Model-Based Preference Optimization:

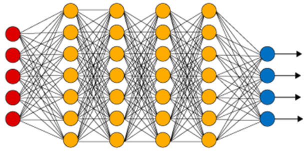
Metric Elicitation

Sanmi Koyejo

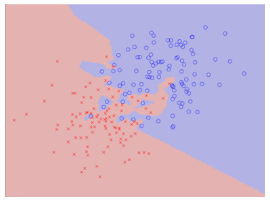
All Sections

Export PDF

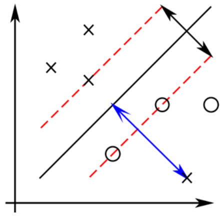
1. Evaluating Models



- 94.1% Accuracy
- 90% false positives
- 5% false negatives



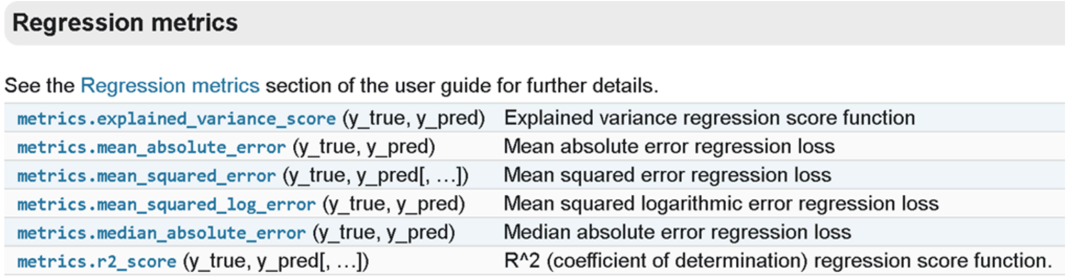
- 89.6% Accuracy
- 50% false positives
- 1% false negatives



- 80.1% Accuracy
- 10% false positives
- 20% false negatives

- What are the tradeoffs of the 3 binary classifiers above? Which would you choose?
- You'll realize that it depends on the context in which the classifier is being used.
- In the case of cancer diagnosis, a false negative could result in death. However, in the context of recidivism prediction, a false positive means unjustly putting someone behind bars for too long.

1. Evaluating Models



Regression metrics

See the Regression metrics section of the user guide for further details.

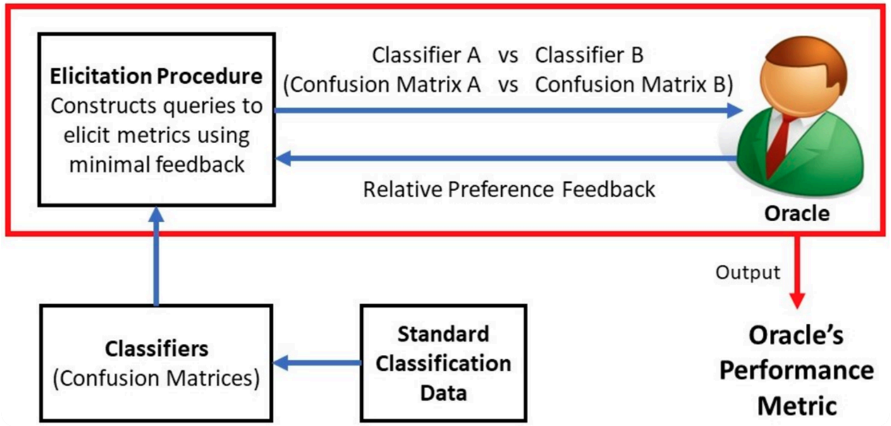
`metrics.explained_variance_score` (y_true, y_pred)	Explained variance regression score function
`metrics.mean_absolute_error` (y_true, y_pred)	Mean absolute error regression loss
`metrics.mean_squared_error` (y_true, y_pred[, …])	Mean squared error regression loss
`metrics.mean_squared_log_error` (y_true, y_pred)	Mean squared logarithmic error regression loss
`metrics.median_absolute_error` (y_true, y_pred)	Median absolute error regression loss
`metrics.r2_score` (y_true, y_pred[, …])	R^2 (coefficient of determination) regression score function.

- When evaluating models, we need to consider the relative cost/benefit of different kinds of errors.
- The **metric** is a quantitative description of tradeoffs.
- For many tasks in regression and classification, a variety of popular metrics are available off the shelf.

1. Evaluating Models

- One must be careful in choosing the appropriate metrics for a given task. Metrics are not exchangeable because different metrics evaluate different areas of model performance.
- Cremonesi, Koren, Turrin (2010) found that the RMSE metric used for evaluating models submitted to the Netflix Prize competition did not translate well to top-N ranking accuracy which more directly impacts what users see.
- In some cases, metrics can even be contradictory. The COMPAS model used to predict rescindivism in criminal cases was calibrated on equal accuracy across demographic groups. However, an analysis by ProPublica found "blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend," despite the calibration.

