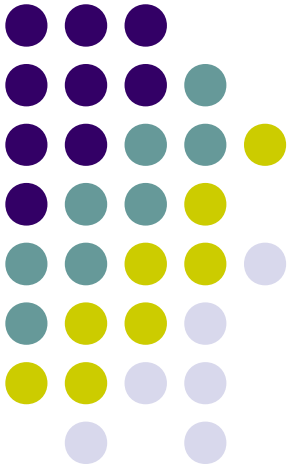
Pattern Recognition

Two-class problems
The ROC curve

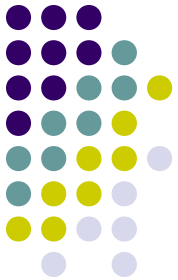
Francesco Tortorella

University of Cassino and Southern Latium

Cassino, Italy



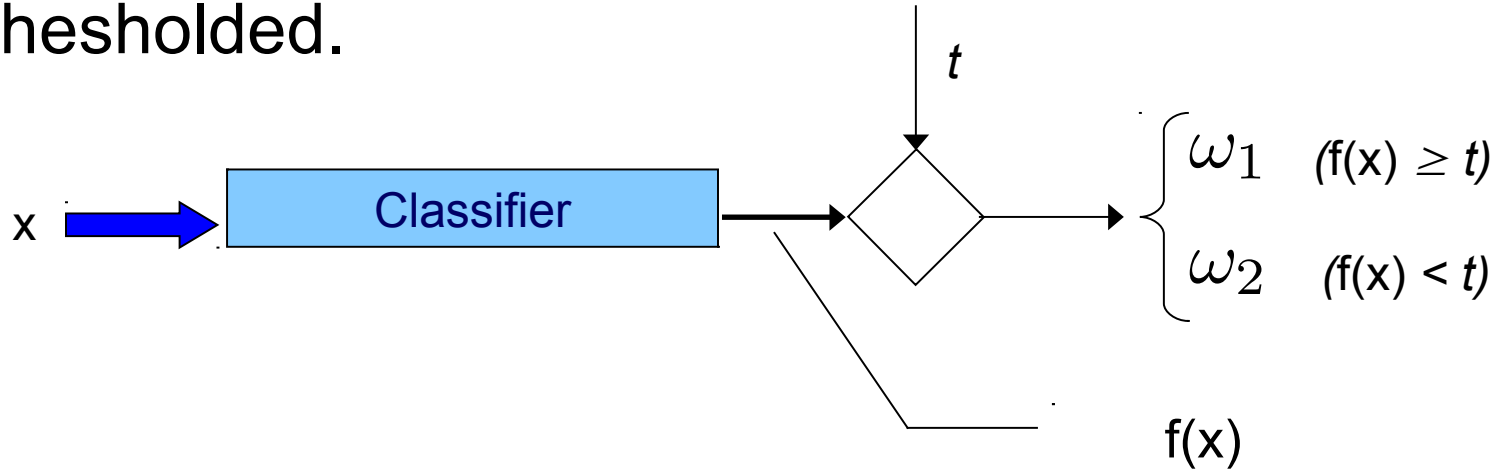
Two class problems performance assessment



- General decision criterion for two-class problems based on Bayes approach:
  - Likelihood ratio: $$\frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \frac{P(\omega_2)}{P(\omega_1)}$$
- It can be written in this form: $$L(\mathbf{x}) \underset{\omega_2}{\overset{\omega_1}{\gtrless}} \gamma$$

Two class problems performance assessment

- This is true also for “real classifier” which tipically provide as output a continuous value to be thesholded.

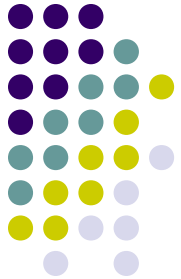


- In this case, the decision depends on the value chosen for the threshold t.

Two class problems performance assessment

- The decision threshold is chosen looking at the priors and at the cost matrix
- The LHS member is independent of such parameters
- When evaluating the decision criterion (the classifier) we have to take into account this.
- Actually, we deal with two (very different) kinds of evaluations:
  - The quality of the classifier (in terms of risk or error probability) for a given setting of the costs and of the priors (for a given *decision bias*)
  - The intrinsic capability of the classifier in discriminating between the two classes, independently of the decision bias

Two class problems performance assessment



- The Bayes criterion minimizes the conditional risk *R*

$$R = \int_{R_1} \left[\lambda_{11} P_1 p(\mathbf{x}|\omega_1) + \lambda_{12} P_2 p(\mathbf{x}|\omega_2) \right] d\mathbf{x} + \int_{R_2} \left[\lambda_{21} P_1 p(\mathbf{x}|\omega_1) + \lambda_{22} P_2 p(\mathbf{x}|\omega_2) \right] d\mathbf{x}$$

which can be written:

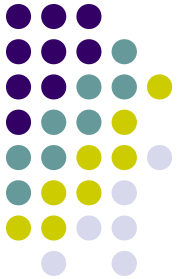
$$R = \lambda_{11} P_1 \delta_1 + \lambda_{12} P_2 \varepsilon_2 + \lambda_{21} P_1 \varepsilon_1 + \lambda_{22} P_2 \delta_2$$

where:

$$\delta_1 = \int_{R_1} p(\mathbf{x}|\omega_1) d\mathbf{x} \qquad \delta_2 = \int_{R_2} p(\mathbf{x}|\omega_2) d\mathbf{x}$$

$$\varepsilon_1 = \int_{R_2} p(\mathbf{x}|\omega_1) d\mathbf{x} \qquad \varepsilon_2 = \int_{R_1} p(\mathbf{x}|\omega_2) d\mathbf{x}$$

Two class problems performance assessment



- The conditional risk can be written as a function of the threshold γ:

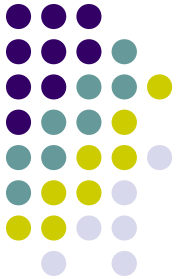
$$R(\gamma) = \lambda_{11} P_1 \delta_1(\gamma) + \lambda_{12} P_2 \varepsilon_2(\gamma) + \lambda_{21} P_1 \varepsilon_1(\gamma) + \lambda_{22} P_2 \delta_2(\gamma)$$

- Alternatives:

$$R(\gamma) = \lambda_{11} P_1 \delta_1(\gamma) + \lambda_{12} P_2 \varepsilon_2(\gamma) + \lambda_{21} P_1 \left[1 - \delta_1(\gamma)\right] + \lambda_{22} P_2 \left[1 - \varepsilon_2(\gamma)\right]$$

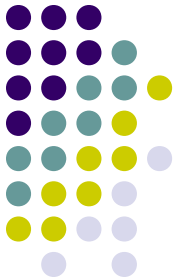
$$R(\gamma) = \lambda_{11} P_1 \left[1 - \varepsilon_1(\gamma)\right] + \lambda_{12} P_2 \left[1 - \delta_2(\gamma)\right] + \lambda_{21} P_1 \varepsilon_1(\gamma) + \lambda_{22} P_2 \delta_2(\gamma)$$

Two class problems performance assessment



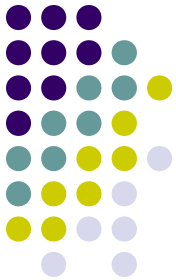
- Two factors:
  - λ_{ij} , P_i : deriving from the particular problem
  - $\delta_1(\gamma)\ \varepsilon_2(\gamma)\ \delta_2(\gamma)\ \varepsilon_1(\gamma)$: intrinsic to the classifier
- In order to evaluating the discriminating quality of the classifier, one can evaluate the behavior of $\delta_1(\gamma)\ \varepsilon_2(\gamma)$ or $\delta_2(\gamma)\ \varepsilon_1(\gamma)$ when γ varies.
- This means to consider the performance of the classifier on each class separately.

Another reason ...



