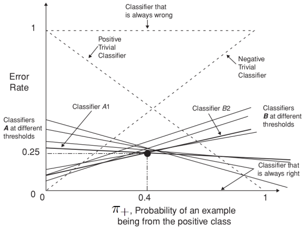
Introduction to Machine Learning

Evaluation: Cost Curves



Learning goals

- Understand cost curves
- As alternative to ROC curves

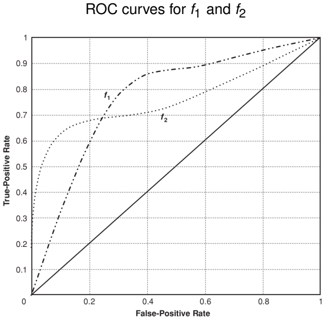
COST CURVES

- Directly plot the misclassif costs / error
- Might be easier to interpret than ROC, especially in case of different misclassif costs or priors

Example:

- f_1 and f_2 with intersecting ROC curves
- f_2 dominates first, then f_1

BUT: Unclear for which thresholds, costs or class distribs f_2 better than f_1



Nathalie Japkowicz (2004): Evaluating Learning Algorithms : A Classification Perspective. (p. 125)

COST CURVES

Simplifying assumption: equal misclassif costs, i.e., $cost_{FN} = cost_{FP}$
$\Rightarrow$ Expected misclassif cost reduces to misclassif error rate
With law of total prob, we write error rate as function of π_+:

$$
\begin{aligned}
\rho_{MCE}(\pi_+) &= (1 - \pi_+) \cdot \mathbb{P}(\hat{y} = 1 | y = 0) + \pi_+ \cdot \mathbb{P}(\hat{y} = 0 | y = 1) \\
&= (1 - \pi_+) \cdot FPR + \pi_+ \cdot FNR \\
&= (FNR - FPR) \cdot \pi_+ + FPR
\end{aligned}
$$

Confusion matrix

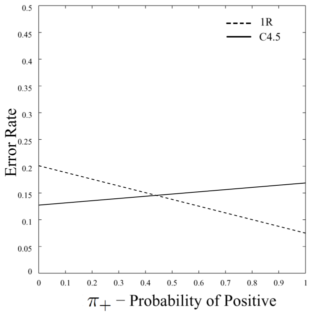
		True class	
		$y = 1$	$y = 0$
Pred.	$\hat{y} = 1$	TP	FP
class	$\hat{y} = 0$	FN	TN

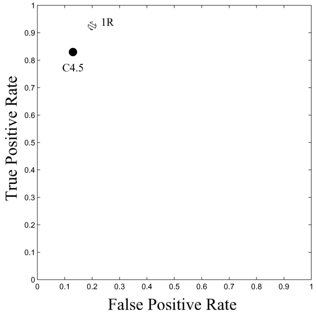
Cost matrix

		True class	
		$y = 1$	$y = 0$
Pred.	$\hat{y} = 1$	0	$cost_{FP}$
class	$\hat{y} = 0$	$cost_{FN}$	0

COST CURVES

- Expected misclassif costs (or error rate if $cost_{FN} = cost_{FP}$) is plotted as function of proportion of positive instances, $\pi_+ = \mathbb{P}(y = 1)$
- Cost curves are point–line duals of ROC curves, i.e., a single classifier is represented by a point in the ROC space and by a line in cost space





Chris Drummond and Robert C. Holte (2006): Cost curves: An improved method for visualizing classifier performance. Machine Learning, 65, 95-130 (URL).

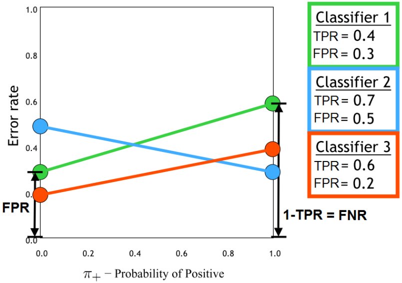
COST LINES

Cost line of a classifier with slope $(FNR - FPR)$ and intercept FPR:

$$\rho_{MCE}(\pi_+) = (FNR - FPR) \cdot \pi_+ + FPR$$

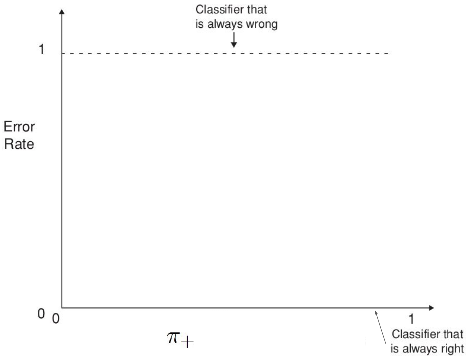
- Hard classifiers are points (TPR, FPR) in ROC space
- The cost line of a classifier connects (π_+, ρ_{MCE})-points at $(0, FPR)$ and $(1, 1 - TPR)$
- Classifier 3 always dominates classifier 1
- Classifier 3 is better than classifier 2 when $\pi_+ < 0.7$

