Post-processing Bias Mitigation

ROC
(Receiver Operating Characteristic)

Slides merging Lectures from:

Bradley Malin - Carnegie Mellon University

And

Ricco RAKOTOMALALA - Université Lumière Lyon 2

Why Should I Care?

- Imagine you have 2 different probabilistic classification models
  - e.g. logistic regression vs. neural network
- How do you know which one is better?
- How do you communicate your belief?
- Can you provide quantitative evidence beyond a gut feeling and subjective interpretation?

Recall Basics: Contingencies

		MODEL PREDICTED	
		It's NOT a Heart Attack	*Heart Attack!!!*
GOLD STANDARD TRUTH	*Was NOT a Heart Attack*	A	B
	Was a Heart Attack	C	D

Some Terms

		MODEL PREDICTED	
		NO EVENT	*EVENT*
GOLD STANDARD TRUTH	*NO EVENT*	TRUE NEGATIVE	B
	EVENT	C	TRUE POSITIVE

Some More Terms

		MODEL PREDICTED	
		NO EVENT	*EVENT*
GOLD STANDARD TRUTH	*NO EVENT*	A	FALSE POSITIVE (Type 1 Error)
	EVENT	FALSE NEGATIVE (Type 2 Error)	D

Accuracy

- What does this mean?
- What is the difference between “accuracy” and an “accurate prediction”?
- Contingency Table Interpretation

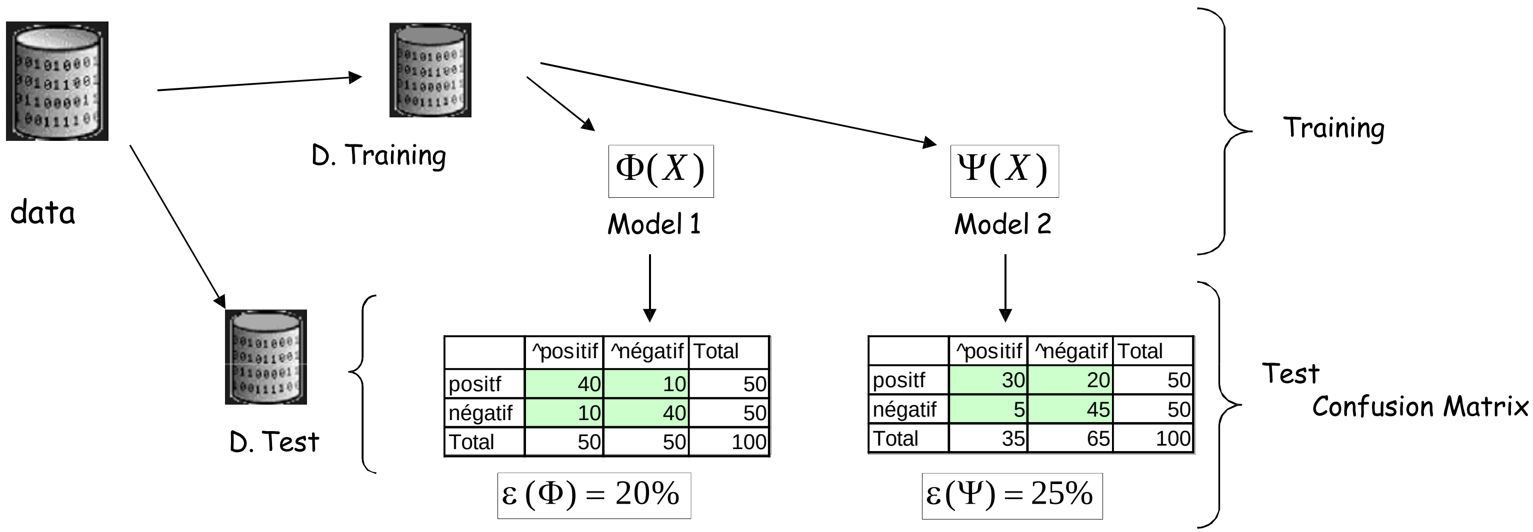
$$\frac{\text{(True Positives)} + \text{(True Negatives)}}{\text{(True Positives)} + \text{(True Negatives)} + \text{(False Positives)} + \text{(False Negatives)}}$$

- Is this a good measure? (Why or Why Not?)

Accuracy?

Accuracy: a too restrictive measure

Standad schema of models' evaluation



	^positif	^négatif	Total
positf	40	10	50
négatif	10	40	50
Total	50	50	100

$\varepsilon(\Phi) = 20\%$

	^positif	^négatif	Total
positf	30	20	50
négatif	5	45	50
Total	35	65	100

$\varepsilon(\Psi) = 25\%$

Conclusion: Model 1 would be better than Model 2

This conclusion –- based on Test dataset – assumes that the Matrix of misclassification cost is sysmetric and unitary

!

Accuracy?

Introduction of Error cost matrix

Misclassification cost non-symetric

	^positive	^negative
Positve	0	1
Negative	10	0

	^positive	^negative	Total
Positive	40	10	50
Negative	10	40	50
Total	50	50	100

$c(\Phi) = 1.1$

	^positive	^negative	Total
Positive	30	20	50
Negative	5	45	50
Total	35	65	100

$c(\Psi) = 0.7$

Conclusion: Model 2 would be better than Model 1 in this case???

?

Cost matrices are often the result of cyclical opportunities.
Should we test all possible cost matrices to compare M1 and M2?

Is it possible to benefit from a system that allows for global comparison of models, regardless of the misclassification cost matrix?

ROC Curve

Objectives of the ROC Curve

1. The ROC curve is a tool for evaluating and comparing models Independent of misclassification cost matrices
   It provides insight into whether M1 will always be better than M2 regardless of the cost matrix

2. Operational even in the case of very unbalanced distributions
   Without the perverse effects of the confusion matrix related to the need to carry out an assignment

3. A graphical tool that allows you to visualize performance
   A single glance should allow us to see the model(s) that are likely to be of interest to us