- For measuring the performance of a classification system, the typical measure is the *accuracy*, defined as the percentage of correctly classifed samples (on a test set).
- When dealing with unbalanced data (very frequent in 2-class problems), accuracy is not adequate.
- Example:
  - $P(\omega_1)=0.95 \quad P(\omega_2)=0.05$
  - $f(x)=\omega_1 \ \forall x \ \Rightarrow$ accuracy=95%

Evaluating the performance



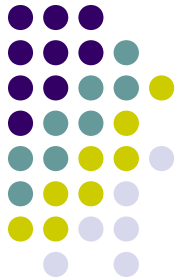
- We must look at the performance on each class.

		Actual class	
		ω_1	ω_2
Predicted class	ω_1	95	5
	ω_2	0	0

Accuracy on class ω_1 = 95/95 = 100%
Accuracy on class ω_2 = 0/5 = 0%

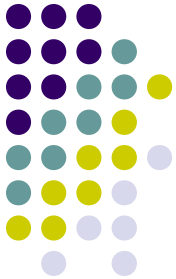
- Let's call P (Positive) and N (Negative) the two classes.

Evaluating the performance



- In the case of two-class problems it is necessary to measure the accuracy for each class separately also for avoiding a bias due to the class skew.
- This is common to several different application contexts:
  - Hypothesis testing: type I error, type II error
  - Radar detection: P_F (false alarm),P_M (miss),P_D (detection) (P_H hit), P_{CR} (correct rejection)
  - Medical diagnosis: TPR, FPR, TNR, FNR (True Positive, False Positive, True Negative, False Negative)

Evaluating the performance



- We can consider some figures:
  - TRUE POSITIVE RATE (TPR):

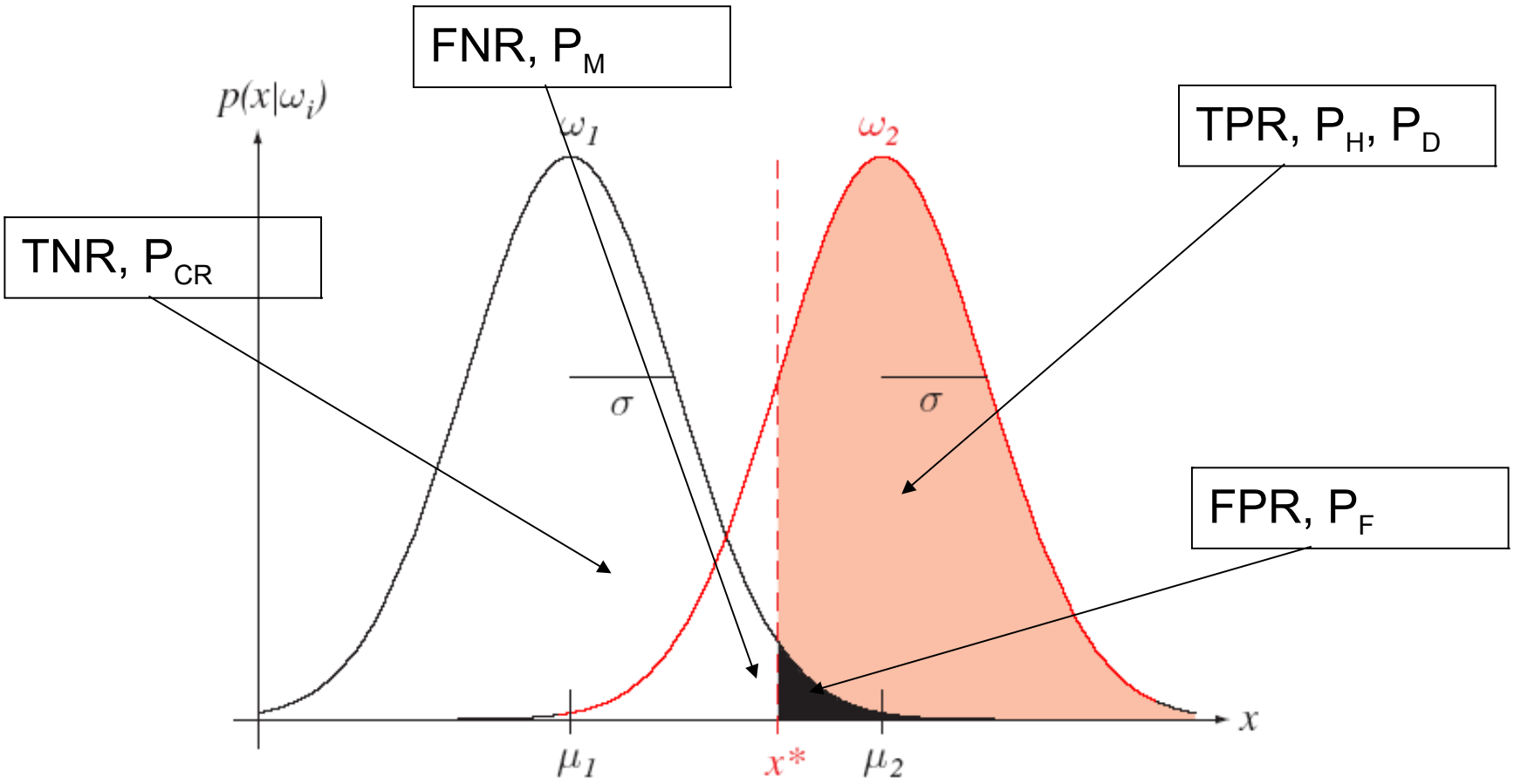
$$\frac{\text{Number of positive samples correctly classified}}{\text{Number of positive samples}}$$

  - FALSE POSITIVE RATE (FPR):

$$\frac{\text{Number of negative samples erroneously classified}}{\text{Number of negative samples}}$$

  - FNR=1-TPR TNR=1-FPR

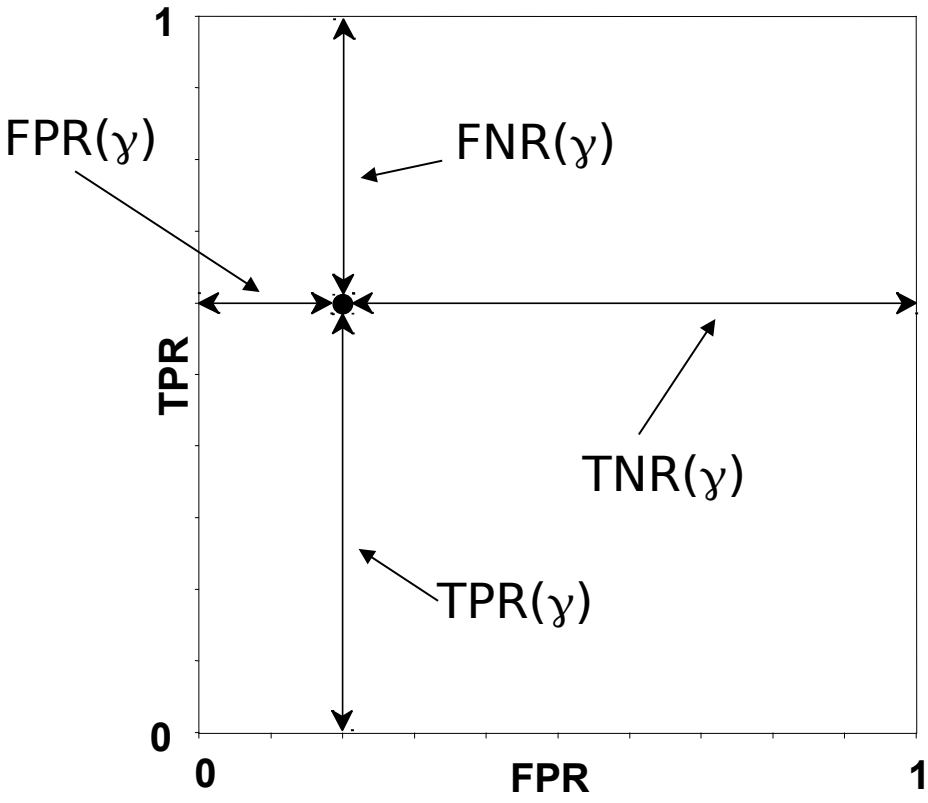
Evaluating the performance



Evaluating the performance

- Our choice:
  - TPR=1-FNR
  - FPR=1-TNR
- Two possible assumptions:
  - TPR(γ)= $\delta_1(\gamma)$ FPR(γ)=$\varepsilon_2(\gamma)$
  - TPR(γ)= $\delta_2(\gamma)$ FPR(γ)=$\varepsilon_1(\gamma)$
- Such values can be drawn on a particular plane (the *ROC plane* or *ROC space).*

the ROC plane



Because of the relations:

