Part II: A broader view

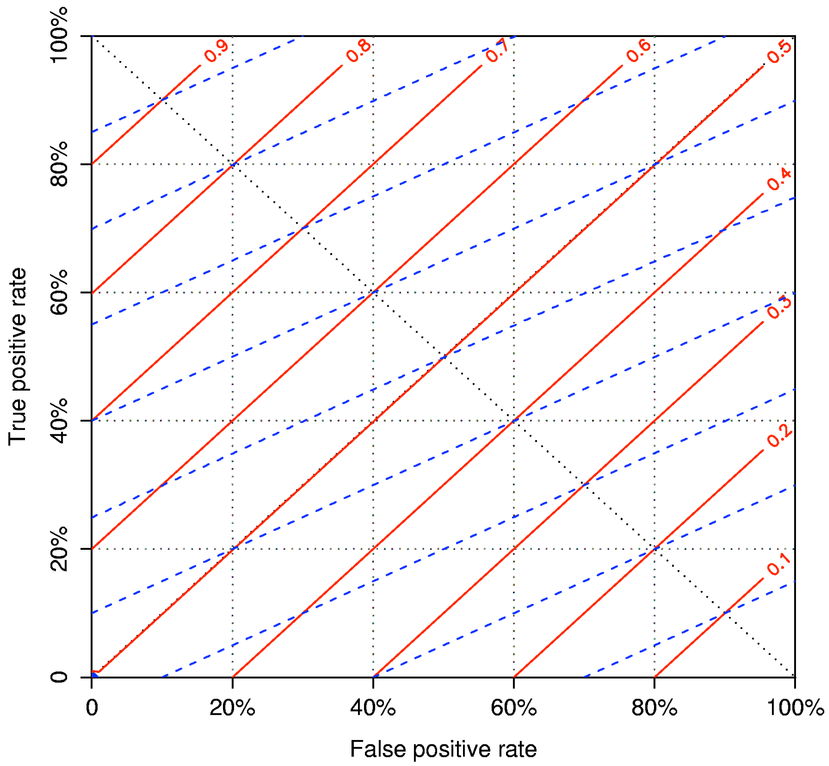
- Understanding ML metrics:
  - isometrics, basic types of linear isometric plots
  - linear metrics and equivalences between them
  - skew-sensitivity
  - non-linear metrics

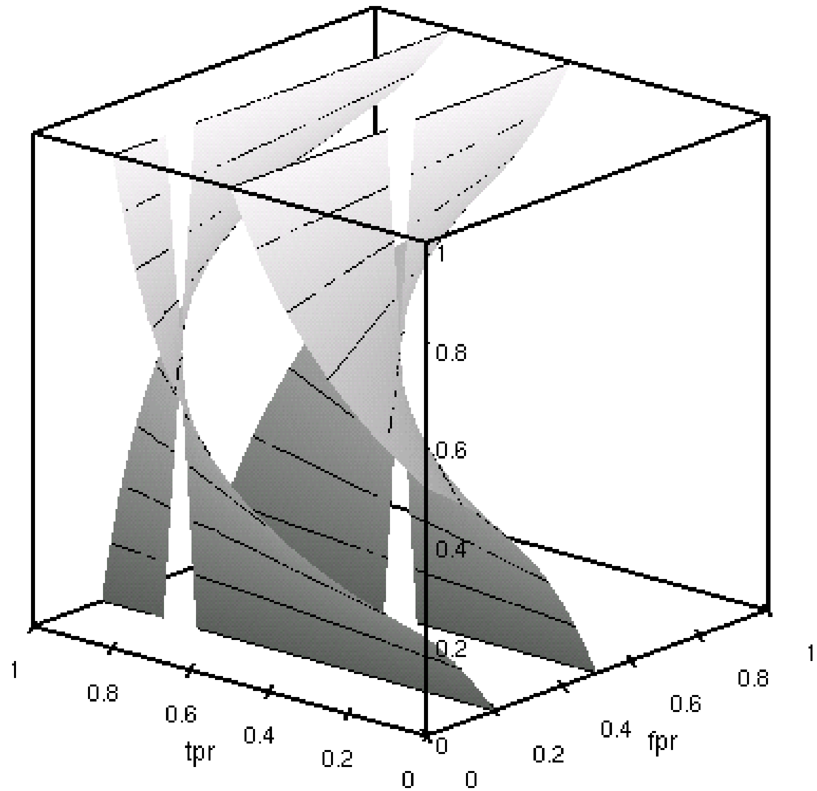
- Model manipulation:
  - obtaining new models without re-training
  - ordering decision tree branches
  - repairing concavities by locally adjusting rankings

Understanding ML metrics

- We are referring here to metrics (or heuristics) that are used to rank (*fpr*,*tpr*) points
  - i.e., classifiers or parts of classifiers
    - NB. different sense of ranking than before!
- Metrics are equivalent if their rankings are the same
  - absolute value of metric not important
- This can be visualised very clearly by means of ROC isometrics
  - additional benefit of studying skew-sensitivity
  - see (Flach, 2003) and (Fürnkranz & Flach, 2003)

Iso-accuracy lines revisited





- In 2D ROC space
  - c = 1, c = 1/2
- In 3D ROC space
  - acc = 0.5, acc = 0.8

Isometrics and skew ratio

- Accuracy is weighted average of true positive/negative rates:

$$acc = pos \cdot tpr + neg \cdot (1 - fpr) = \frac{tpr + c \cdot (1 - fpr)}{c + 1}$$

- Skew ratio indicates relative importance of negatives over positives
  - without costs: $c = neg/pos$

- Isometric plots show contour lines in 2D ROC space for a given metric with skew ratio as parameter

Skew-sensitivity

- Strongly skew-insensitive metric is independent of skew ratio
  - isometric surfaces in 3D ROC space are vertical
  - can be obtained for any metric by fixing c
- Weakly skew-insensitive metric has the same isometric landscape for different values of c
  - any collection of ROC points is ranked the same way, regardless of c
- Line of skew-indifference: points where the metric is independent of c
  - for accuracy, this is the line tpr+fpr–1=0

Types of isometric plots

a) Parallel linear isometrics
- accuracy, weighted relative accuracy (WRAcc)

b) Rotating linear isometrics
- precision, lift, F-measure

c) Non-linear isometrics
- decision tree splitting criteria

Symmetries

- Inverting predictions of classifier
  - ROC space: point-mirroring through (0.5, 0.5)
  - contingency table: swapping columns
- Inverting test labels
  - ROC space: mirroring along ascending diagonal
  - contingency table: swapping rows
    - affects skew ratio (c becomes 1/c), so a test for skew-insensitivity
- Inverting both predictions and test labels
  - ROC space: mirroring along descending diagonal
  - contingency table: swapping rows and columns