Cost lines plot different values of π_+ vs. $\rho_{MCE}(\pi_+)$



COST LINES - EXAMPLE

- Horizontal dashed line: worst classifier (100% error rate for all π_+) $\Rightarrow FNR = FPR = 1$
- x-axis: perfect classifier (0% error rate for all π_+) $\Rightarrow FNR = FPR = 0$



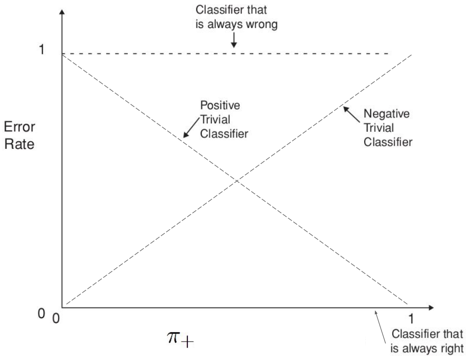
$$\rho_{MCE} = (FNR - FPR) \cdot \pi_+ + FPR$$

Confusion matrix

		True class	
		$y = 1$	$y = 0$
Pred.	$\hat{y} = 1$	TP	FP
class	$\hat{y} = 0$	FN	TN

COST LINES - EXAMPLE

- Horizontal dashed line: worst classifier (100% error rate for all π_+) $\Rightarrow$ *FNR* $=$ *FPR* $= 1$
- x-axis: perfect classifier (0% error rate for all π_+) $\Rightarrow$ *FNR* $=$ *FPR* $= 0$
- Dashed diagonal lines: trivial classifiers, i.e., ascending diagonal always predicts negative instances ($\Rightarrow$ *FNR* $= 1$ and *FPR* $= 0$) and vice versa

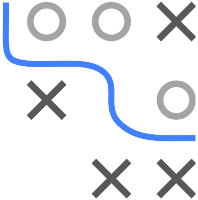


$$\rho_{MCE} = (FNR - FPR) \cdot \pi_+ + FPR$$

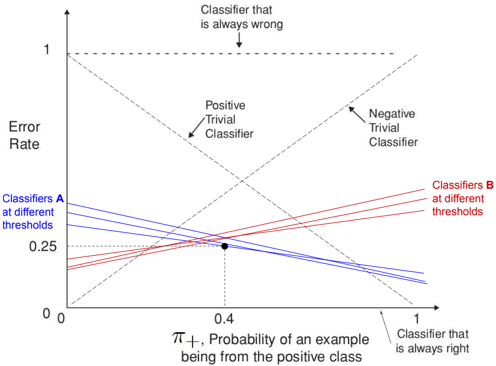
Confusion matrix

		True class	
		$y = 1$	$y = 0$
Pred.	$\hat{y} = 1$	TP	FP
class	$\hat{y} = 0$	FN	TN

COST LINES - EXAMPLE



- Horizontal dashed line: worst classifier (100% error rate for all π_+) $\Rightarrow FNR = FPR = 1$
- x-axis: perfect classifier (0% error rate for all π_+) $\Rightarrow FNR = FPR = 0$
- Dashed diagonal lines: trivial classifiers, i.e., ascending diagonal always predicts negative instances ($\Rightarrow FNR = 1$ and $FPR = 0$) and vice versa
- Descending/ascending bold lines: two families of classifiers A and B (represented by points in their respective ROC curves)



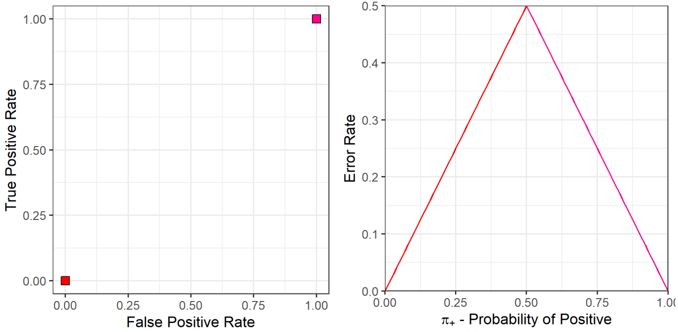
$$\rho_{MCE} = (FNR - FPR) \cdot \pi_+ + FPR$$