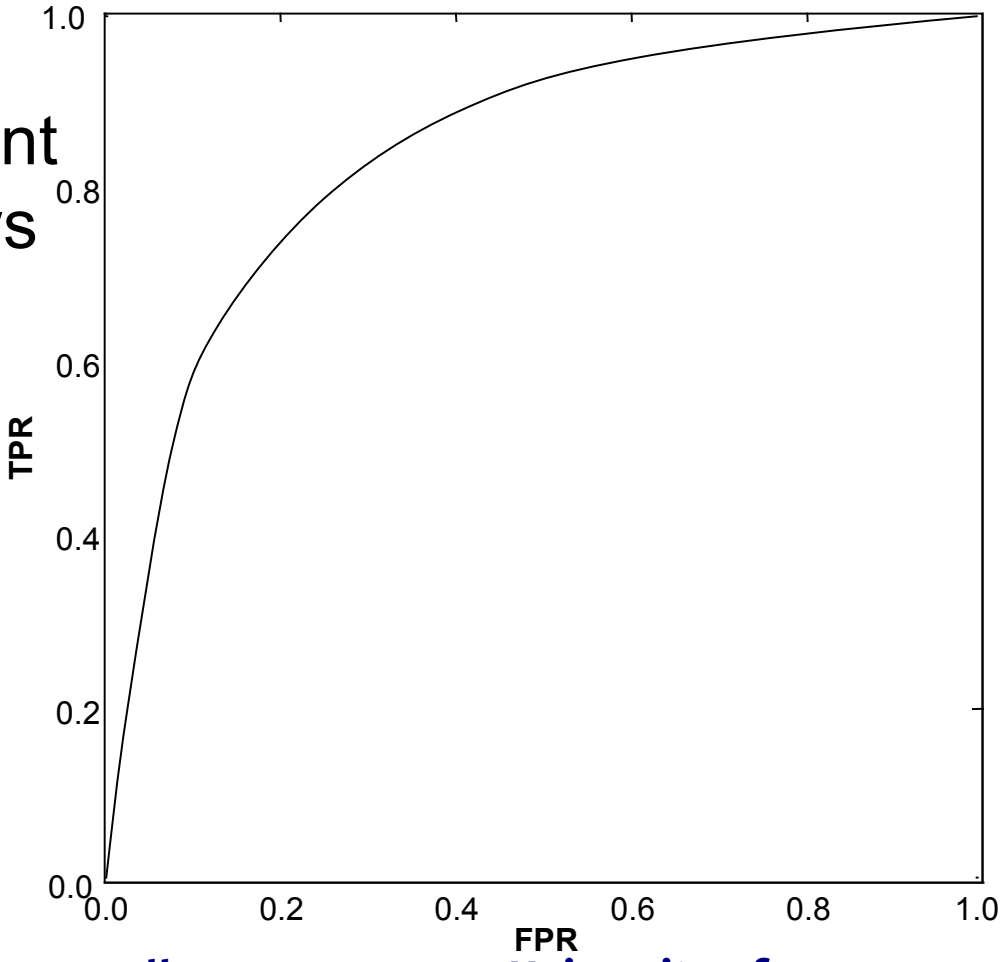
FNR(t)=1- TPR(γ)

FPR(t)=1- TNR(γ)

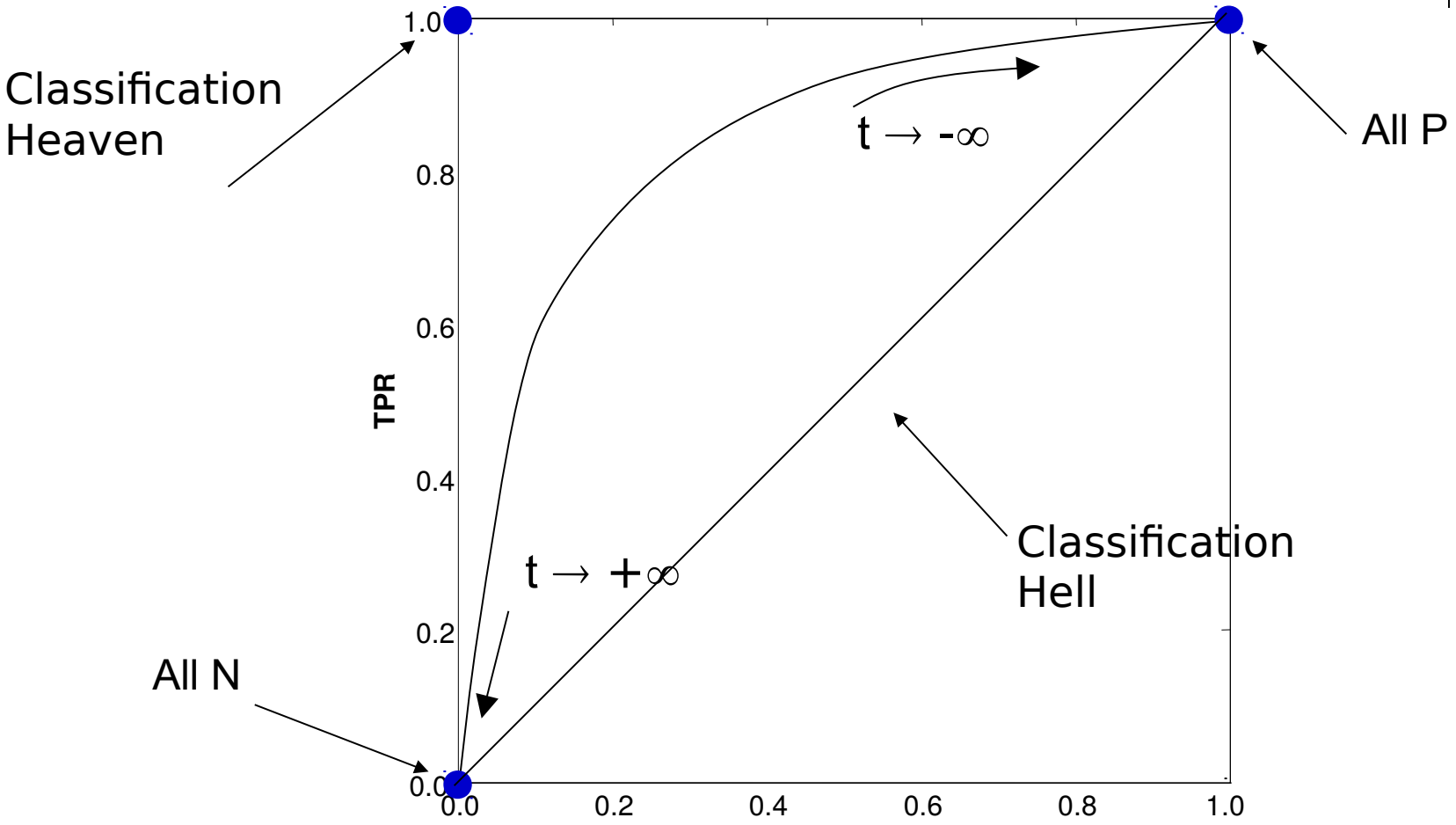
only two values are sufficient to have a complete evaluation of the classifier.
E.g. FPR(γ) and TPR(γ)