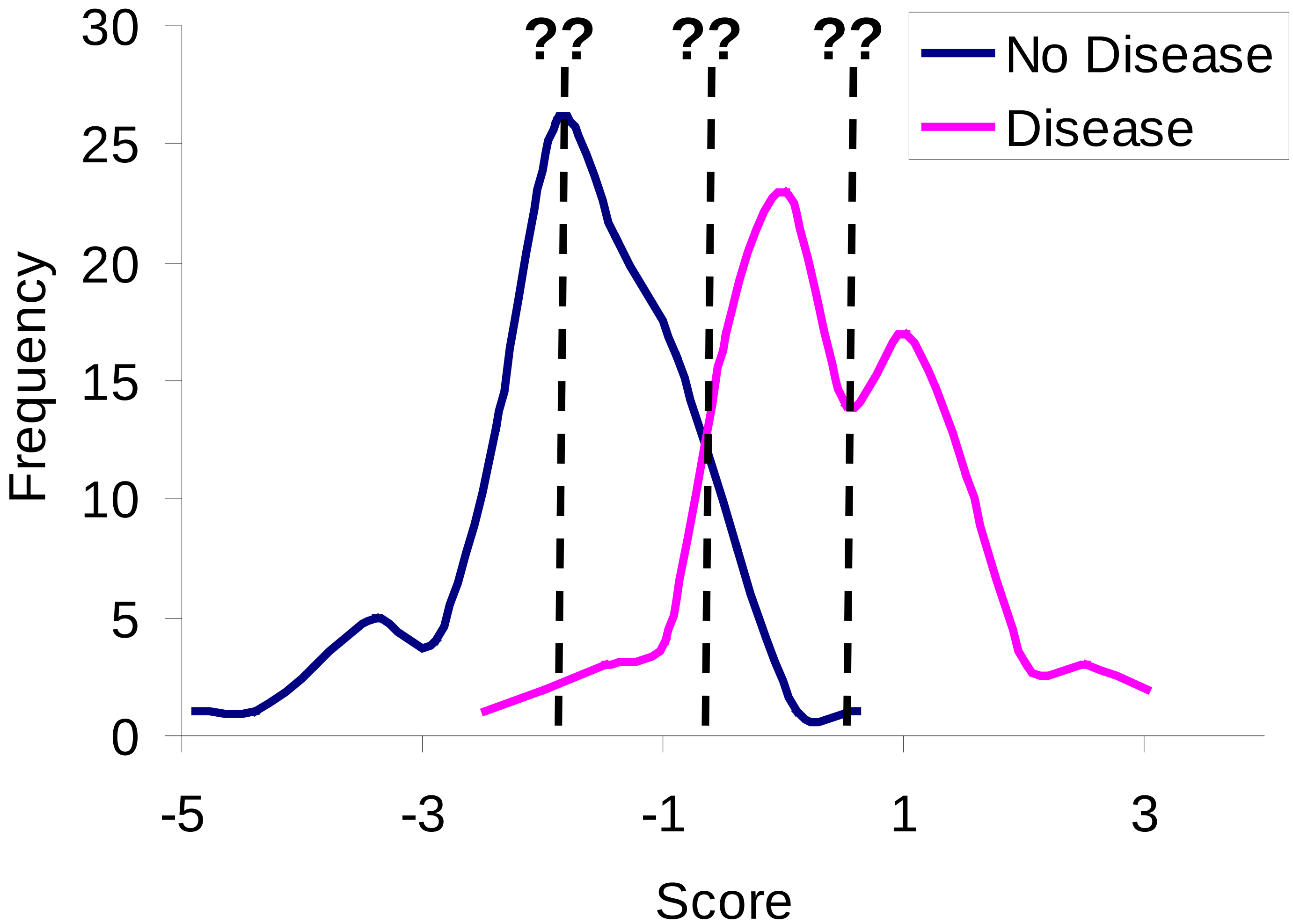
4. An associated synthetic indicator
   Easily interpretable

Its scope goes far beyond the interpretations (indicators) derived from the analysis of the confounding matrix. !

Note on Discrete Classes

- TRADITION … Show contingency table when reporting predictions of model.
- *BUT* … probabilistic models do not provide discrete calculations of the matrix cells!!!
- *IN OTHER WORDS* … Regression does not report the number of individuals predicted positive (*e.g. has a heart attack*) … ***well, not really***
- *INSTEAD* … report probability the output will be certain variable (e.g. 1 or 0)

Visual Perspective



ROC Curves

- Originated from signal detection theory
  - Binary signal corrupted by Guassian noise
  - What is the optimal threshold (i.e. operating point)?

- Dependence on 3 factors
  - Signal Strength
  - Noise Variance
  - Personal tolerance in Hit / False Alarm Rate

ROC Curves

- Receiver operator characteristic
- Summarize & present performance of any binary classification model
- Models ability to distinguish between false & true positives

Use Multiple Contingency Tables

- Sample contingency tables from range of threshold/probability.
- *TRUE POSITIVE RATE* (also called *SENSITIVITY*)

$$\frac{\text{True Positives}}{(\text{True Positives}) + (\text{False Negatives})}$$

- *FALSE POSITIVE RATE* (also called 1 - *SPECIFICITY*)

$$\frac{\text{False Positives}}{(\text{False Positives}) + (\text{True Negatives})}$$

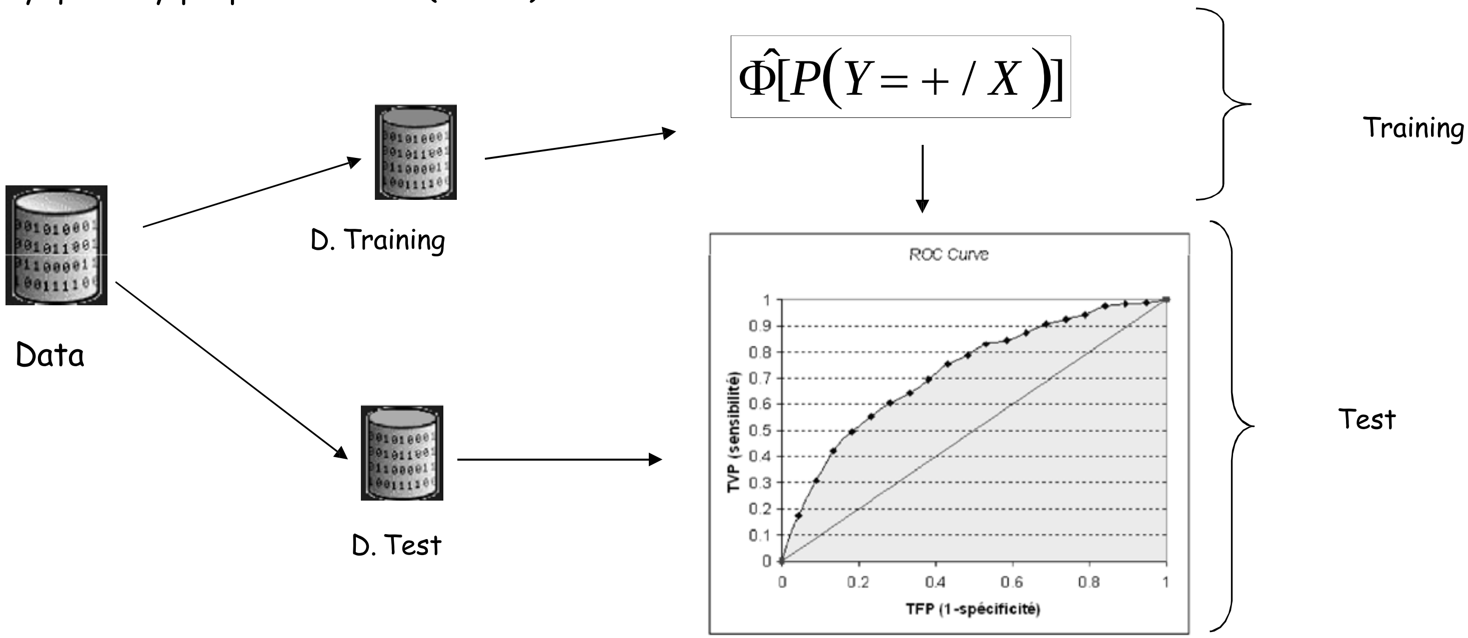
- Plot Sensitivity vs. (1 – Specificity) for sampling and you are done

ROC Curve

Framework for Using the ROC Curve

We are in a 2-class problem
In fact, in all cases where we have the possibility to define the modality (+) of the variable to be predicted

The prediction model provides P(Y=+/X)
Or any quantity proportional to P(Y=+/X) that will allow the observations to be classified



ROC Curve

ROC Curve Principle

Confusion Matrix

	^positive	^negative
positve	TP	FN
Negative	FP	TN

TPR = Rappel = Sensitivity = TP/Positives

FPR = 1 – Specificity = FP/Negatives

ROC Curve Principle
P(Y=+/X) >= P(Y=-/X) Equivalent to an assignment rule P(Y=+/X) >= 0.5 (threshold = 0.5)
This assignment rule provides an MC1 confusion matrix, and thus 2 TPR1 and FPR1 indicators

The idea of the ROC curve is to vary the "threshold" from 1 to 0 and, for each case, calculate the TPR and the FPR that are plotted in a graph: on the x-axis the FPR, on the y-axis the TPR.