2. Metric Elicitation through Preferences



Elicitation Procedure
Constructs queries to elicit metrics using minimal feedback

Classifier A vs Classifier B
(Confusion Matrix A vs Confusion Matrix B)

Relative Preference Feedback

Oracle

Output

Oracle's Performance Metric

Classifiers
(Confusion Matrices)

Standard Classification Data

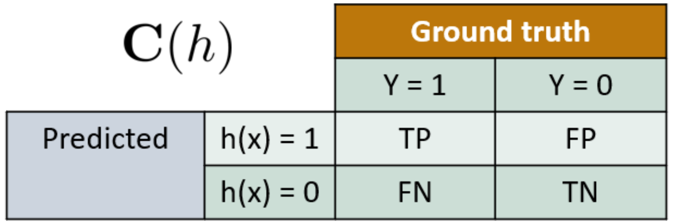
- Determining metrics by interacting with individual stakeholders
  - Hiranandani et al. "Performance metric elicitation from pairwise classifier comparisons."
  - Hiranandani et. al "Multiclass Performance Metric Elicitation"
  - Hiranandani et. al., "Fair Performance Metric Elicitation"
- Metric elicitation from stakeholder groups
  - Robertson et. al., "Probabilistic Performance Metric Elicitation

2. Metric Elicitation through Preferences

- Preference elicitation is studied in economics and psychology (Samuelson, 1938; Varian, 2005)
- Elicitation is related to (contextual dueling) bandits, when focused on recovering reward function (Yue et. al. 2012; Dudik et. al. 2015)
- Inverse reinforcement learning generalizes elicitation when reward is transportable (Amin, 2017)

2. Metric Elicitation through Preferences

$\mathbf{C}(h)$



		Ground truth	
		Y = 1	Y = 0
Predicted	h(x) = 1	TP	FP
	h(x) = 0	FN	TN

- We study work by Hiranandani et al. (YEAR) about eliciting linear binary classification metrics in the noise-free setting.
- Assume we have a data generating distribution $\mathcal{P}$ across the space of inputs X and outputs $Y \in \{0, 1\}$. The set of all classifiers is denoted by $\mathcal{H} = \{h : X \mapsto [0, 1]\}$. For a given classifier h, its confusion matrix entries $\mathbf{C}(h)$ are
  - $TP = \Pr(Y = 1, h = 1)$
  - $FP = \Pr(Y = 0, h = 1)$
  - $FN = \Pr(Y = 1, h = 0)$
  - $TN = \Pr(Y = 0, h = 0)$

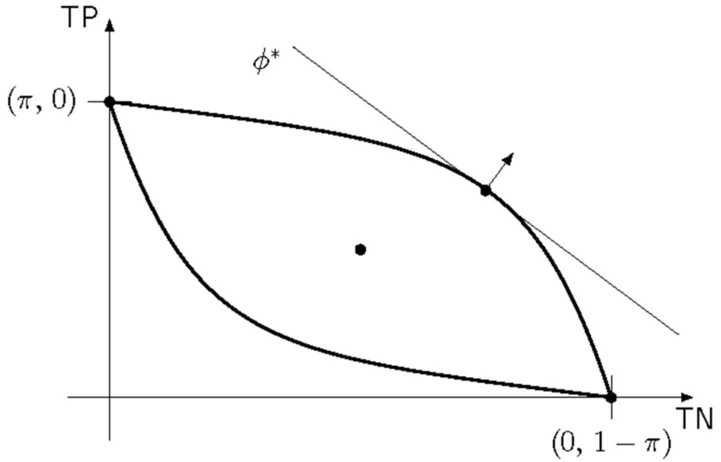
2. Metric Elicitation through Preferences

- The metric used by the **oracle** to evaluate a classifier h is given by

$$\phi^*(h) = 1 - (a_1^* FP(h) + a_2^* FN(h)).$$

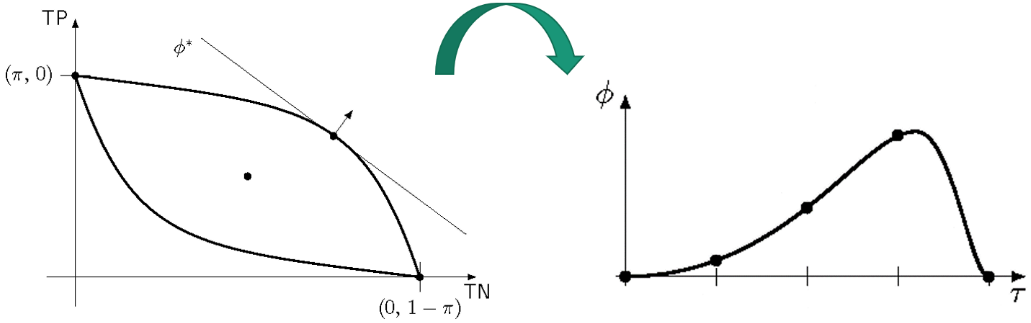
- We do not know a_1^* and a_2^* (which weigh the relative cost of errors between FP and FN) a priori. The objective is to find $\mathbf{argmax}_h \phi^*(h)$ by querying the oracle with pairs of classifiers confusion matrices $\mathcal{C}(h_1), \mathcal{C}(h_2)$ upon which the oracle returns the preferred classifier w.r.t. ϕ^*. Ideally, we want to minimize the number of queries used (i.e., the query/sample complexity).

2. Metric Elicitation through Preferences



- The above plot illustrates the region of attainable TP and TN values across all classifiers ($\mathcal{H}$). We note that given TP and TN, FN and FP are derivable independent of classifier so there are 2 degrees of freedom.
- The region can be shown to be convex. Since ϕ^* is a linear function on it, the optimal classifier must lie on the boundary!

2. Metric Elicitation through Preferences



- As a result, we unroll the boundary and use a binary search style algorithm to construct queries for the oracle.
- The authors found that the resulting classifier was guaranteed to be ϵ-accurate in $O(\log \frac{1}{\epsilon})$ queries. This was true even under system noise (e.g., noisy responses from the oracle using probabilistic extensions).
- Work that came after extended to multiclass classification, more complex metrics, and stakeholder groups.