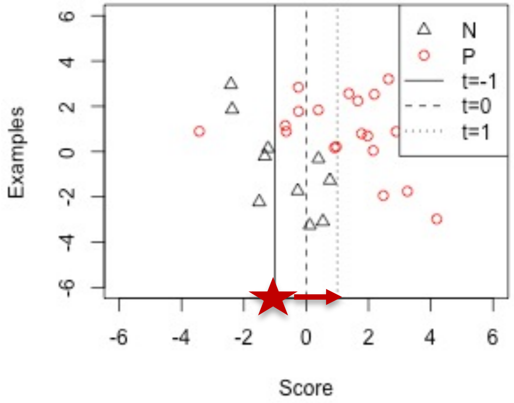
A Receiver Operating Characteristic (ROC) curve is a plot of sensitivity vs 1-specificity.

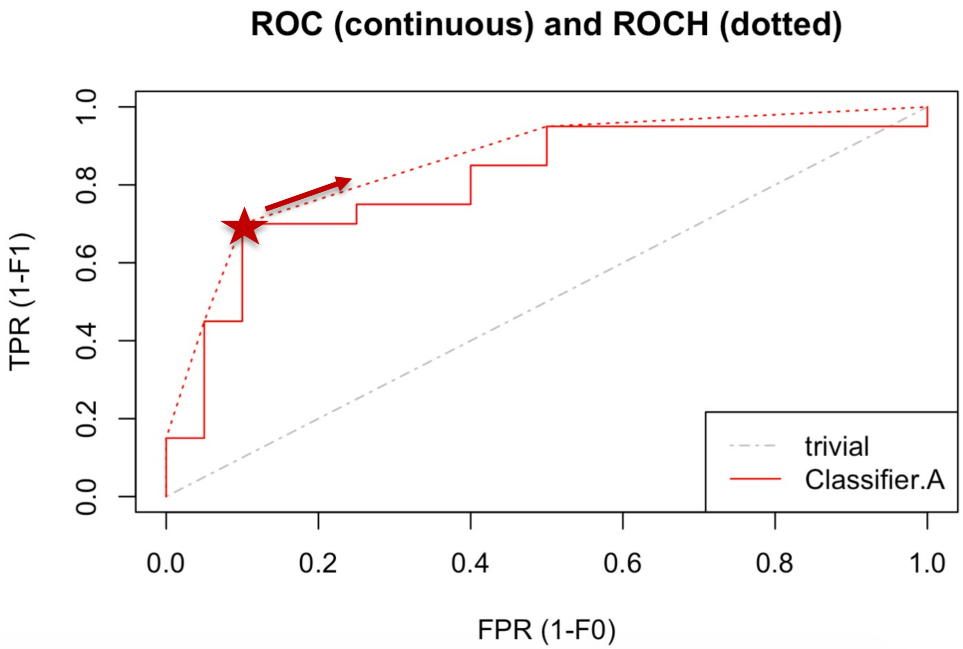
The diagonal line represents a coin toss, and the perfect classifier would be the line from (0,0) to (0,1) to (1,1).



ROC curves

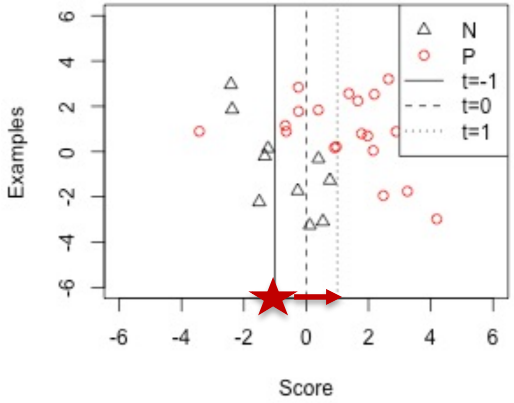
As we move the threshold from left to right, we obtain a new pair of *increasing* TPR, FPR values.