ROC Curve

ROC Curve Construction (1/2)

Sort data according to a descending score

Individu	Score (+)	Classe
1	1	+
2	0.95	+
3	0.9	+
4	0.85	-
5	0.8	+
6	0.75	-
7	0.7	-
8	0.65	+
9	0.6	-
10	0.55	-
11	0.5	-
12	0.45	+
13	0.4	-
14	0.35	-
15	0.3	-
16	0.25	-
17	0.2	-
18	0.15	-
19	0.1	-
20	0.05	-

Positifs = 6
Négatifs = 14

Seuil = 1

	^positif	^négatif	Total
positf	1	5	6
négatif	0	14	14
Total	1	19	20

TVP = 1/6 = 0.2 ; TFP = 0/14 = 0

Seuil = 0.95

	^positif	^négatif	Total
positf	2	4	6
négatif	0	14	14
Total	2	18	20

TVP = 2/6 = 0.33 ; TFP = 0/14 = 0

Seuil = 0.9

	^positif	^négatif	Total
positf	3	3	6
négatif	0	14	14
Total	3	17	20

TVP = 3/6 = 0.5 ; TFP = 0/14 = 0

Seuil = 0.85

	^positif	^négatif	Total
positf	3	3	6
négatif	1	13	14
Total	4	16	20

TVP = 3/6 = 0.5 ; TFP = 1/14 = 0.07

Seuil = 0

	^positif	^négatif	Total
positf	6	0	6
négatif	14	0	14
Total	20	0	20

TVP = 6/6 = 1 ; TFP = 14/14 = 1

ROC Curve

Construction de la courbe ROC (2/2)

Mettre en relation
FPR (x-axis) and TPR (y-axis)

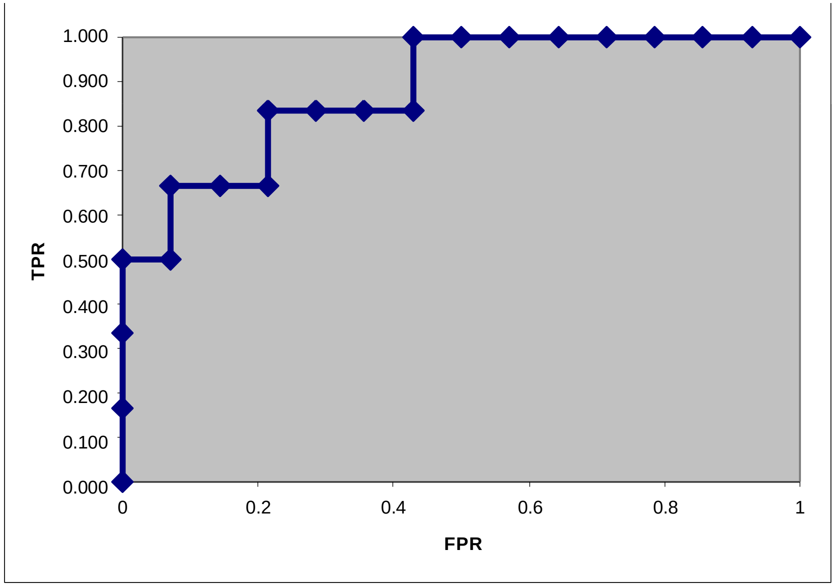
Individu	Score (+)	Classe	TFP	TVP
			0	0.000
1	1	+	0.000	0.167
2	0.95	+	0.000	0.333
3	0.9	+	0.000	0.500
4	0.85	-	0.071	0.500
5	0.8	+	0.071	0.667
6	0.75	-	0.143	0.667
7	0.7	-	0.214	0.667
8	0.65	+	0.214	0.833
9	0.6	-	0.286	0.833
10	0.55	-	0.357	0.833
11	0.5	-	0.429	0.833
12	0.45	+	0.429	1.000
13	0.4	-	0.500	1.000
14	0.35	-	0.571	1.000
15	0.3	-	0.643	1.000
16	0.25	-	0.714	1.000
17	0.2	-	0.786	1.000
18	0.15	-	0.857	1.000
19	0.1	-	0.929	1.000
20	0.05	-	1.000	1.000

Practical Calculation

FPR (i) = Number of negatives among the first "i's" / (total number of negatives)

TPR (i) = Number of positives among the first i's / (total number of positives)

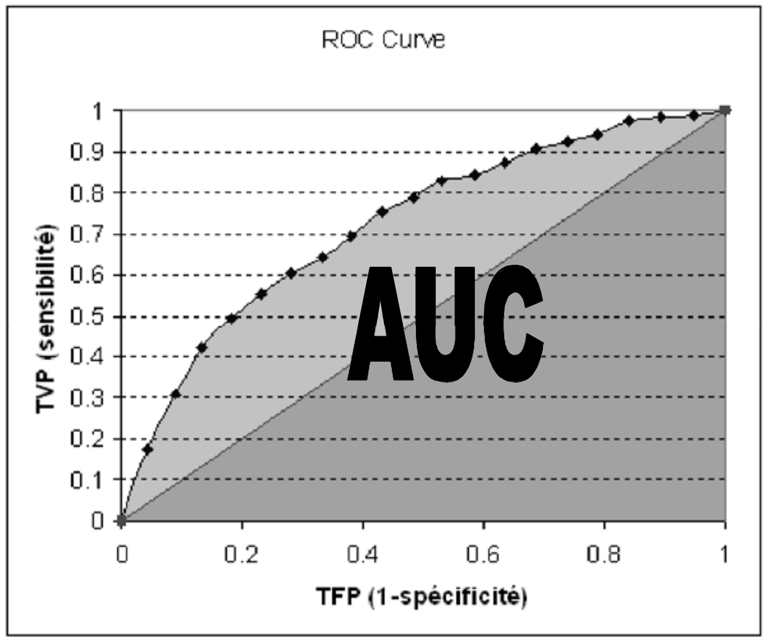
ROC Curve



ROC Curve

Interpretation: AUC, the area under the curve

AUC indicates the probability that the SCORE function will place a positive before a negative (best-case AUC = 1) **!**



If SCORE randomly classifies individuals (i.e., the prediction model is useless), AUC = 0.5

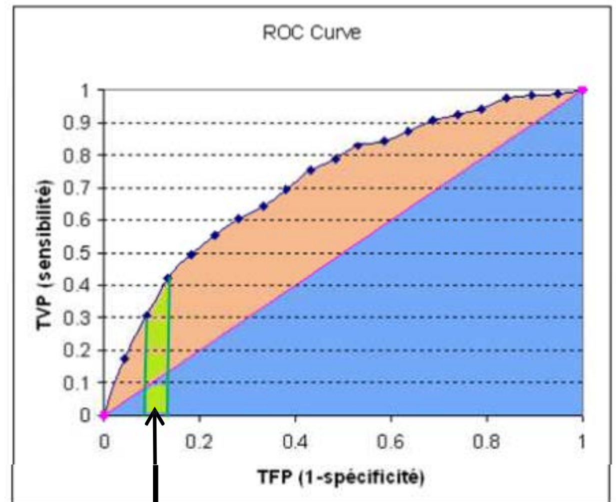
Symbolized by the main diagonal in the graph