The ROC curve

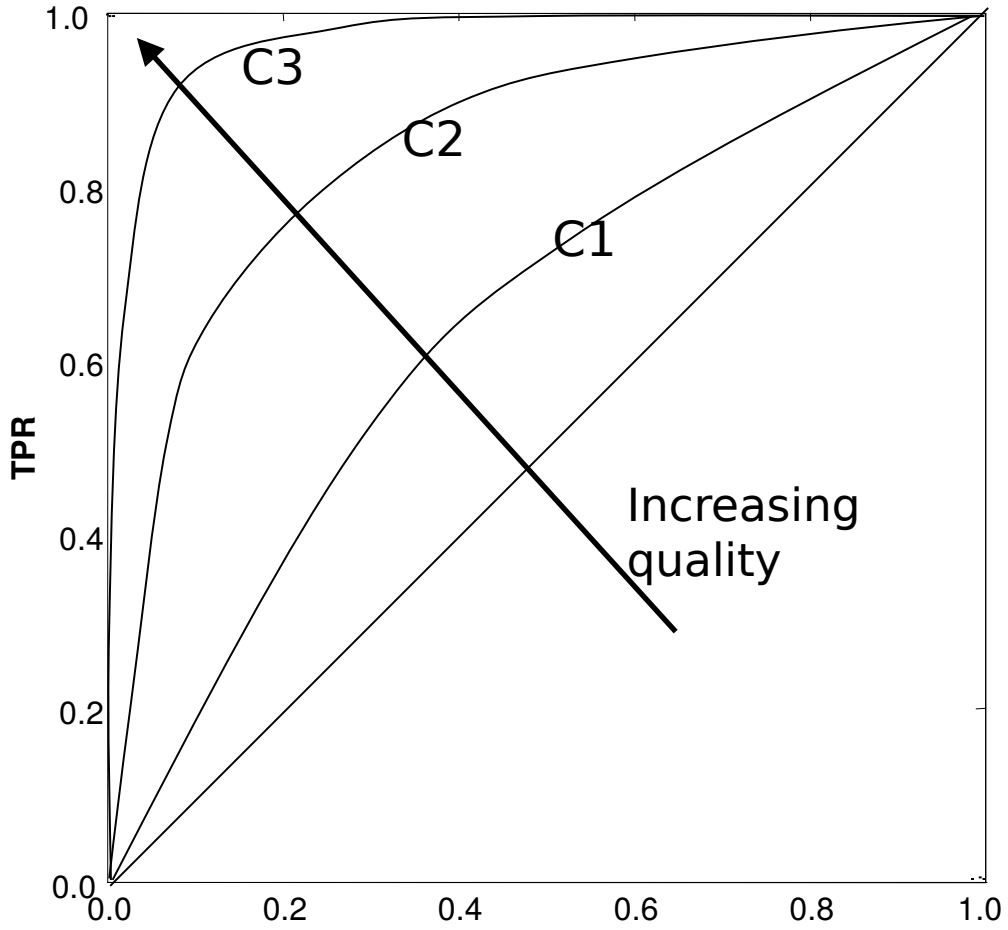
- When the threshold varies, the moving point (FPR(γ), TPR(γ) draws the ROC (*Receiver Operating Characteristic*) curve of the classifier.



The Receiver Operating Characteristic (ROC) curve



Comparing classifiers



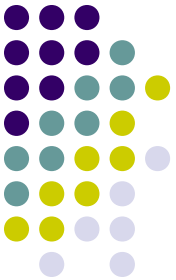
We can say that C2 is definitely better than C1 if the ROC curve of C2 dominates the curve of C1

Es. C3 > C2 > C1

ROC curve and LRT: properties

- The ROC curve is concave downward
- It is above the line TPR=FPR.
- The slope of a curve in a ROC at a particular point is equal to the value of the threshold $\gamma = \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \frac{P(\omega_2)}{P(\omega_1)}$ required to achieve the TPR and FPR of that point.

Isocost lines



- Let us assume TPR(γ)= $\delta_1(\gamma)$ e FPR(γ)= $\varepsilon_2(\gamma)$ and consider the equation of the risk as a function of the threshold γ:

$$R(\gamma) = \lambda_{11} P_1 \delta_1(\gamma) + \lambda_{12} P_2 \varepsilon_2(\gamma) + \lambda_{21} P_1 \left[1 - \delta_1(\gamma)\right] + \lambda_{22} P_2 \left[1 - \varepsilon_2(\gamma)\right]$$

- The risk associated to a point (FPR,TPR) on the ROC plane will be:

$$\lambda_{11} P_1 \cdot TPR + \lambda_{12} P_2 \cdot FPR + \lambda_{21} P_1 \left[1 - TPR\right] + \lambda_{22} P_2 \left[1 - FPR\right]$$

that can be written as:

$$P_1 \cdot (\lambda_{11} - \lambda_{21}) \cdot TPR + P_2 \cdot (\lambda_{12} - \lambda_{22}) \cdot FPR + P_1 \cdot \lambda_{21} + P_2 \cdot \lambda_{22}$$

Isocost lines

- If two different points (FPR_1,TPR_1) and (FPR_2,TPR_2) have associated the same value for the risk, we have:

$$P_1 \cdot (\lambda_{11} - \lambda_{21}) \cdot TPR_1 + P_2 \cdot (\lambda_{12} - \lambda_{22}) \cdot FPR_1 =$$
$$P_1 \cdot (\lambda_{11} - \lambda_{21}) \cdot TPR_2 + P_2 \cdot (\lambda_{12} - \lambda_{22}) \cdot FPR_2$$

and thus:

$$\frac{TPR_2 - TPR_1}{FPR_2 - FPR_1} = \frac{P_2}{P_1} \cdot \frac{(\lambda_{12} - \lambda_{22})}{(\lambda_{21} - \lambda_{11})}$$

Isocost lines

- The equation defines the slope of an *isocost line*. In other words, all the points (FPR,TPR) of the ROC plane on the line