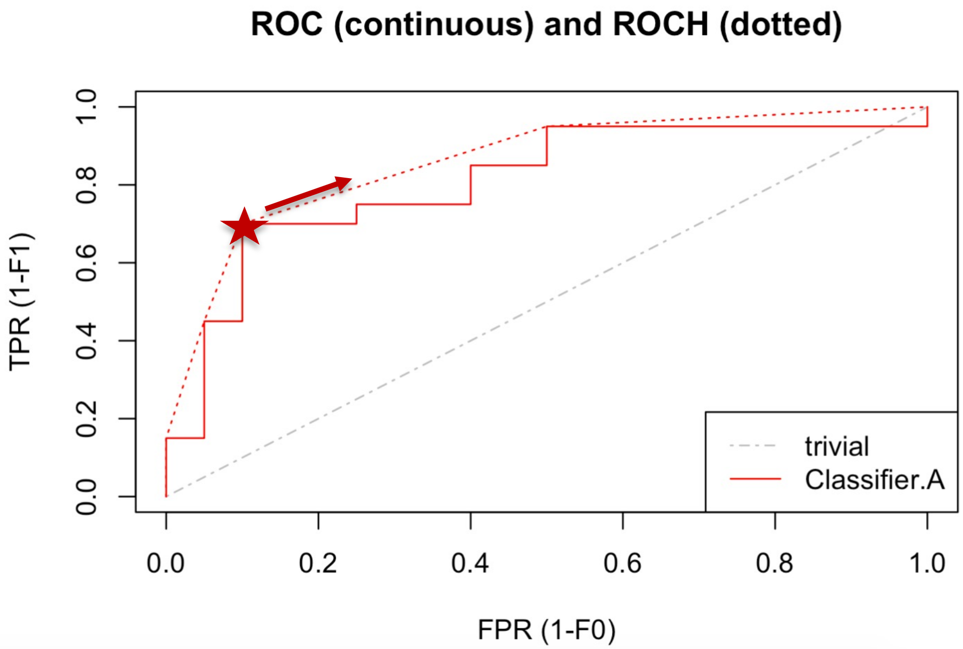




ROC curves

The objective is to increase TPR while making FPR increase too much, though inevitably it eventually will

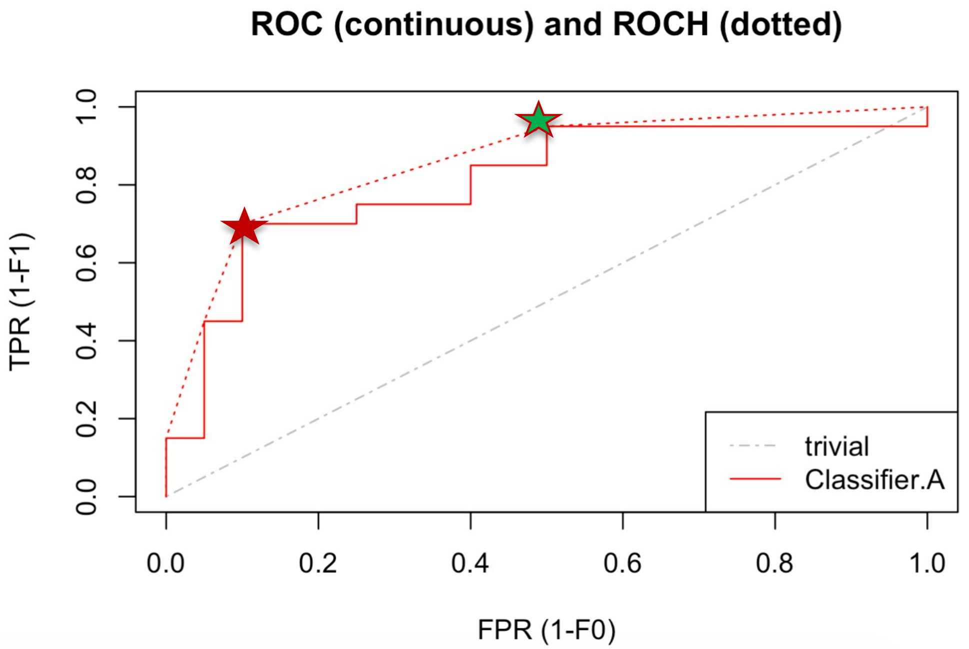




ROC curves

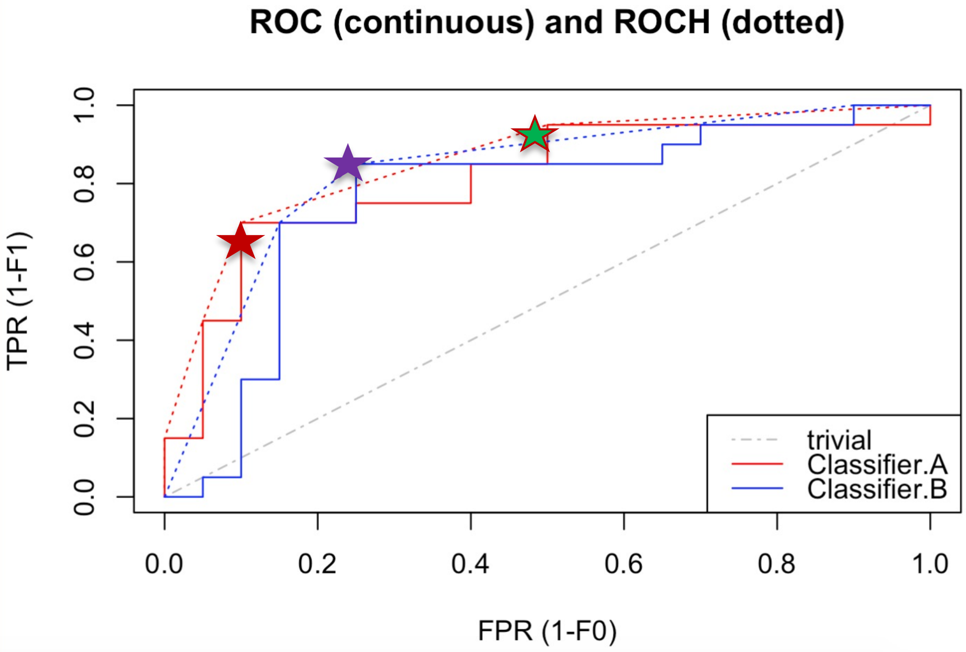
Selecting between different points on the ROC curve (i.e., different classification thresholds) is entirely dependent on the relative misclassification cost of false positives over false negatives