ROC Curve

AUC – Practical Calculation – Integration with the Trapezoid Method

Derived directly from the definition: surface = integral

Individu	Score (+)	Classe	TFP	TVP	Largeur	Hauteur	Surface
			0	0.000			
1	1	+	0.000	0.167	0.000	0.083	0.000
2	0.95	+	0.000	0.333	0.000	0.250	0.000
3	0.9	+	0.000	0.500	0.000	0.417	0.000
4	0.85	-	0.071	0.500	0.071	0.500	0.036
5	0.8	+	0.071	0.667	0.000	0.583	0.000
6	0.75	-	0.143	0.667	0.071	0.667	0.048
7	0.7	-	0.214	0.667	0.071	0.667	0.048
8	0.65	+	0.214	0.833	0.000	0.750	0.000
9	0.6	-	0.286	0.833	0.071	0.833	0.060
10	0.55	-	0.357	0.833	0.071	0.833	0.060
11	0.5	-	0.429	0.833	0.071	0.833	0.060
12	0.45	+	0.429	1.000	0.000	0.917	0.000
13	0.4	-	0.500	1.000	0.071	1.000	0.071
14	0.35	-	0.571	1.000	0.071	1.000	0.071
15	0.3	-	0.643	1.000	0.071	1.000	0.071
16	0.25	-	0.714	1.000	0.071	1.000	0.071
17	0.2	-	0.786	1.000	0.071	1.000	0.071
18	0.15	-	0.857	1.000	0.071	1.000	0.071
19	0.1	-	0.929	1.000	0.071	1.000	0.071
20	0.05	-	1.000	1.000	0.071	1.000	0.071
						AUC	0.881



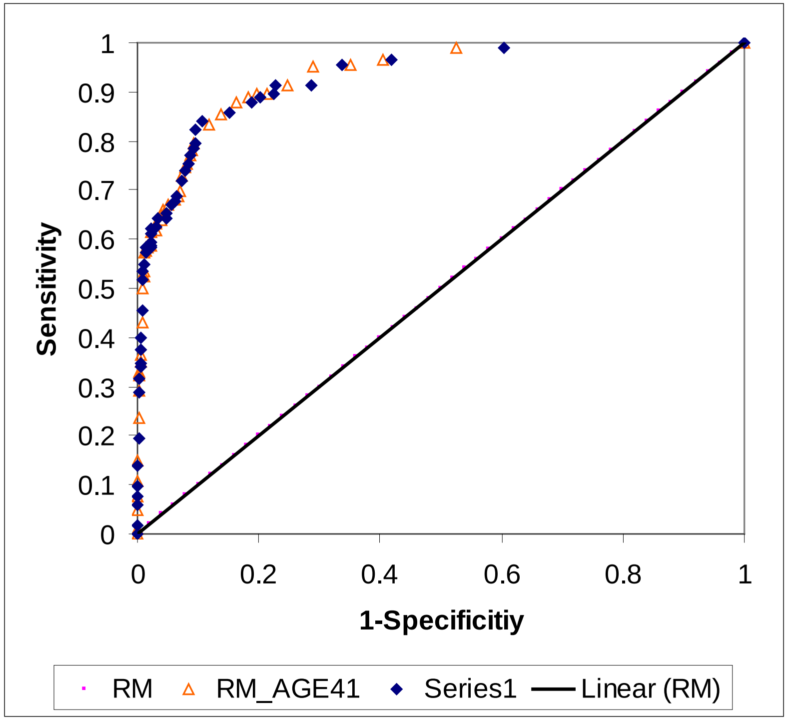
$$s_i = (FPR_i - FPR_{i-1}) \times \frac{TPR_i + TPR_{i-1}}{2}$$

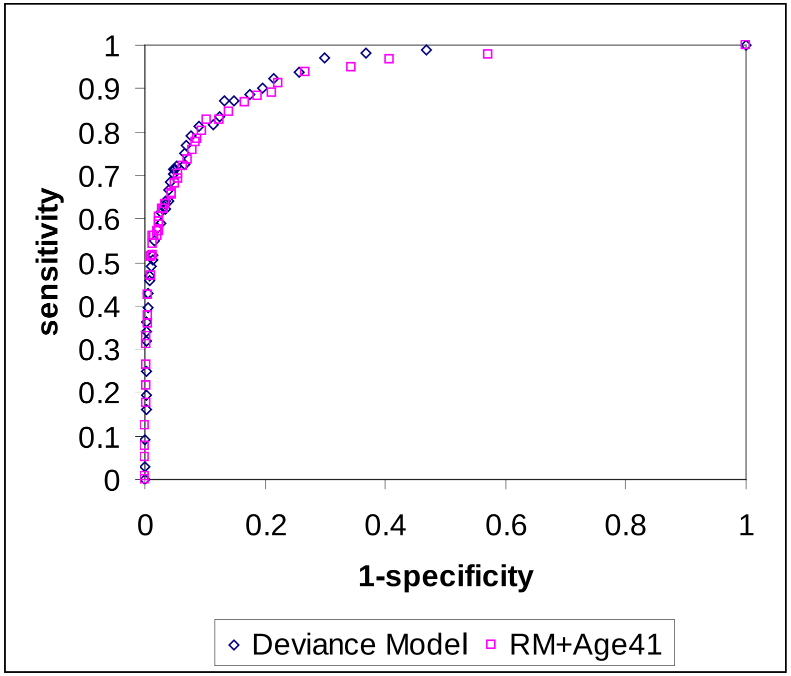
Surface of a trapezoid

$$AUC = \sum_i s_i$$

AUC = SUM (Trapezoid Area)

Sidebar: Use More Samples



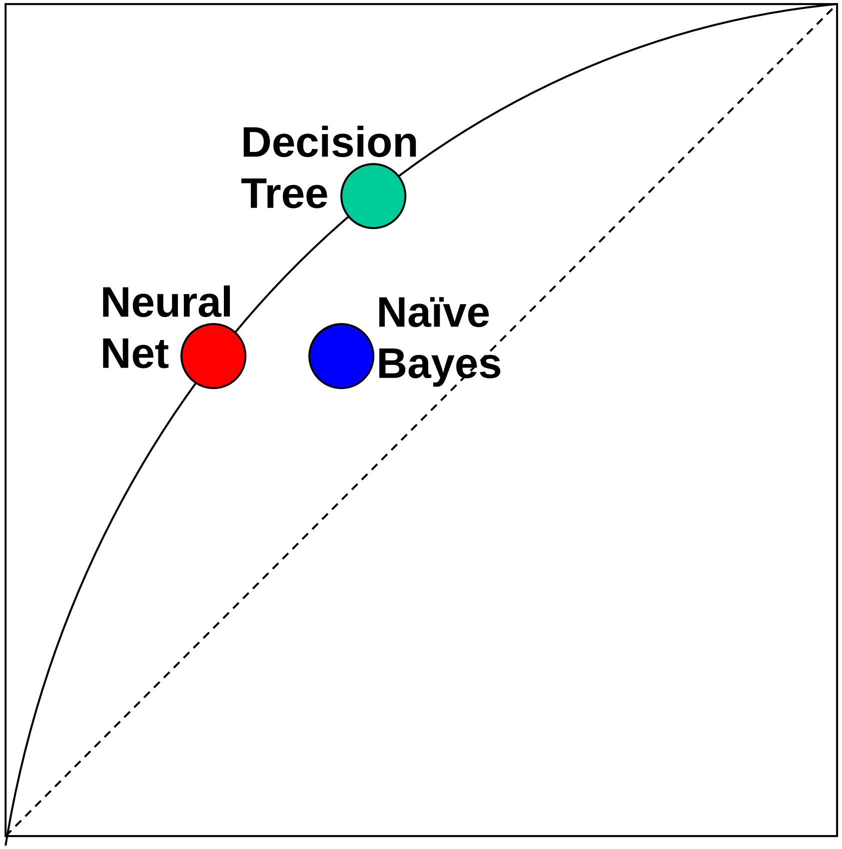


(These are plots from a much larger dataset
– See Malin 2005)