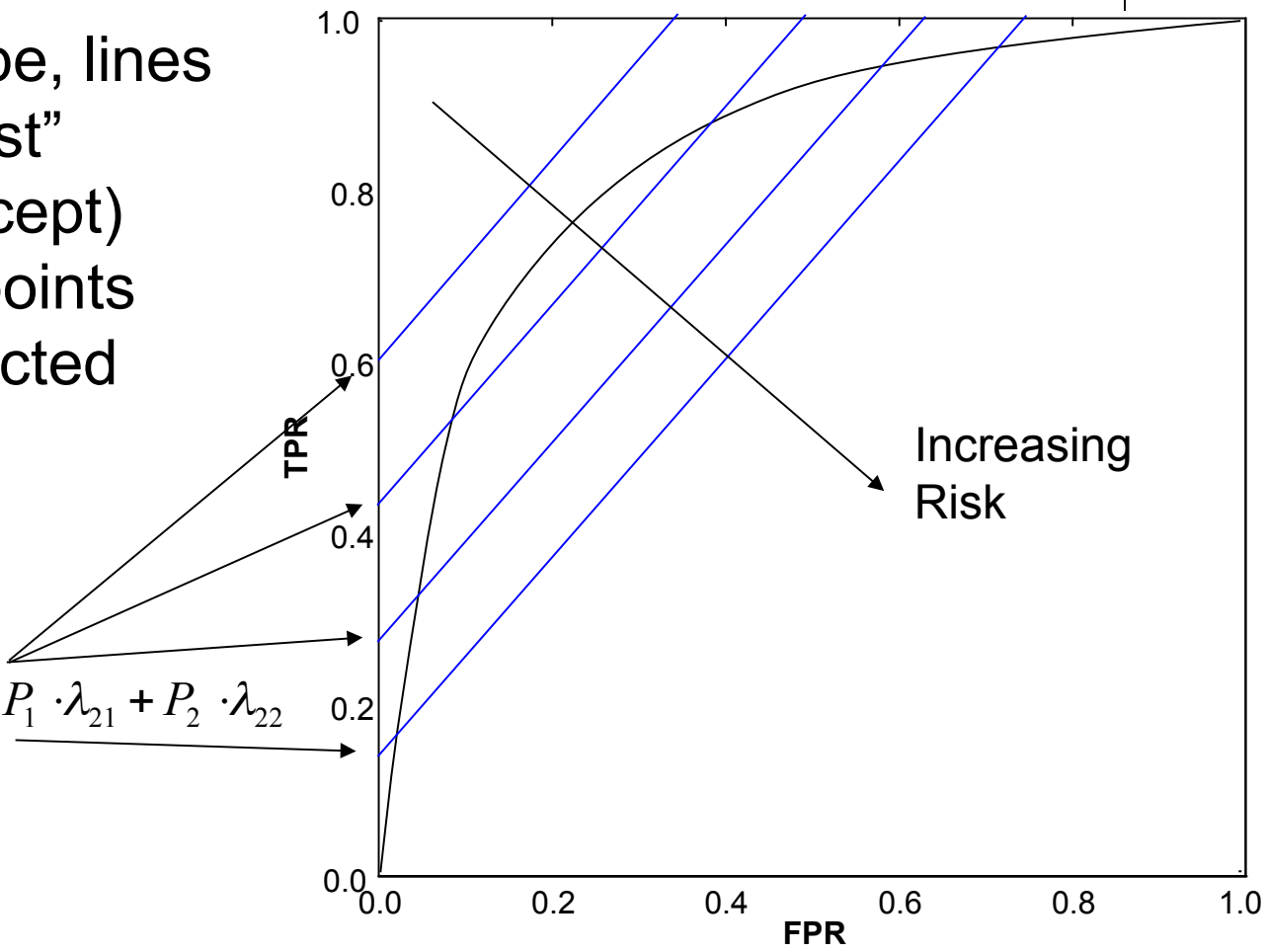
$$\frac{TPR - TPR_1}{FPR - FPR_1} = \frac{P_2}{P_1} \cdot \frac{(\lambda_{12} - \lambda_{22})}{(\lambda_{21} - \lambda_{11})}$$

  will provide the same conditional risk .

- Each combination of priors and costs defines a family of parallel straight lines.

Isocost lines

For a given slope, lines more “north-west” (larger TP intercept) correspond to points with lower expected cost.

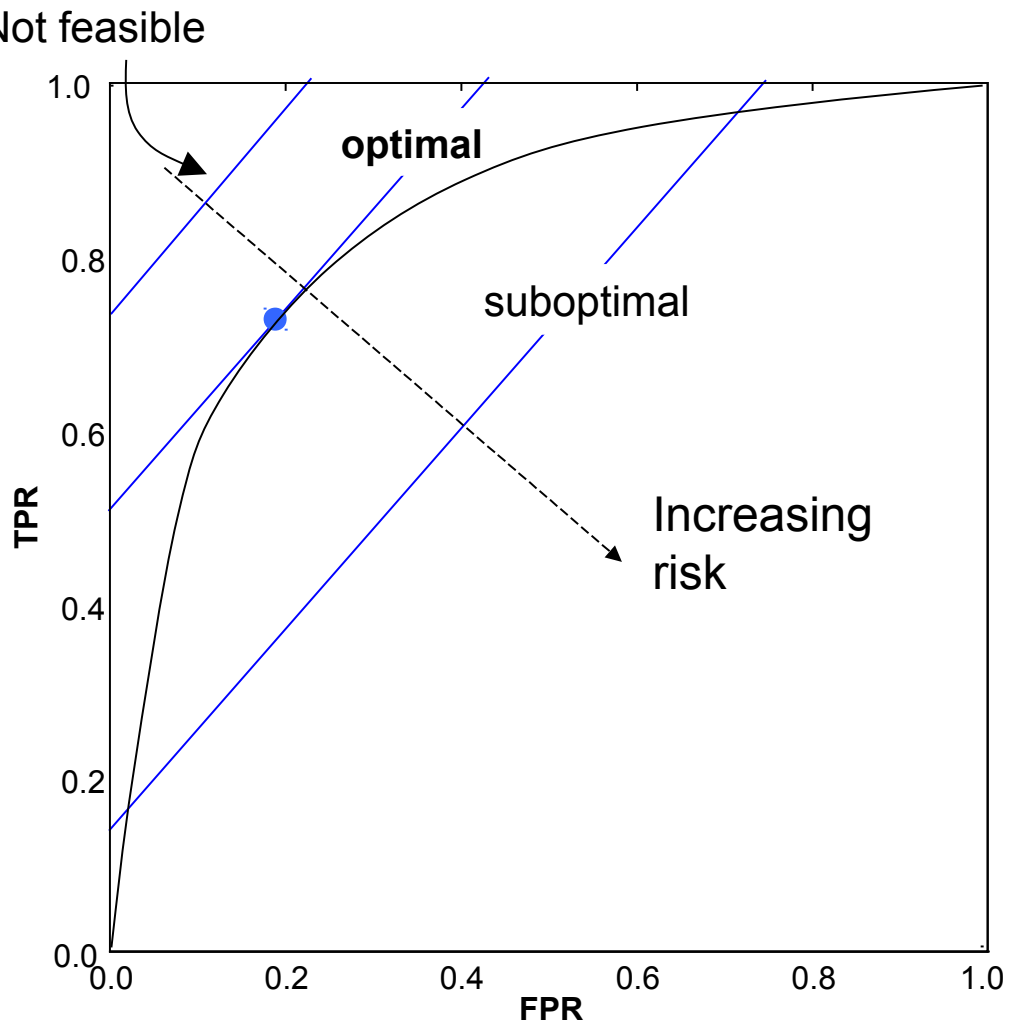


$$R = P_1 \cdot (\lambda_{11} - \lambda_{21}) \cdot TPR_0 + P_1 \cdot \lambda_{21} + P_2 \cdot \lambda_{22}$$