Weighted relative accuracy

- Original definition:

$$wracc / 4 = P(x) \cdot [P(+ \mid x) - P(+)] = P(x,+) - P(x) \cdot P(+)$$

- In ROC notation: $wracc = \dfrac{4c}{(c+1)^2}(tpr - fpr)$

- Weakly skew-insensitive: isometrics are parallel to diagonal
  - strongly skew-insensitive version: *tpr–fpr*

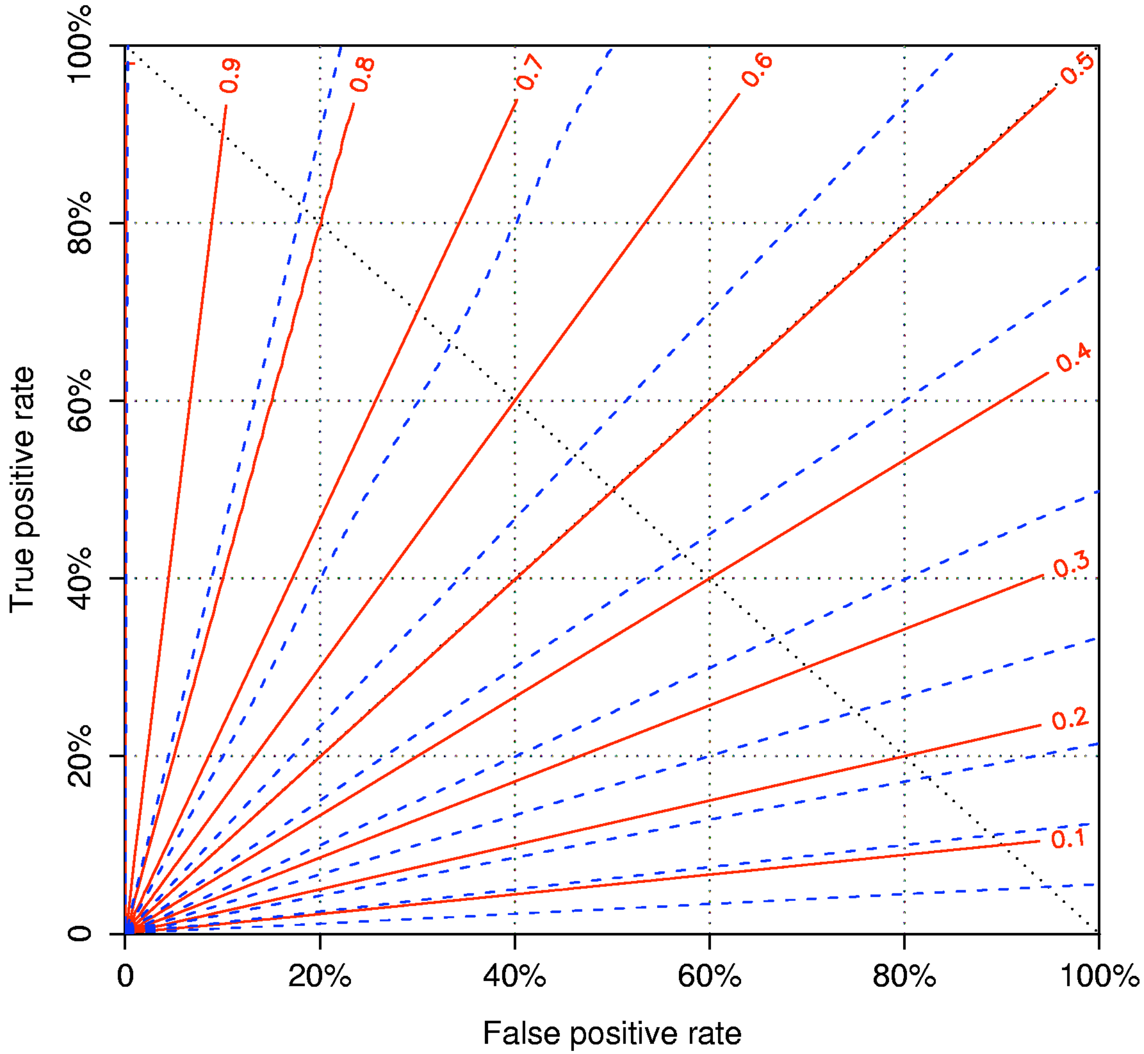
Precision or confidence

- Precision is defined as

$$prec = \frac{pos \cdot tpr}{pos \cdot tpr + neg \cdot fpr} = \frac{tpr}{tpr + c \cdot fpr}$$

- Weakly skew-insensitive, rotating isometrics
  - on $tpr = fpr$ diagonal, $prec = pos$
- Two variants with fixed value on diagonal
  - relative precision: $relprec = prec - pos$
  - lift: $lift = prec / pos$