Theory: Model Optimality

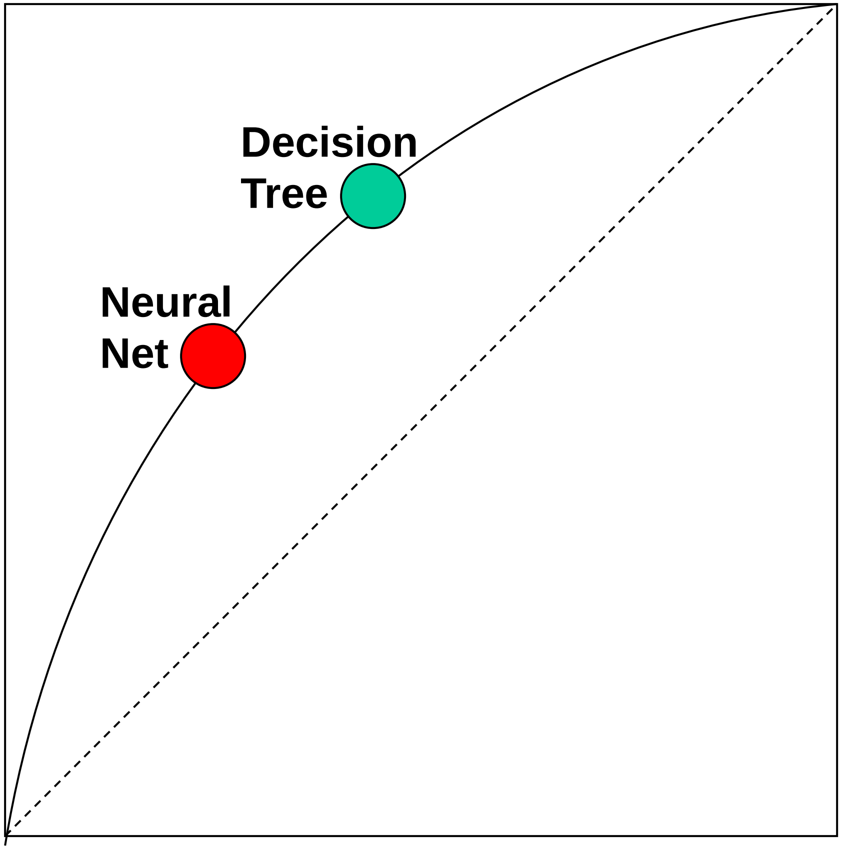
- Classifiers on convex hull are always "optimal"
  - e.g. Net & Tree

- Classifiers below convex hull are always "suboptimal"
  - e.g. Naïve Bayes



Building Better Classifiers

- Classifiers on convex hull can be combined to form a *strictly dominant* hybrid classifier
  - ordered sequence of classifiers
  - can be converted into “ranker”



Some Statistical Insight

- Curve Area:
  - Take random healthy patient → score of *X*
  - Take random heart attack patient → score of *Y*
  - Area estimate of *P* [*Y* > *X*]

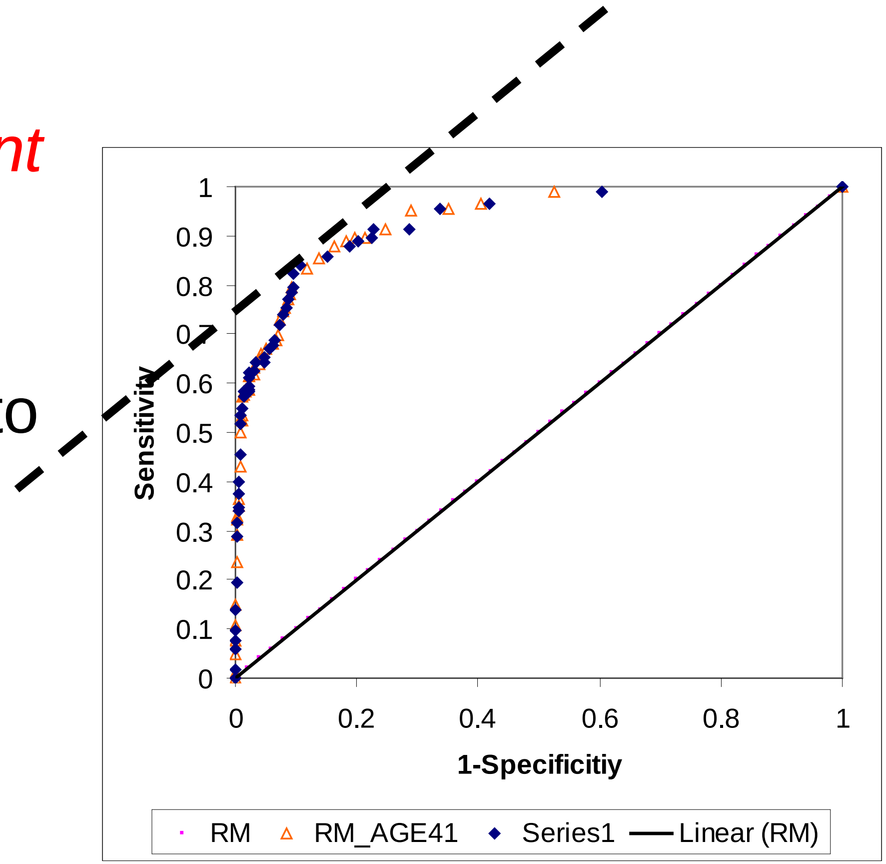
- Slope of curve is equal to likelihood:

$$\frac{P\,(score \mid \mathbf{Signal})}{P\,(score \mid \mathbf{Noise})}$$

- ROC graph captures all information in conting. table
  - *False negative & true negative rates are complements of true positive & false positive rates, resp.*

Can Always Quantify Best Operating Point

- When misclassification costs are equal, *best operating point* is …
- 45° tangent to curve closest to (0,1) coord.
- Verify this mathematically (*economic interpretation*)
- Why?



Quick Question

- Are ROC curves always appropriate?

- Subjective operating points?

- Must weight the tradeoffs between false positives and false negatives
  - *ROC curve plot is independent of the class distribution or error costs*

- This leads into utility theory (not touching this today)

Equality of Opportunity
(Hardt et al. method, 2016)

Principle

- This method, necessarily true as a post-processing method, requires only knowledge about membership in a protected class as well as the predictions of a method.
- It then imposes a fairness model of equality of opportunity, meaning that for each group, of the individuals who should have a positive label (ground truth), their probability of having a positive label is the same.

How it works?

Definition 2.1 (Equalized odds)**.** We say that a predictor $\widehat{Y}$ satisfies *equalized odds* with respect to protected attribute A and outcome Y, if $\widehat{Y}$ and A are independent conditional on Y.

$$\Pr\left\{\widehat{Y} = 1 \mid A = 0, Y = y\right\} = \Pr\left\{\widehat{Y} = 1 \mid A = 1, Y = y\right\}, \quad y \in \{0, 1\} \tag{2.1}$$

Definition 2.2 (Equal opportunity)**.** We say that a binary predictor $\widehat{Y}$ satisfies *equal opportunity* with respect to A and Y if $\Pr\{\widehat{Y} = 1 \mid A = 0, Y = 1\} = \Pr\{\widehat{Y} = 1 \mid A = 1, Y = 1\}$.

Equal opportunity is a weaker, though still interesting, notion of non-discrimination, and can thus allows for better utility.

Definition 3.1 (Derived predictor)**.** A predictor $\widetilde{Y}$ is *derived from a random variable R and the protected attribute A* if it is a possibly randomized function of the random variables (R, A) alone. In particular, $\widetilde{Y}$ is independent of X conditional on (R, A).

In designing a derived predictor from binary $\widehat{Y}$ and A we can only set four parameters: the conditional probabilities $p_{ya} = \Pr\{\widetilde{Y} = 1 \mid \widehat{Y} = a, A = a\}$. These four parameters, $p = (p_{00}, p_{01}, p_{10}, p_{11})$, together specify the derived predictor $\widetilde{Y}_p$. To check whether $\widetilde{Y}_p$ satisfies equalized odds we need to verify the two equalities specified by (2.1), for both values of y. To this end, we denote

$$\gamma_a(\widetilde{Y}) \stackrel{\text{def}}{=} \left(\Pr\{\widetilde{Y} = 1 \mid A = a, Y = 0\}, \Pr\{\widetilde{Y} = 1 \mid A = a, Y = 1\}\right). \tag{3.1}$$

How it works?