Choosing the optimal operating point

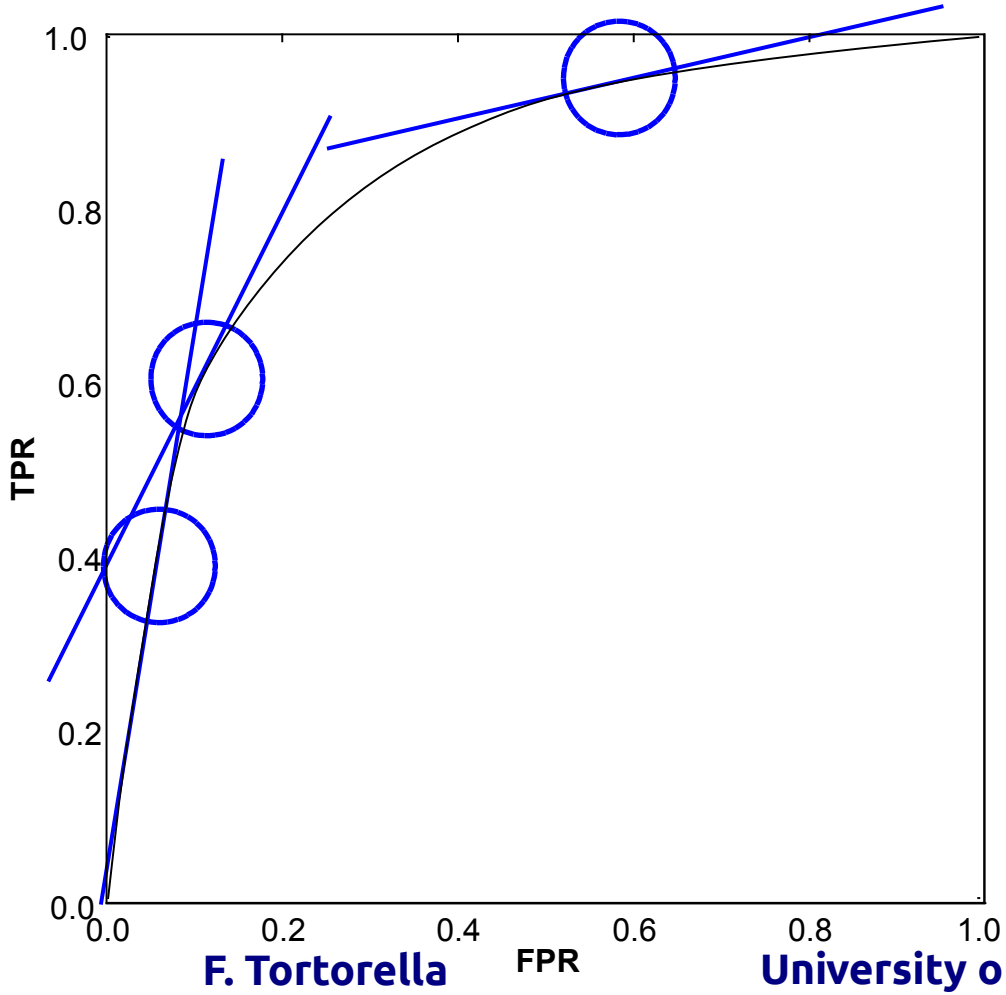
- For a given combination of priors and costs, which is the optimal operating point on a ROC curve?
- Such point must belong
  - to the ROC curve
  - to the “most north-west” line

Choosing the optimal operating point

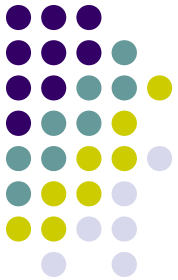


Choosing the optimal operating point

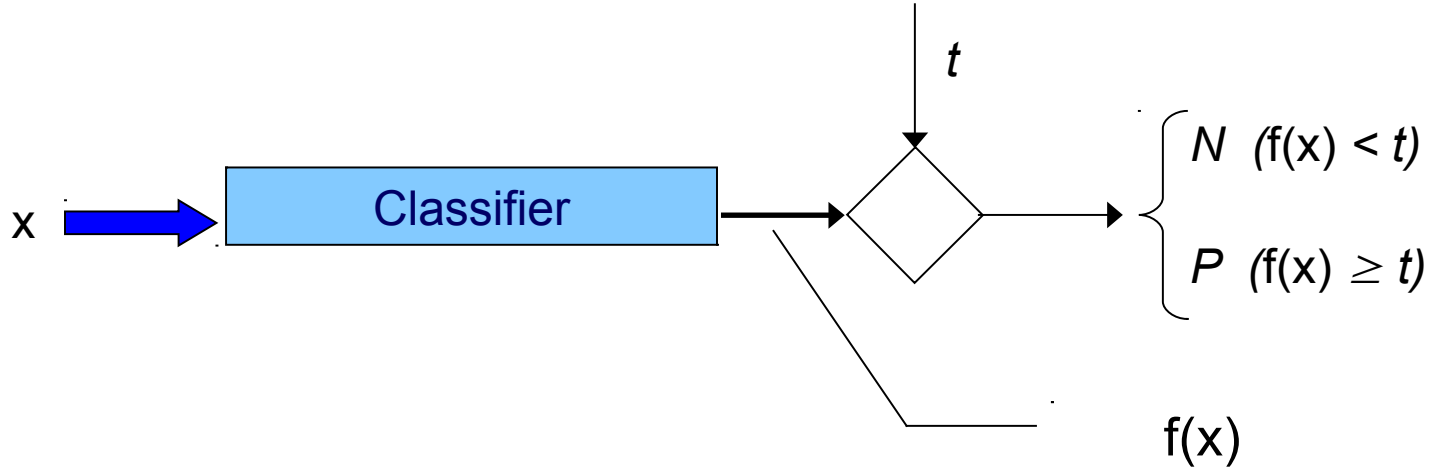
When the prior ratio and/or the cost ratio change, also the optimal operating point changes.



Evaluating the performance for a real classifier



- Assume that our classifier provides as output a continuous value to be thesholded (e.g. LR).



- In this case, TPR and FPR depend on the value chosen for the threshold t.

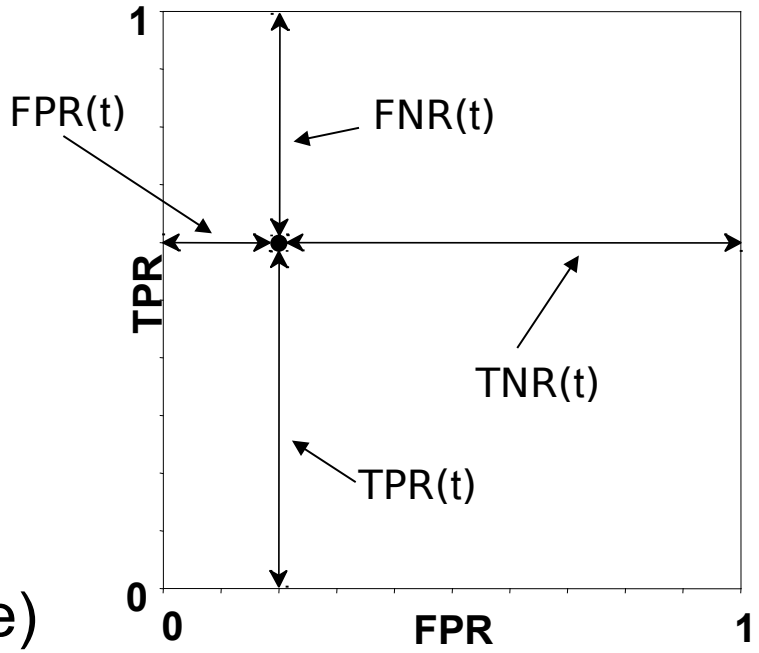
Evaluating the performance

$$TPR(t) = \frac{|x \in P, f(x) \geq t|}{|P|}$$

		Actual class	
		N	**P**
Predicted class	**N**	TNR(t)	FPR(t)
	P	FNR(t)	TPR(t)

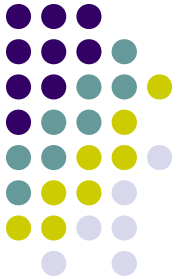
$$TNR(t) = \frac{|x \in N, f(x) < t|}{|N|}$$

$$FPR(t) = \frac{|x \in N, f(x) \geq t|}{|N|}$$

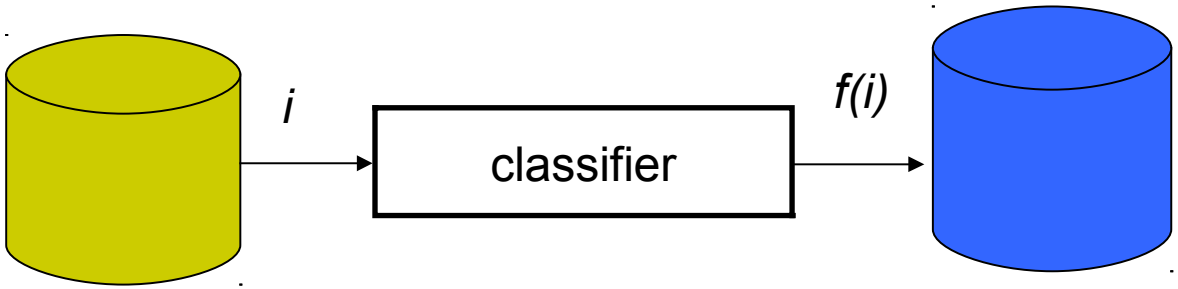


- We can visualize two of the four measures on a plane (ROC plane)