- ★ Cancer diagnosis
- ★ Spam filter



ROC curves

The same principle applies when comparing multiple classifiers. Though in some cases one might dominate, often that is not the case, and the “best” classifier will depend on the relative misclassification cost, too.

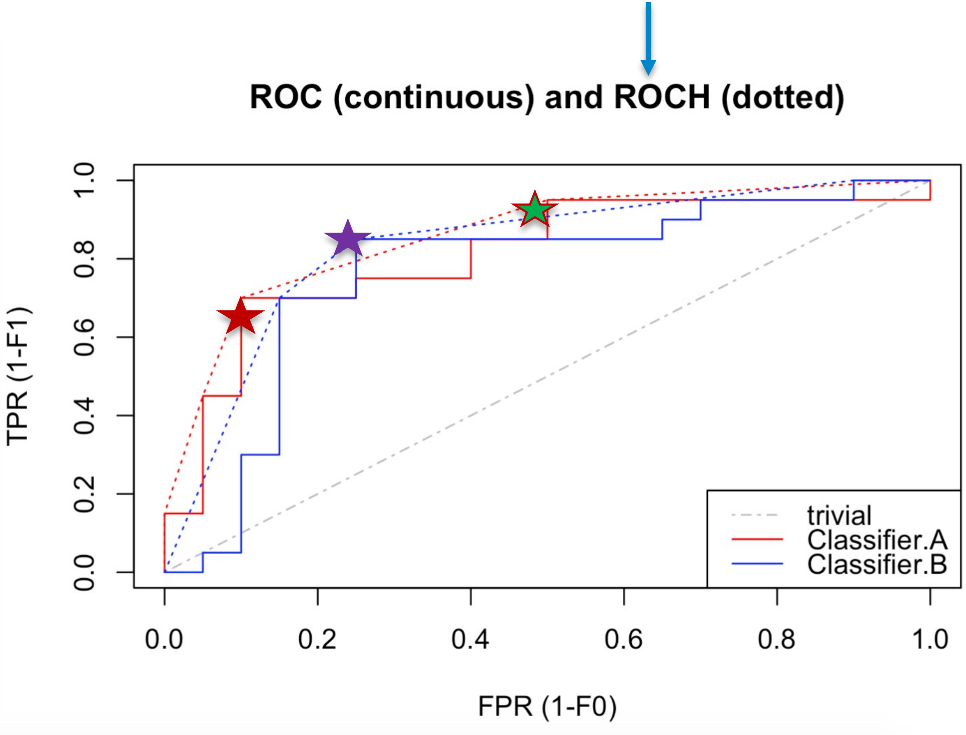


ROCH and Pareto

ROCH is the convex hull of the ROC curve, which is the tightest fitting convex curve to this data.

More generally, this is known as a Pareto front: we would not want to ever select a solution that could be improved in at least one direction without a deterioration in the other.

It is possible to use randomization to construct a classifier that has this precise performance.