Precision isometrics



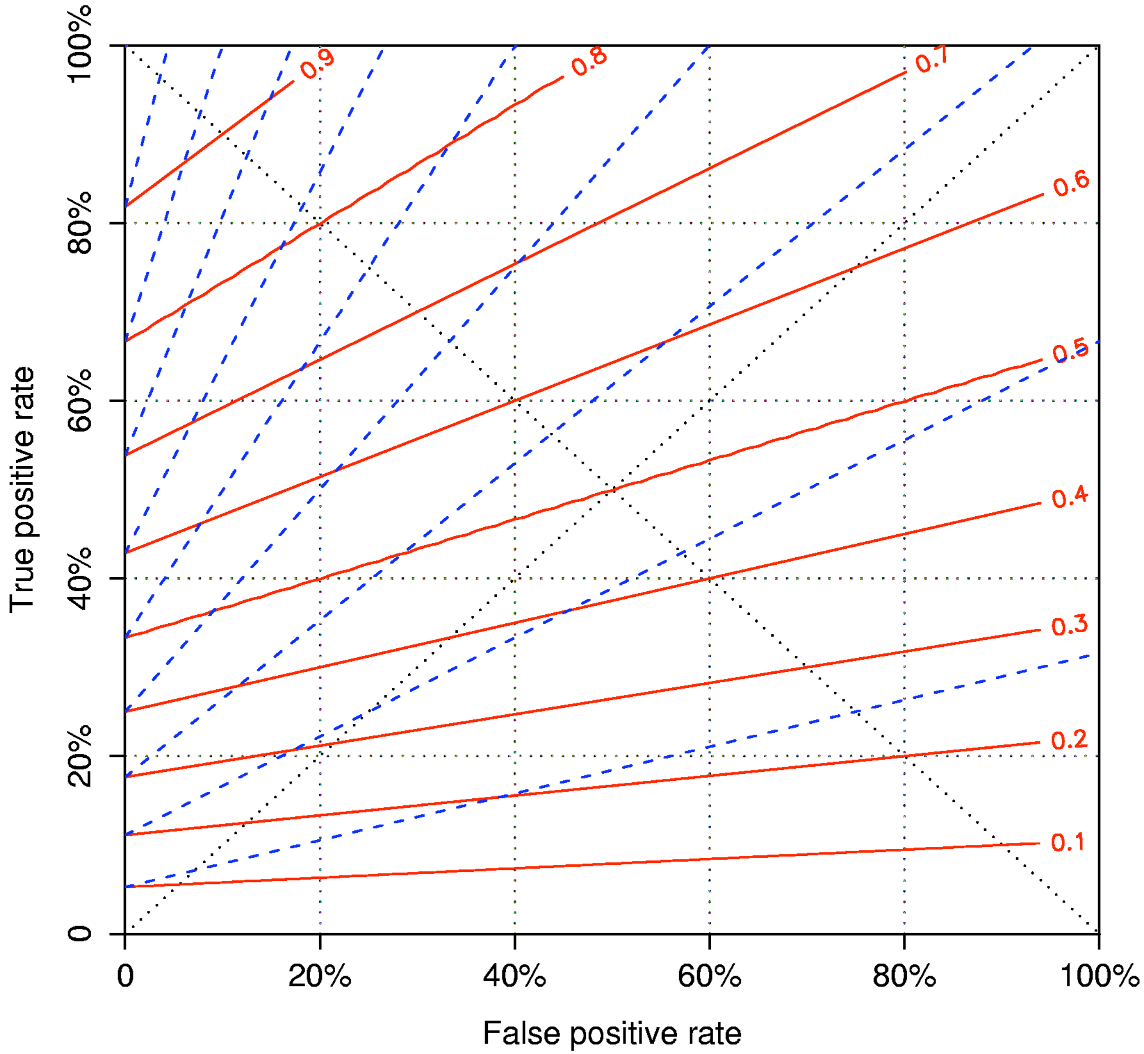
c = 1,

c = 1/2

F-measure

- F-measure is harmonic average of precision and recall
  - alternatively, F-measure = precision (recall) with FP (FN) replaced with (FP+FN)/2
- In ROC notation: $F = \dfrac{2tpr}{tpr + c \cdot fpr + 1}$
- Equivalent but simpler: $G = \dfrac{tpr}{c \cdot fpr + 1}$
- *fpr*=0 is line of skew-indifference