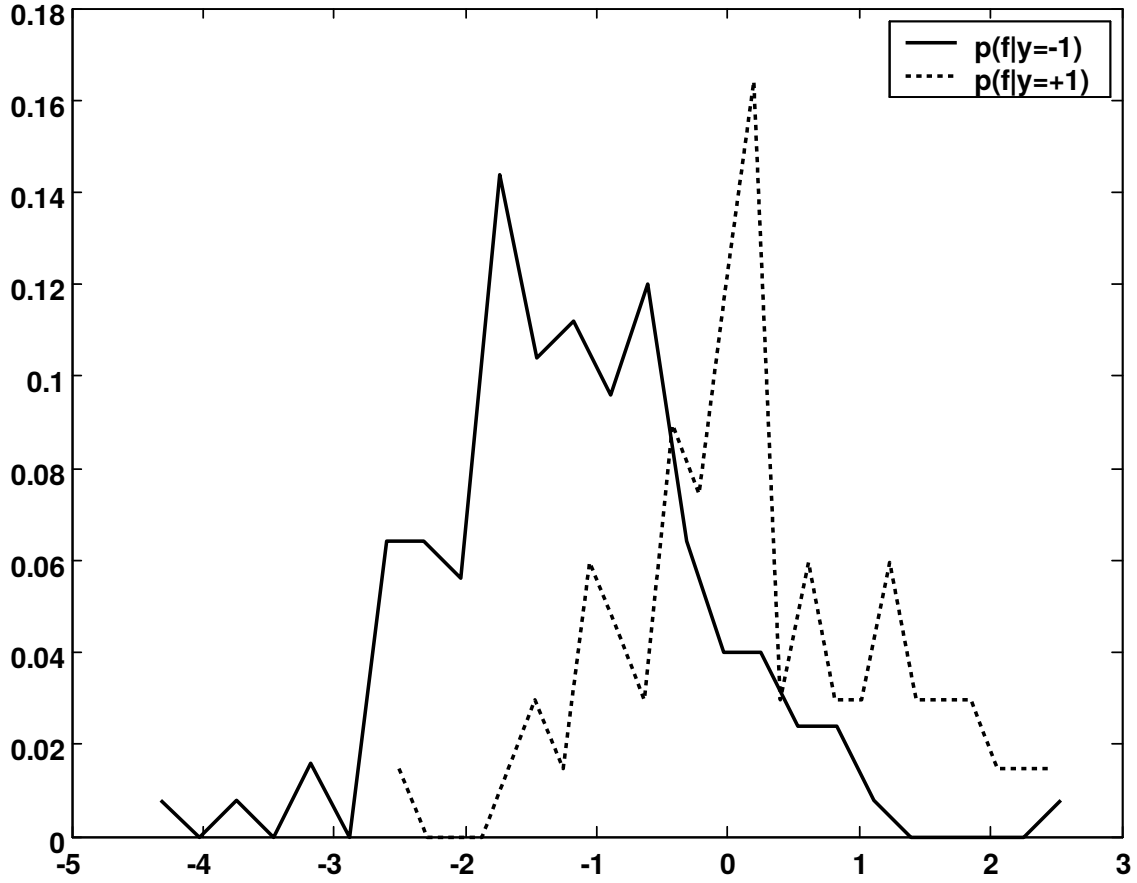
Generating the ROC curve for a classifier



- Assume a set L of samples (P positive samples and N negative samples) is available. The samples were not used for building the classifier.
- Submit each sample *i* to the classifier, and let *f(i)* be the score provided by the classifier.



Generating the ROC curve for a classifier



Generating the ROC curve for a classifier

Algorithm 1 Efficient method for generating ROC points

Inputs: L, the set of test examples; $f(i)$, the probabilistic classifier's estimate that example i is positive; P and N, the number of positive and negative examples.

Outputs: R, a list of ROC points increasing by *fp rate*.

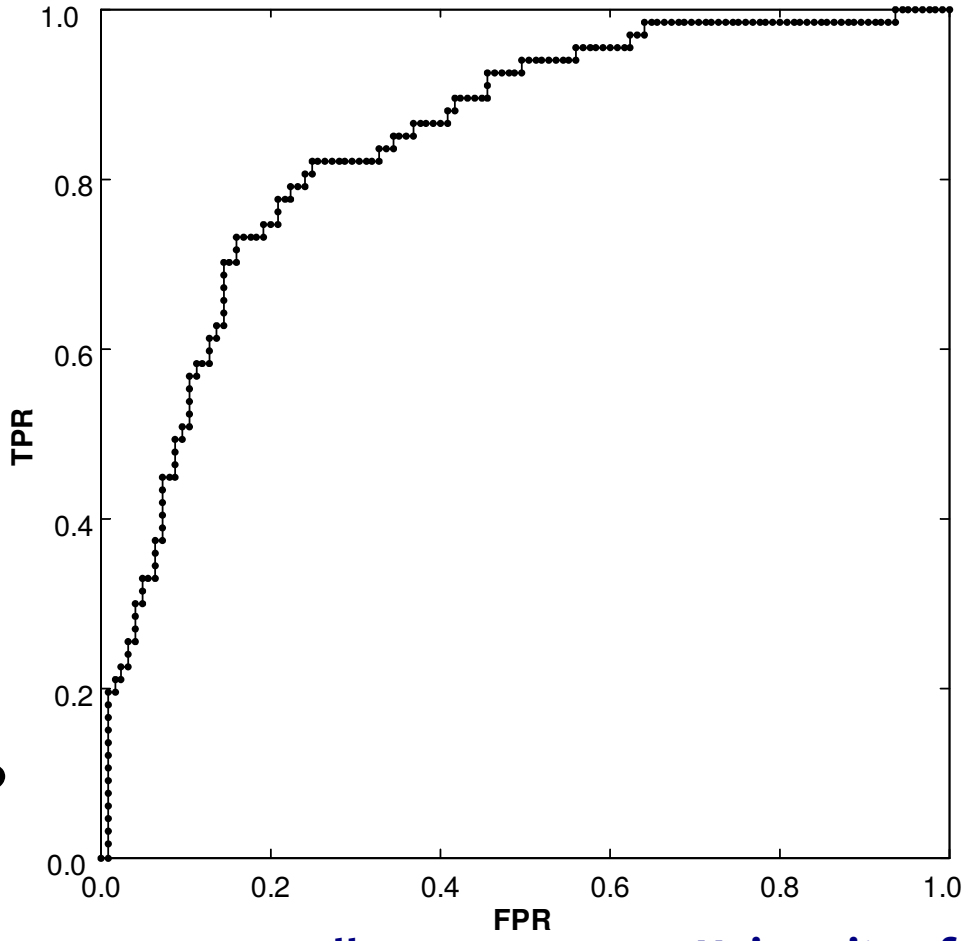
Require: $P > 0$ and $N > 0$

1: $L_{sorted} \leftarrow L$ sorted decreasing by f scores
2: $FP \leftarrow TP \leftarrow 0$
3: $R \leftarrow \langle\rangle$
4: $f_{prev} \leftarrow -\infty$
5: $i \leftarrow 1$
6: **while** $i \leq |L_{sorted}|$ **do**
7: **if** $f(i) \neq f_{prev}$ **then**
8: push $\left(\frac{FP}{N}, \frac{TP}{P}\right)$ onto R
9: $f_{prev} \leftarrow f(i)$
10: **end if**
11: **if** $L_{sorted}[i]$ is a positive example **then**
12: $TP \leftarrow TP + 1$
13: **else** /* i is a negative example */
14: $FP \leftarrow FP + 1$
15: **end if**
16: $i \leftarrow i + 1$
17: **end while**
18: push $\left(\frac{FP}{N}, \frac{TP}{P}\right)$ onto R /* This is (1,1) */
19: **end**

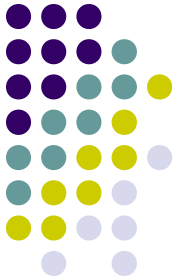
Generating the ROC curve for a classifier

Empirical ROC curve.
Estimated by applying the classifier to a finite set of samples.