Confusion matrix

		True class	
		$y = 1$	$y = 0$
Pred.	$\hat{y} = 1$	TP	FP
class	$\hat{y} = 0$	FN	TN

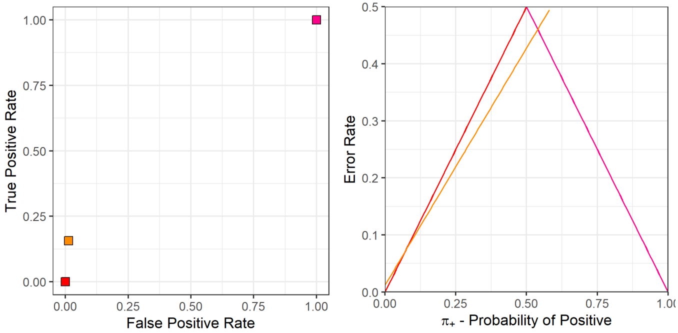
VISUALIZE COST CURVE - LOWER ENVELOPE

- Left: TPR & FPR of a classifier for different prob thresholds
- Right: Corresponding cost lines



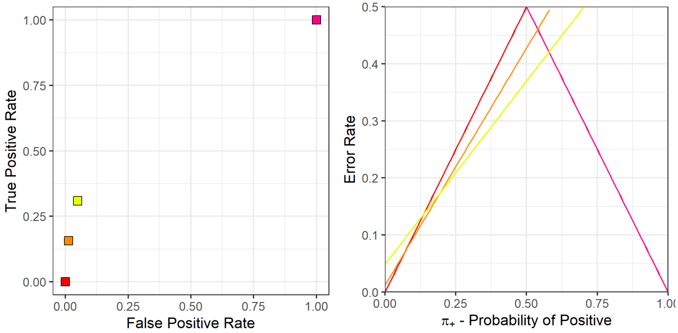
VISUALIZE COST CURVE - LOWER ENVELOPE

- Left: TPR & FPR of a classifier for different prob thresholds
- Right: Corresponding cost lines



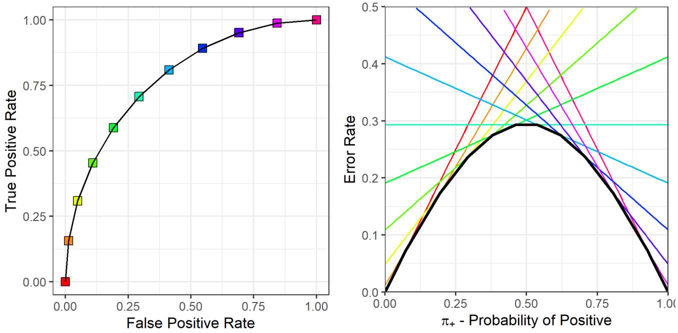
VISUALIZE COST CURVE - LOWER ENVELOPE

- Left: TPR & FPR of a classifier for different prob thresholds
- Right: Corresponding cost lines

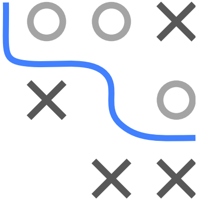


VISUALIZE COST CURVE - LOWER ENVELOPE

- Left: TPR & FPR of a classifier for different prob thresholds
- Right: Corresponding cost lines
- **Cost curve** (right: black line) is lower envelope of **cost lines**
  $\hat{=}$ pointwise minimum of error rate (as function of π_+)



CONSIDER COSTS



Now: Assume unequal misclassif costs, i.e., $cost_{FN} \neq cost_{FP}$ and generalize error rate to **expected costs** (as function of π_+):

$$Costs(\pi_+) = (1 - \pi_+) \cdot FPR \cdot cost_{FP} + \pi_+ \cdot FNR \cdot cost_{FN}$$

Maximum of expected costs happens when

$$FPR = FNR = 1 \;\Rightarrow\; Costs_{max} = (1 - \pi_+) \cdot cost_{FP} + \pi_+ \cdot cost_{FN}$$

Consider **normalized costs** (as function of π_+):

$$\begin{aligned} Costs_{norm}(\pi_+) &= \frac{(1-\pi_+)\cdot FPR\cdot cost_{FP} + \pi_+\cdot FNR\cdot cost_{FN}}{(1-\pi_+)\cdot cost_{FP} + \pi_+\cdot cost_{FN}} \\ &= \frac{(1-\pi_+)\cdot cost_{FP}\cdot FPR}{(1-\pi_+)\cdot cost_{FP} + \pi_+\cdot cost_{FN}} + \frac{\pi_+\cdot cost_{FN}\cdot FNR}{(1-\pi_+)\cdot cost_{FP} + \pi_+\cdot cost_{FN}} \end{aligned}$$