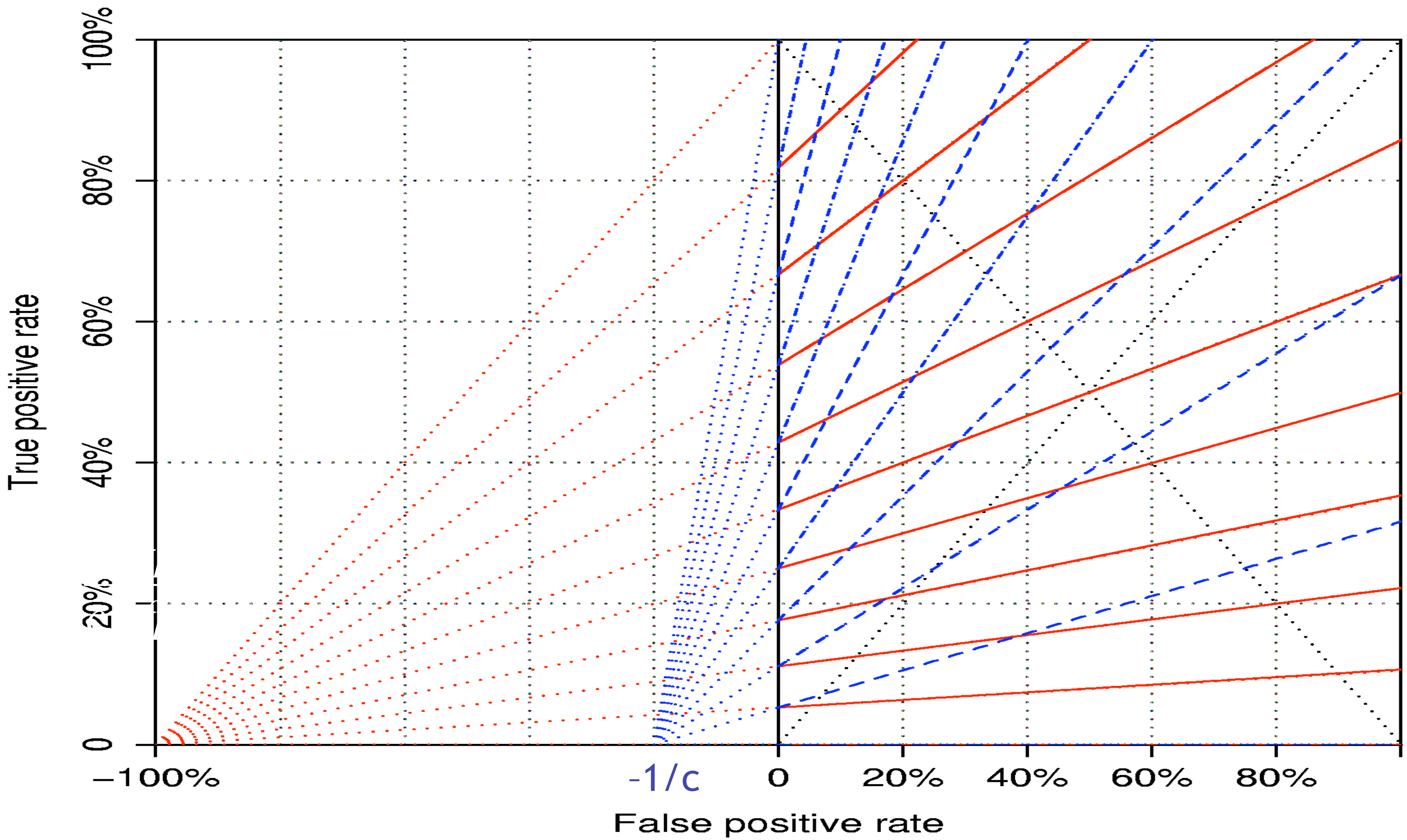
F-measure isometrics



c = 1,

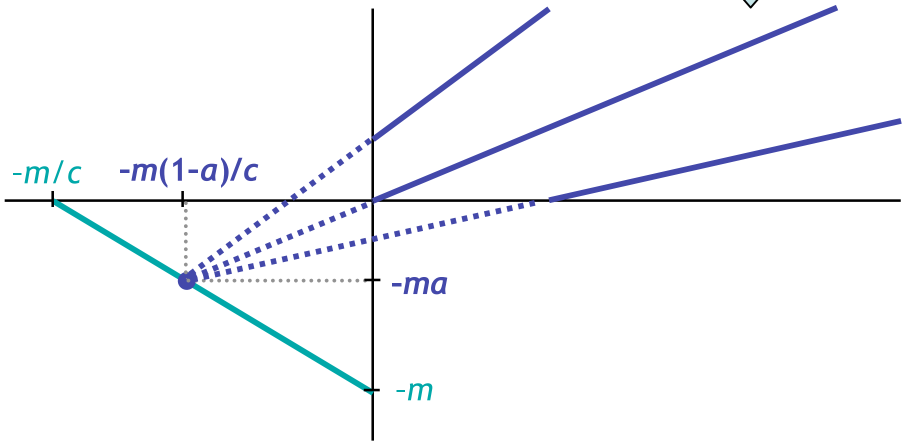
c = 5

F-measure isometrics



Generalised linear isometrics

- Laplace correction and *m*-estimate are other examples which translate the rotation point
- General form: $$\frac{tpr + ma}{tpr + c \cdot fpr + m}$$
  - m=0: precision
  - $m \rightarrow \infty$: parallel isometrics with slope $\frac{ac}{1-a}$
    - e.g. accuracy: a=1/2



Linear metrics: summary

Metric	Formula	Skew-insensitive version	Isometric slope
Accuracy	$\frac{tpr + c(1-fpr)}{c+1}$	$\frac{(tpr + 1 - fpr)}{2}$	c
WRAcc*	$\frac{4c}{(c+1)^2}(tpr - fpr)$	$tpr - fpr$	1
Precision*	$\frac{tpr}{tpr + c \cdot fpr}$	$\frac{tpr}{tpr + fpr}$	
Lift*	$\frac{c+1}{2}\frac{tpr}{tpr + c \cdot fpr}$	$\frac{tpr}{tpr + fpr}$	$\frac{tpr}{fpr}$
Relative precision*	$\frac{2c}{c+1}\frac{(tpr - fpr)}{tpr + c \cdot fpr}$	$\frac{tpr - fpr}{tpr + fpr}$	
F-measure	$\frac{2tpr}{tpr + c \cdot fpr + 1}$	$\frac{2tpr}{tpr + fpr + 1}$	$\frac{tpr}{fpr + 1/c}$
G-measure	$\frac{tpr}{c \cdot fpr + 1}$	$\frac{tpr}{fpr + 1}$	

All metrics are re-scaled such that the strongly skew-insensitive version is in [0,1] or [-1,1]. An asterisk (*) denotes weak skew-insensitivity.