The components of $\gamma_a(\widetilde{Y})$ are the *false positive rate* and the *true positive rate* within the demographic $A = a$. Following (2.1), $\widetilde{Y}$ satisfies equalized odds iff $\gamma_0(\widetilde{Y}) = \gamma_1(\widetilde{Y})$. But $\gamma_a(\widetilde{Y}_p)$ is just a linear function of p, with coefficients determined by the joint distribution of $(Y, \widehat{Y}, A)$. Since the expected loss $\mathbb{E}\ell(\widetilde{Y}_p, Y)$ is also linear in p, we have that the optimal derived predictor can be obtained as a solution to the following linear program with four variables and two equality constraints:

$$\min_p \quad \mathbb{E}\ell(\widetilde{Y}_p, Y) \tag{3.2}$$

$$\text{s.t.} \quad \gamma_0(\widetilde{Y}_p) = \gamma_1(\widetilde{Y}_p) \tag{3.3}$$

$$\forall_{y,a} 0 \leqslant p_{ya} \leqslant 1 \tag{3.4}$$

To better understand this linear program, let us understand the range of values $\gamma_a(\widetilde{Y}_p)$ can take:

Claim 3.2. $\{\gamma_a(\widetilde{Y}_p) \mid 0 \leqslant p \leqslant 1\} = P_a(\widehat{Y}) \stackrel{\text{def}}{=} \text{convhull}\left\{(0,0), \gamma_a(\widehat{Y})\gamma_a(1-\widehat{Y}), (1,1)\right\}$

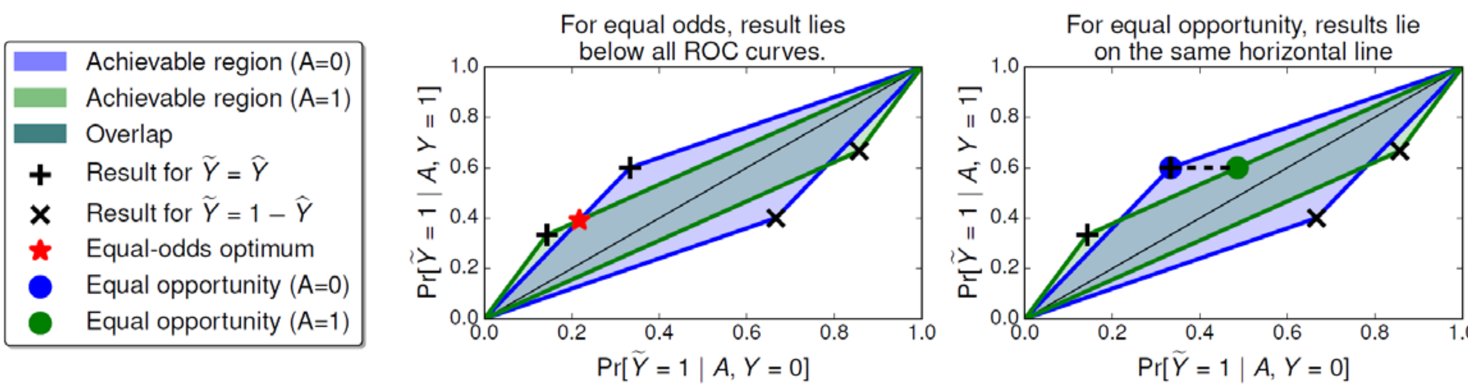


Figure 1: Finding the optimal equalized odds predictor (left), and equal opportunity predictor (right).

ROC: Rejection Option-based Classification

(Kamiran et al. method, 2012)

Rejection Option-based Classification

How to achieve independence?:

- algorithm postprocessing
- data preprocessing (representation/feature learning)

- e.g., information theory approach

$$Z = \phi(X, A), \quad \text{with} \quad \max I(X; Z) \text{ and } \min I(A; Z)$$

then use $\hat{Y} = g(Z, A)$ rather than $\hat{Y} = g(X, A)$.

Problems:

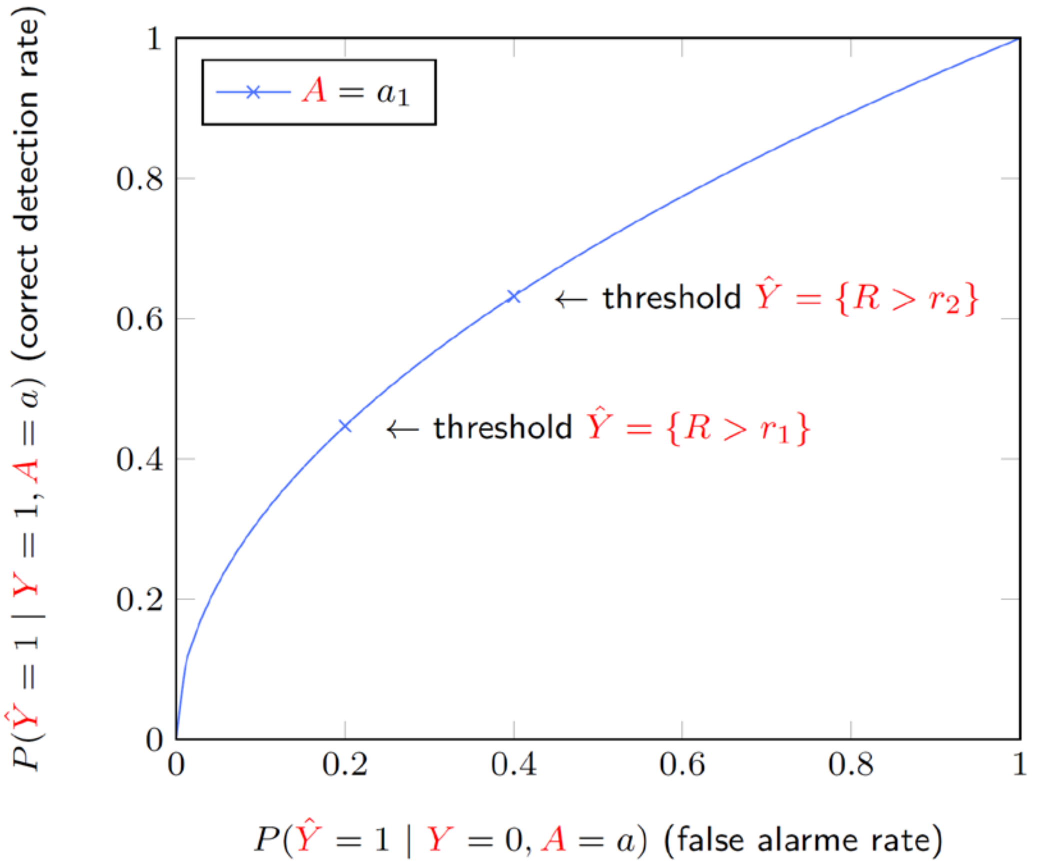
- ignores possible correlations between Y and A
- **Example**: since more male SWE than female SWE, even with Z independent of A, Y relates highly to A.

  $\Rightarrow$ Perfect predictor $\hat{Y} = Y$ unreachable.

- creates random assignments in one group to avoid discrimination (if for all males, $Y = 0$ (no male candidate suitable), solution is to pick males randomly ($\hat{Y} = 1$) to avoid discrimination!)