Let "probability times cost" $PC(+)$ be normalized version of $\pi_+ \cdot cost_{FN}$:

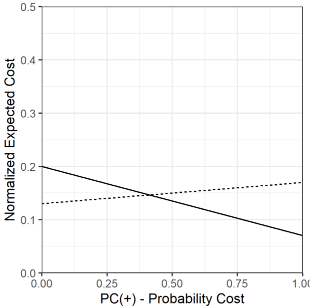
$$PC(+) = \frac{\pi_+\cdot cost_{FN}}{(1-\pi_+)\cdot cost_{FP}+\pi_+\cdot cost_{FN}} \text{ and } 1 - PC(+) = \frac{(1-\pi_+)\cdot cost_{FP}}{(1-\pi_+)\cdot cost_{FP}+\pi_+\cdot cost_{FN}}$$

CONSIDER COSTS / 2

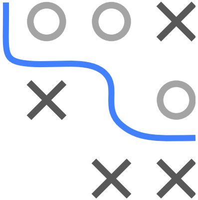
To obtain cost lines, we need a function with slope $(FNR - FPR)$ and intercept $FPR \Rightarrow$ Rewrite $Costs_{norm}(\pi_+)$ as function of $PC(+)$:

$$
\begin{aligned}
Costs_{norm}(PC(+)) &= (1 - PC(+)) \cdot FPR + PC(+) \cdot FNR \\
&= (FNR - FPR) \cdot PC(+) + FPR \\
&= \begin{cases} FPR, \text{ if } PC(+) = 0 \\ FNR, \text{ if } PC(+) = 1 \end{cases}
\end{aligned}
$$

- Plot is similar to simplified case with $cost_{FN} = cost_{FP}$
- Axes' labels and their interpretation have changed
- Normalized cost vs. "probability times cost"

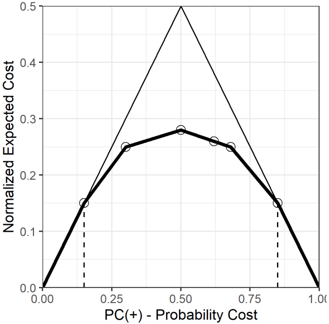


COMPARE WITH TRIVIAL CLASSIFIERS



- Operating range of a classifier is a set of $PC(+)$ values (operating points) where classifier performs better than both trivial classifiers
- Intersection of cost curves and trivial classifiers' diagonals determine operating range
- At any $PC(+)$ value, the vertical distance of trivial diagonal to a classifer's cost curve within operating range shows advantage in performance (normalized costs) of classifier

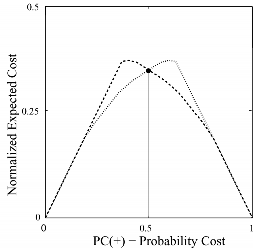
Example: Dotted lines are operating range of a classifier (here: $[0.14, 0.85]$)



COMPARING CLASSIFIERS

- If classifier C1's expected cost is lower than classifier C2's at a $PC(+)$ value, C1 outperforms C2 at that operating point
- The two cost curves of C1 and C2 may cross, which indicates C1 outperforms C2 for a certain operating range and vice versa
- The vertical distance between the two cost curves of C1 and C2 at any $PC(+)$ value directly indicates the performance difference between them at that operating point

Example: Dotted cost curve has lower expected cost as dashed cost curve for $PC(+) < 0.5$ and hence outperforms dashed one in this operating range and vice versa



Chris Drummond and Robert C. Holte (2006): Cost curves: An improved method for visualizing classifier performance. Machine Learning, 65, 95-130 (URL)

ROC CURVES VS. COST CURVES

- A point/line in ROC space is represented by a line/point in cost space, and vice versa
- Area under an ROC curve is a ranking measure while area under a cost curve is the expected cost of the classifier (assuming that all possible $PC(+)$ values are equally likely)
- ROC curves do not indicate for which prob threshold classifier A is superior to another classifier B, cost curves can do exactly that!
  $\Rightarrow$ Cost curves practically more useful than ROC curves
- Cost curves allows users to measure quantitative performance difference between multiple classifiers at any given operating point
  $\Rightarrow$ Not so easy to do that with ROC curve