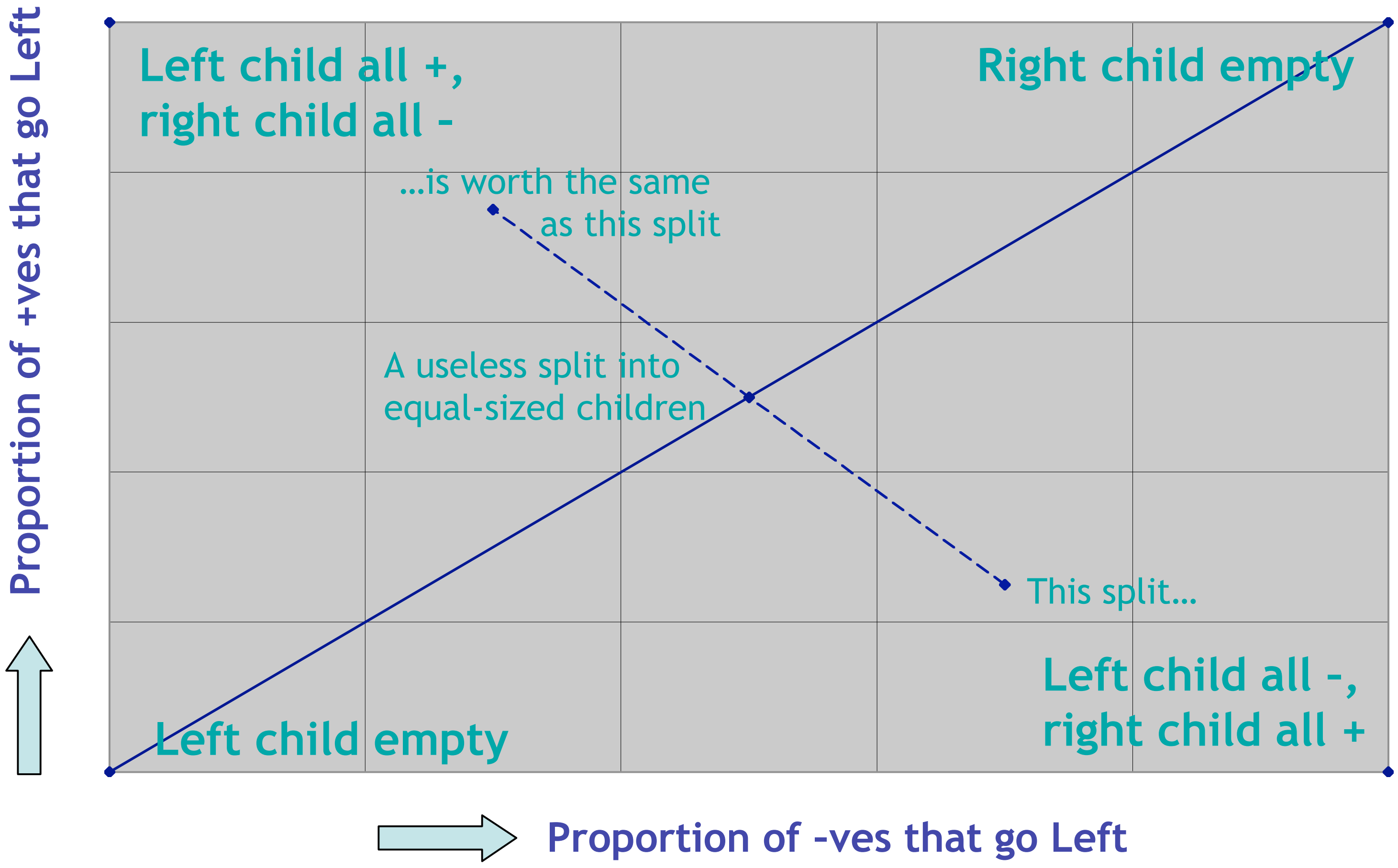
Splitting criteria

	Children		
Parent	TP	FN	Pos
	FP	TN	Neg
	Left	Right	N

- Splitting criteria are invariant under swapping columns, i.e. point-mirroring through (0,0)
  - if cost-insensitive then isometrics are symmetric across both diagonals
- They compare impurity of the parent with weighted average impurity of the children:

$$Imp(Pos/N, Neg/N) - Left/N \cdot Imp(TP/Left, FP/Left) - Right/N \cdot Imp(FN/Right, TN/Right)$$

ROC space for splitting criteria



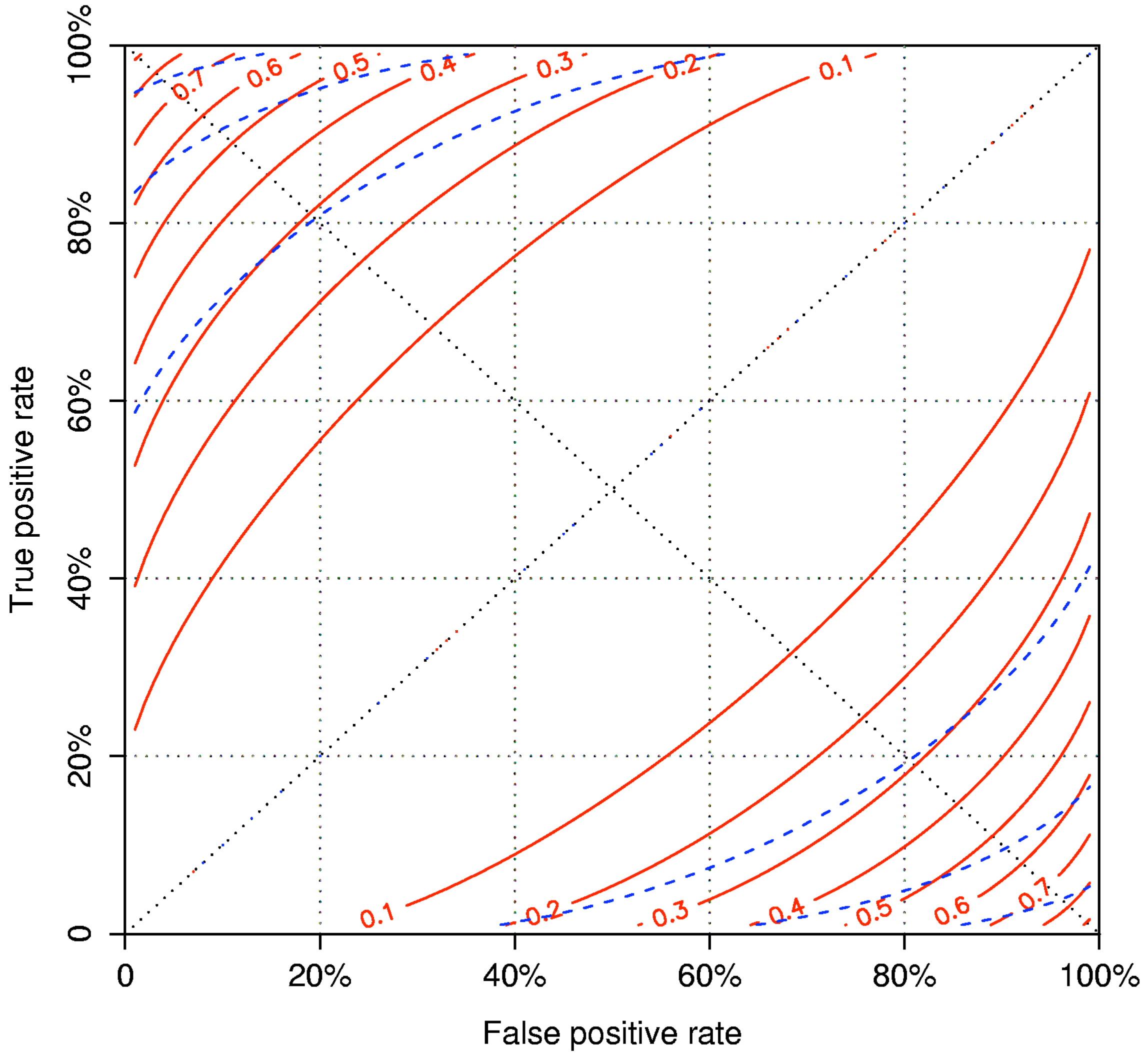
Different impurity measures

- relative impurity is defined as weighted impurity of (left) child in proportion to impurity of parent

Impurity	$Imp(p,n)$	Relative impurity
Entropy	$-p\log p - n\log n$	
Gini index	$4pn$	$\frac{(1+c)\cdot tpr\cdot fpr}{tpr + c\cdot fpr}$
DKM	$2\sqrt{pn}$	$\sqrt{tpr\cdot fpr}$