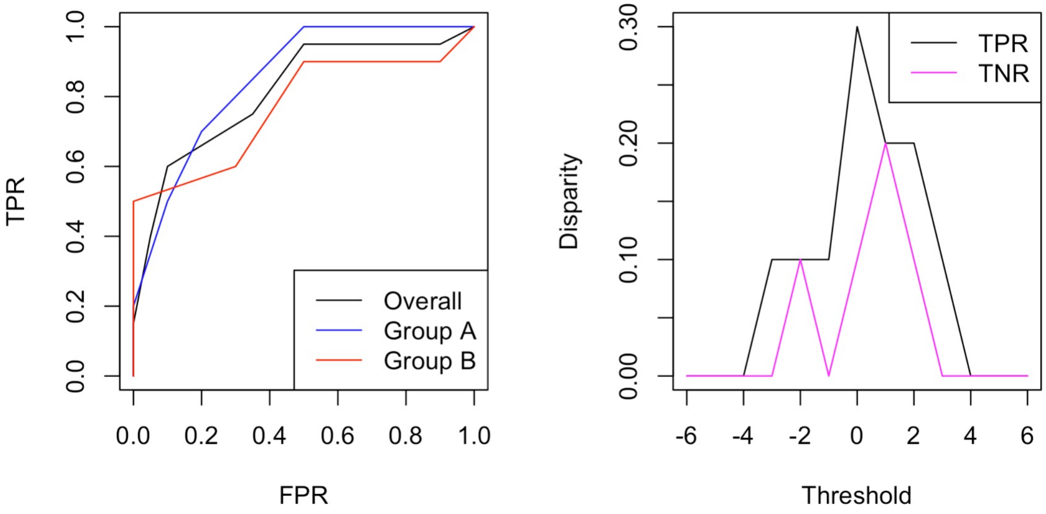
Group-specific ROC curves

ROC curves demonstrate that:

- A given classifier can produce a whole sequence of (FPR,TPR) pairs.
- We can choose among those pairs by setting the classification threshold.

This also suggests that different choices of the threshold will result in different degrees of discrepancy between FPR for men and for women, for example.

Group-specific ROC curves



Explicit tradeoffs

- Minimise total cost, where c is the cost of a false positive and (1-c) the cost of a false negative:

$$L = c \cdot Pr(\hat{Y} = 1 \mid Y = 0) Pr(Y = 0) + (1 - c) \cdot Pr(\hat{Y} = 0 \mid Y = 1) Pr(Y = 1)$$

- Minimise total cost subject to keeping difference between group FPRs below ε:

$$|Pr(\hat{Y} = 1 \mid Y = 1, A = 1) - Pr(\hat{Y} = 1 \mid Y = 1, A = 0)| < \epsilon$$

In general, we might decide to:

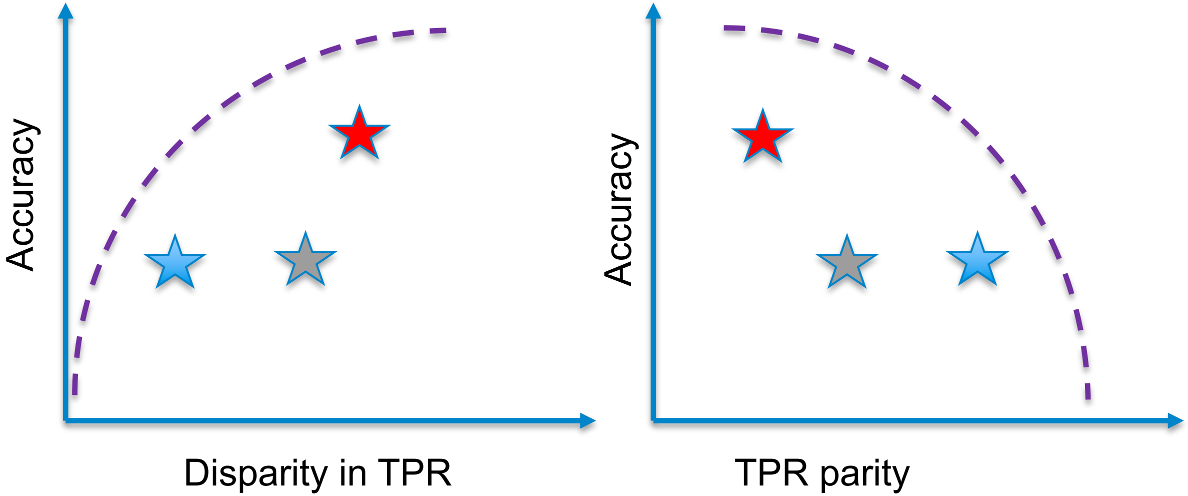
- Choose a classification threshold that minimizes the TPR discrepancy
- Choose different thresholds for males than for females
- Fit altogether different classifiers to males than females
- Change the loss function that the classifier is optimizing to capture a fairness constraint

Explicit tradeoffs

★ : fairness at the expense of accuracy

★ : accuracy at the expense of fairness

★ : inappropriate



Summary

- We introduced the notion of a ROC curve as an explicit representation of the tradeoffs between false positives and false negatives, as a function of the choice of classification threshold
- We briefly introduced the notion of a Pareto front as the set of dominant solutions in a multi-objective optimisation solution space, and explained how it motivates the variant of ROC called ROCH
- We then explained how in fairness, ROC curve analysis must take into account the group-specific ROC curves, which might motivate a different choice of threshold or classifier in total, or per group.