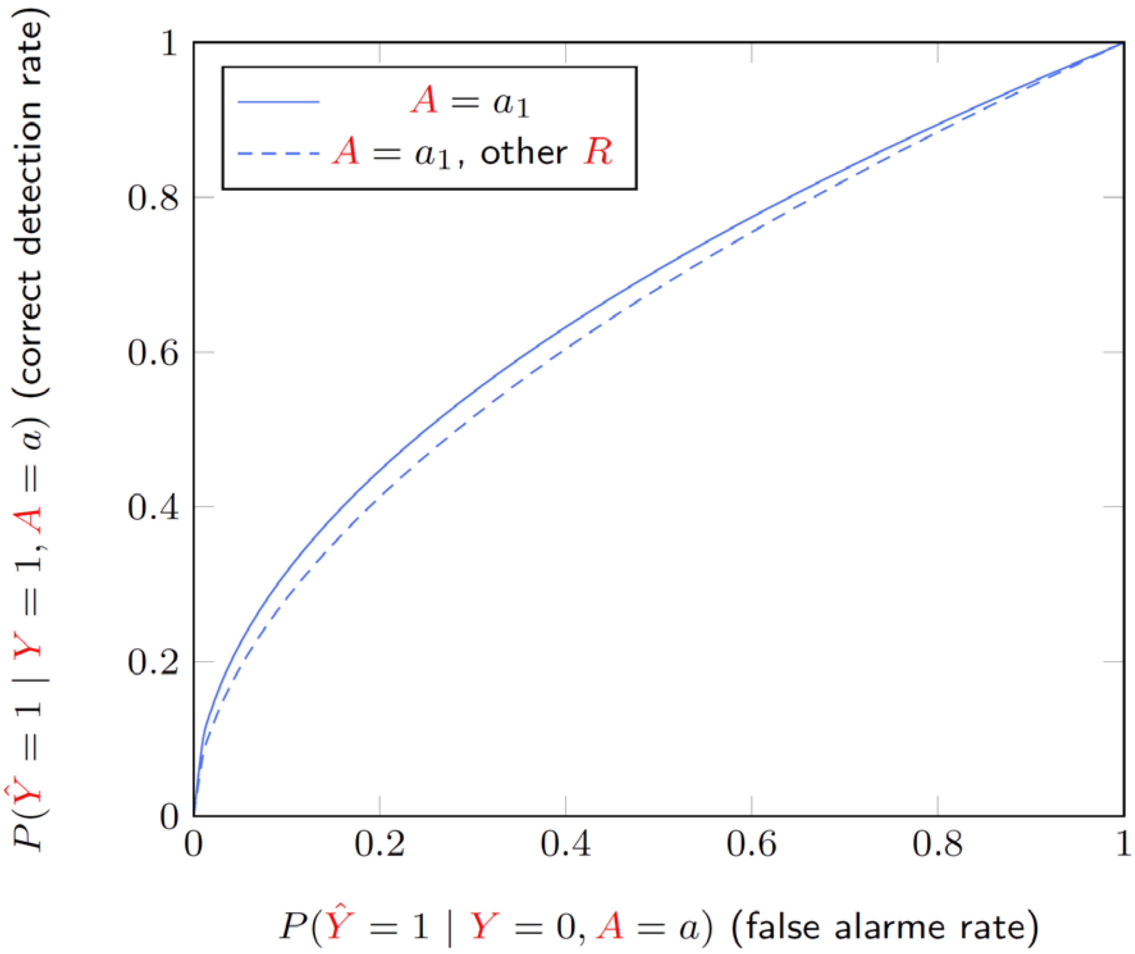
Rejection Option-based Classification

Postprocessing: ROC curve (receiver operator curve)



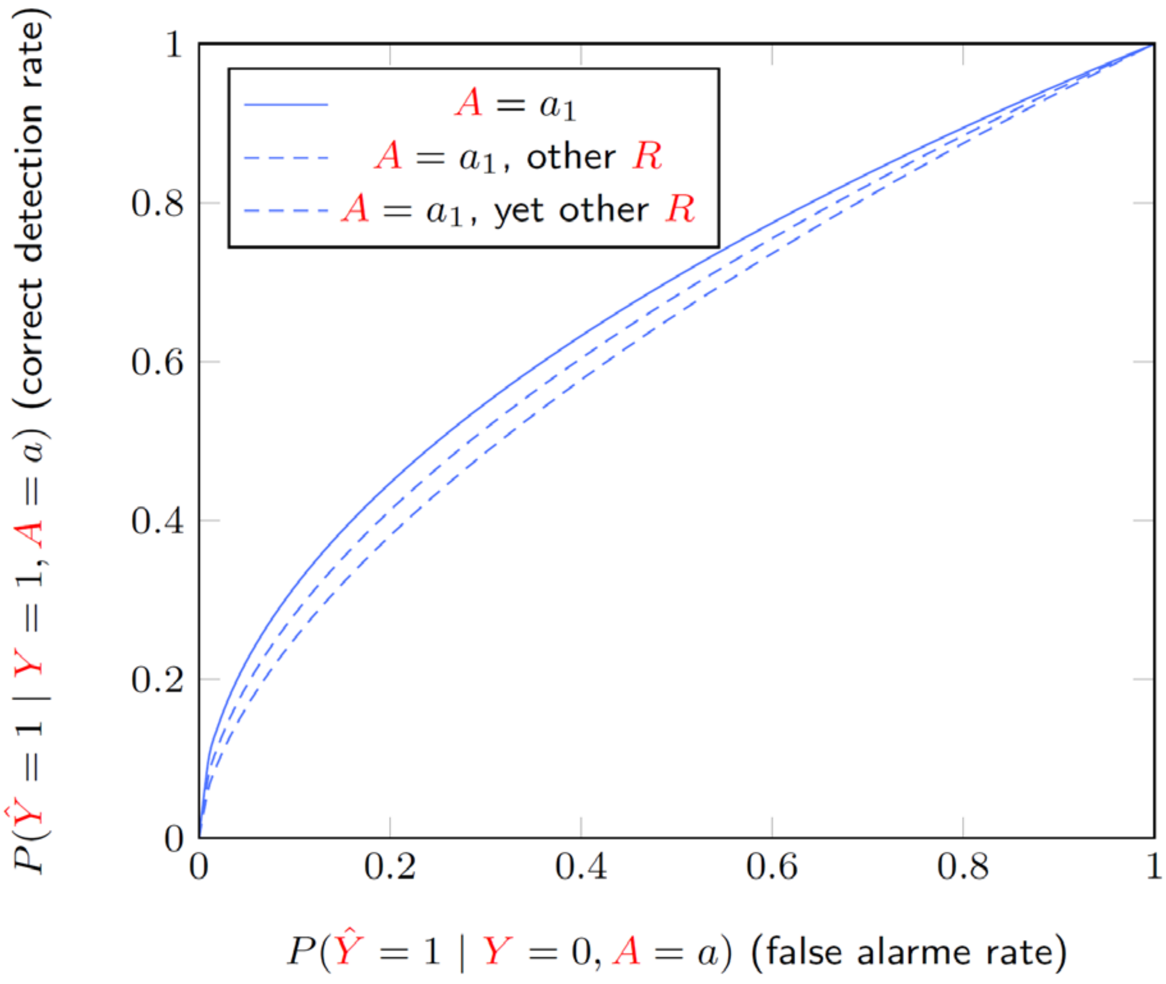
Rejection Option-based Classification

Postprocessing: ROC curve (receiver operator curve)