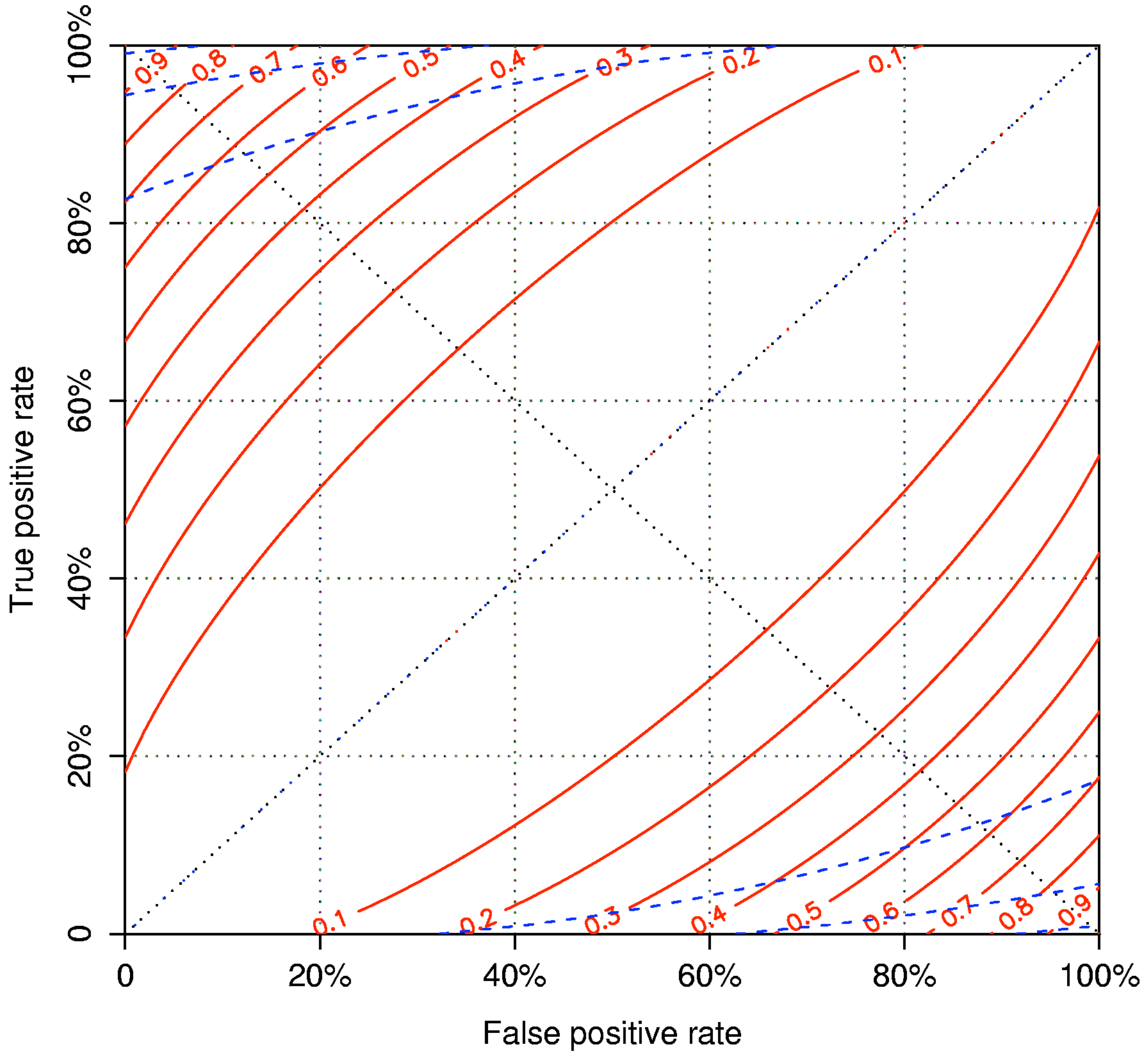
All impurity measures are re-scaled to [0,1]. DKM refers to (Dietterich, Kearns & Mansour, 1996). The cost-insensitivity of DKM-split for binary splits was shown by (Drummond & Holte, 2000).

Information gain isometrics



c = 1,
c = 1/10

Gini-split isometrics



c = 1,
c = 1/10

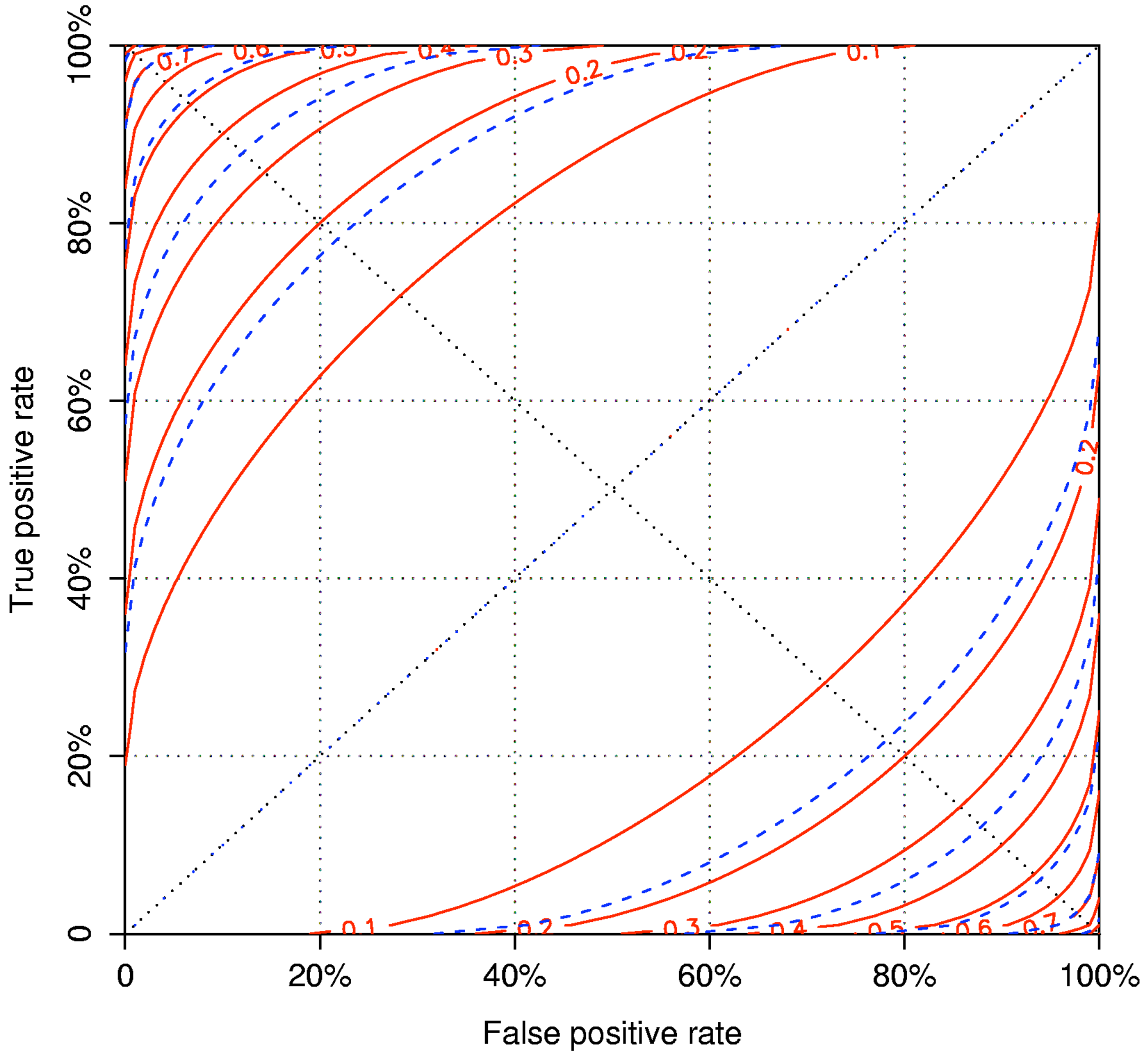
Comments on Gini-split

- More skew-sensitive than information gain
- Equivalent to two-by-two χ^2 normalised by sample size (i.e., ϕ^2)
- Strongly skew-insensitive version obtained by setting c=1:

$$Gini-ROC = 1 - \frac{2 \cdot tpr \cdot fpr}{tpr + fpr} - \frac{2 \cdot (1 - tpr) \cdot (1 - fpr)}{1 - tpr + 1 - fpr}$$

  - impurity of child takes impurity of parent into account
  - no need to weight the impurity of children

DKM-split isometrics



c = 1,

c = 1/10

Skew-insensitive splitting

- The best splits do well on both classes, even with highly unbalanced data sets
- Inflating a class does not change split quality
  - bar rounding errors and tie-breaking
- Skew-sensitivity comes into play when pruning a decision tree

ROC-based model manipulation

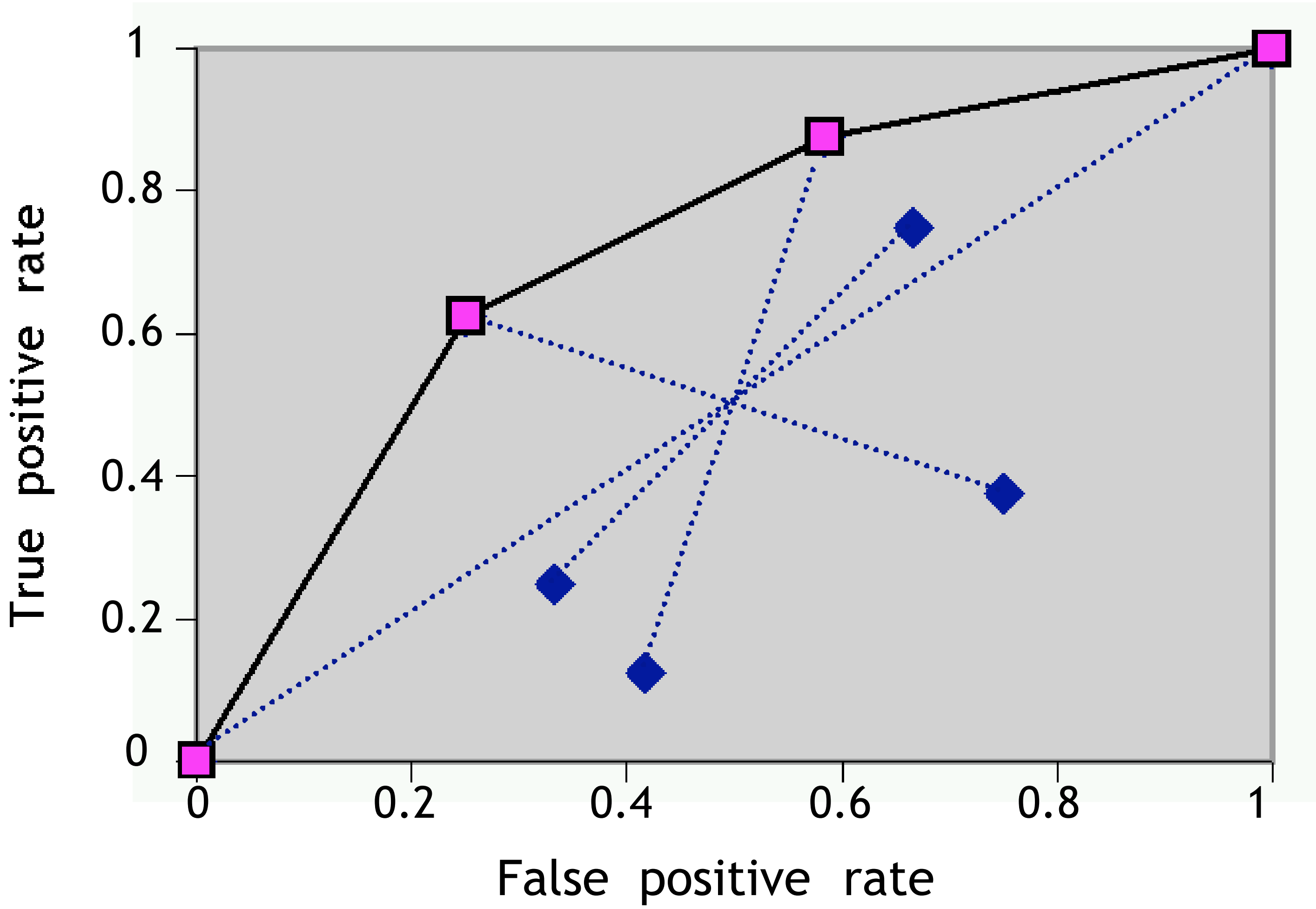
- ROC analysis allows creation of model variants without re-training
  - (Part I) manipulating ranker thresholds
- Example: re-labelling decision trees
  - (Ferri et al., 2002)
- Example: locally adjusting rankings
  - (Flach & Wu, 2003)

Re-labelling decision trees

- A decision tree can be seen as an unlabelled tree (a clustering tree):
  - Given n leaves and 2 classes, there are 2^n possible labellings, each representing a classifier
- Use ROC analysis to select the best labellings

	Training Distribution		Labellings							
	+	-								
Leaf 1	40	20	-	-	-	-	+	+	+	+
Leaf 2	50	10	-	-	+	+	-	-	+	+
Leaf 3	30	50	-	+	-	+	-	+	-	+