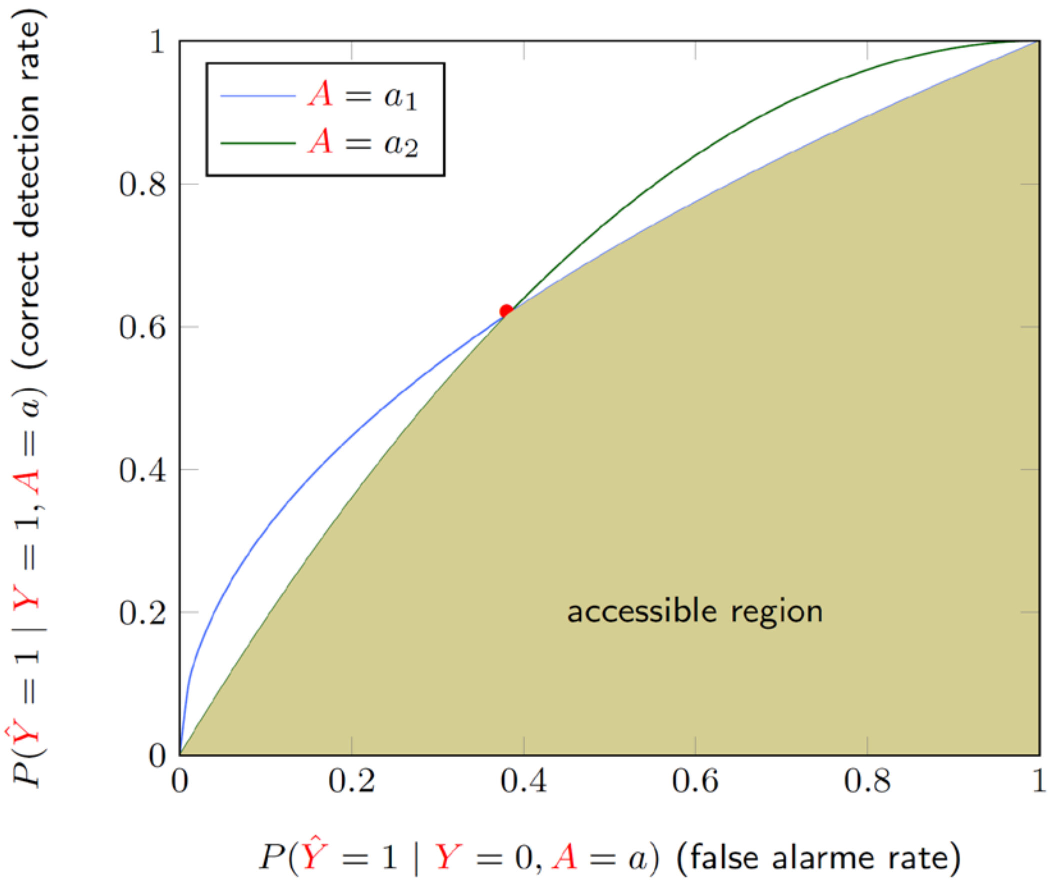
Rejection Option-based Classification

Postprocessing: ROC curve (receiver operator curve)



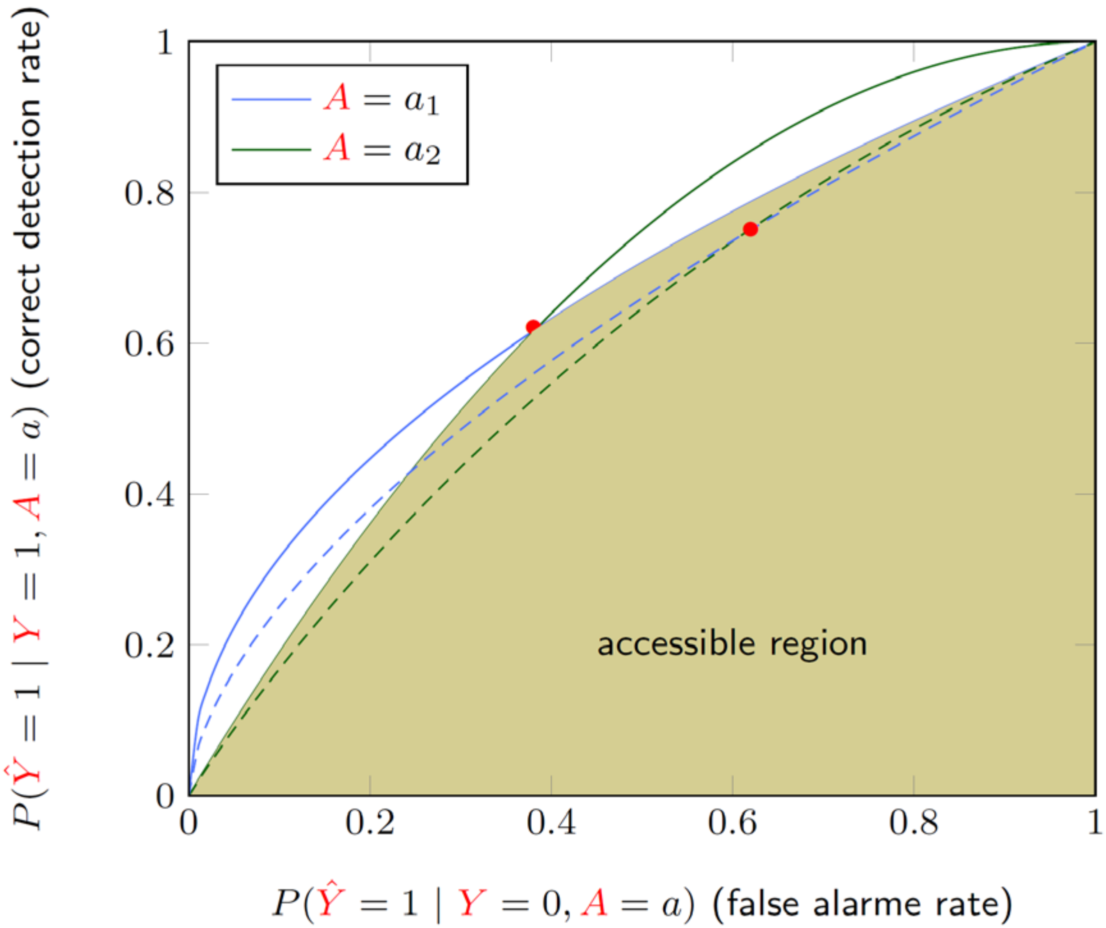
Rejection Option-based Classification

Postprocessing: ROC curve (receiver operator curve)



Rejection Option-based Classification

Postprocessing: ROC curve (receiver operator curve)



Rejection Option-based Classification

Postprocessing: ROC curve (receiver operator curve)

- choose decision threshold r such that (recall $R = r(X, A)$)

$$\mathbb{P}(r(X, A = a) > r \mid Y = y, A = a) = \mathbb{P}(r(X, A = b) > r \mid Y = y, A = b)$$

- $\Rightarrow$ crossing point of two conditional decision rules in ROC curve.
- **Careful**! Requires score reparametrization **or** different thresholds $R > r_a \mid A = a$.

Reparametrization: assume two intersecting ROC curves

$$f_a(r) = (x_a(r), y_a(r)) = (\mathrm{FAR}_a(r), \mathrm{CDR}_a(r)) \quad \text{for} \quad r = r(X, A = a)$$
$$f_b(r) = (x_b(r), y_b(r)) = (\mathrm{FAR}_b(r), \mathrm{CDR}_b(r)) \quad \text{for} \quad r = r(X, A = b)$$

(in particular, $f_\cdot(0) = 0,\ f_\cdot(1) = 1$)

- intersection defined as

$$f_a(r_1) = f_b(r_2) \quad \text{for some} \quad r_1, r_2.$$

- **Unlikely that $r_1 = r_2$!** Depends on parametrization.
- Reparametrization: When intersecting couple (r_1, r_2) found, **scale parameters** $r \to r' = h(r)$ **so that** $f_a \to f_a'$, $f_b \to f_b'$ and

$$f_a'(r) = f_a(h(r_1)) = f_a(r_1) = f_b(r_2) = f_b(h_b(r)) = f_b'(r).$$