DT labellings in ROC space

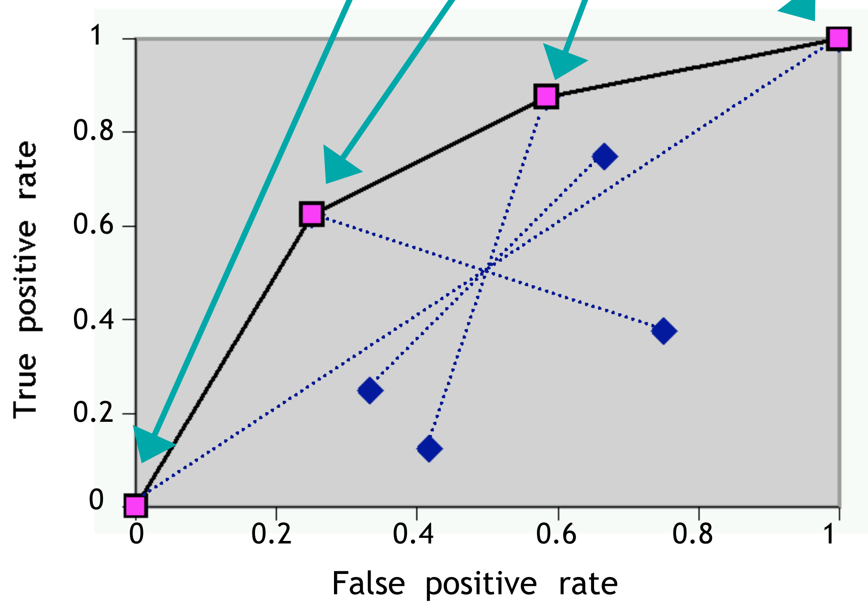


Selecting optimal labellings

1. Rank leaves by likelihood ratio P(l|+)/P(l|–)

2. For each possible split point, label leaves before split + and after split –

	+	-				
Leaf 2	50	10	-	+	+	+
Leaf 1	40	20	-	-	+	+
Leaf 3	30	50	-	-	-	+



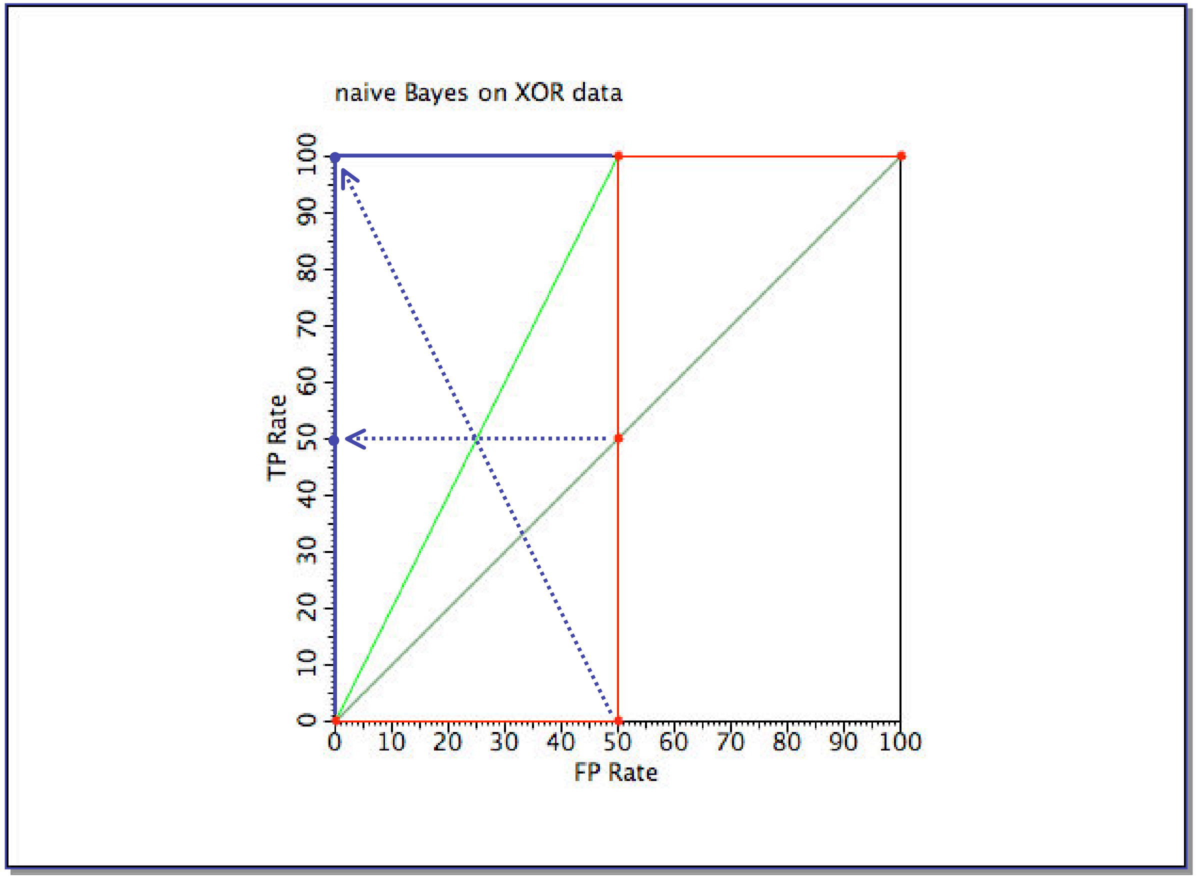
Why does it work?

- Decision trees are rankers if we use class distributions in the leaves
  - Probability Estimation Trees (Provost & Domingos, 2003)
- ROC curve can be constructed by sliding threshold
  - just as with naïve Bayes
- Equivalently, we can order instances, which boils down do ordering leaves
  - because all instances in a leaf are ranked together
- NB. Curve may not be convex on test set

Repairing concavities

- Concavities in ROC curves from rankers indicate worse-than-random segments in the ranking
- Idea 1: use binned ranking (aka discretised scores) —> convex hull
- Idea 2: invert ranking in segment
- Need to avoid overfitting

Example



- Effectively introduces a second decision boundary

Summary of Part II

- Isometric plots visualise the behaviour of machine learning metrics
  - equivalences, skew-sensitivity, skew-insensitive versions
- One model can be many models
  - ROC analysis can be used to obtain alternative labellings of trees, adjust rankings, etc.