Problems?



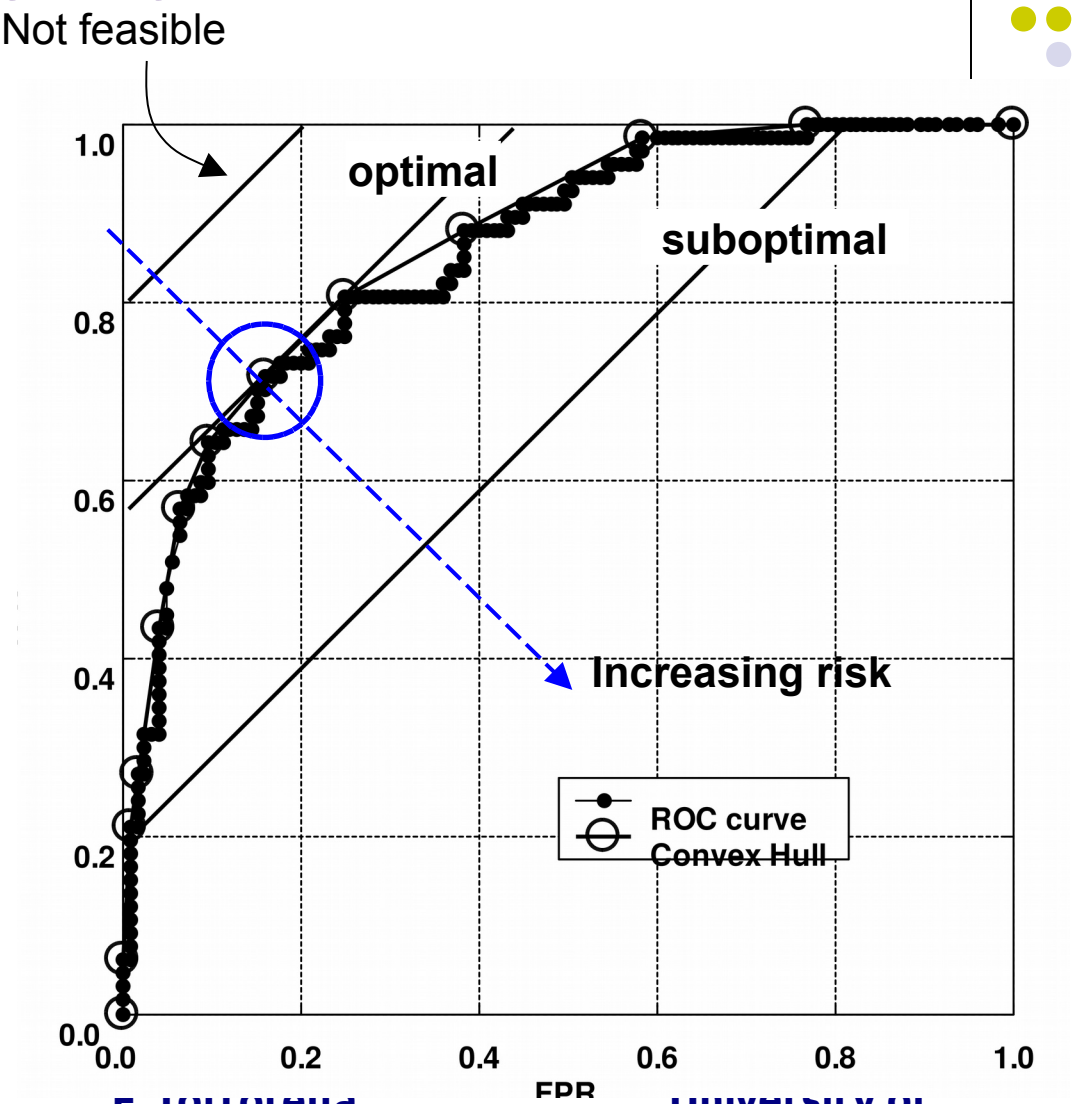
Choosing the optimal operating point



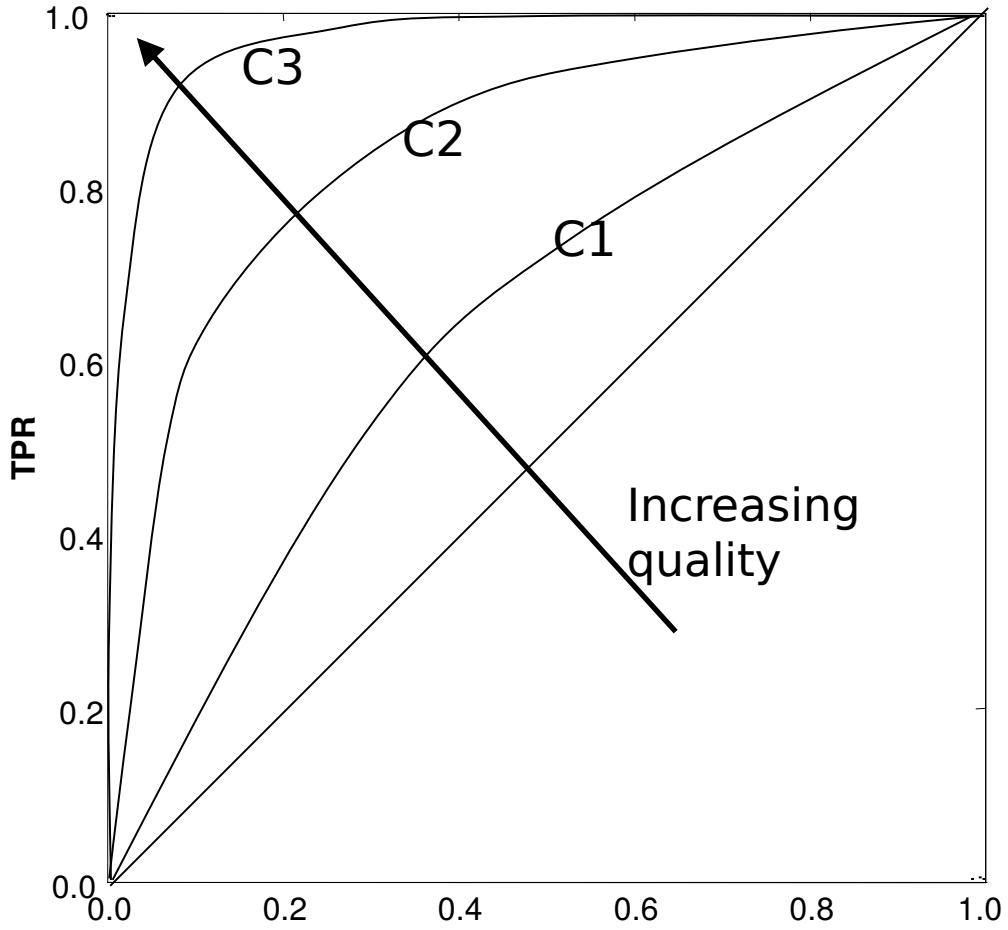
- Also in this case the combination of priors and costs defines a family of parallel isocost lines.
- Also in this case the optimal operating point must belong
  - to the ROC curve
  - to the “most north-west” line
- BUT in this case not all the points of the ROC curve can be chosen. The operating point belongs to the *convex hull* of the ROC curve.

Choosing the optimal operating point

Only the points belonging to the *convex hull* of the ROC curve can be touched by the isocost lines.



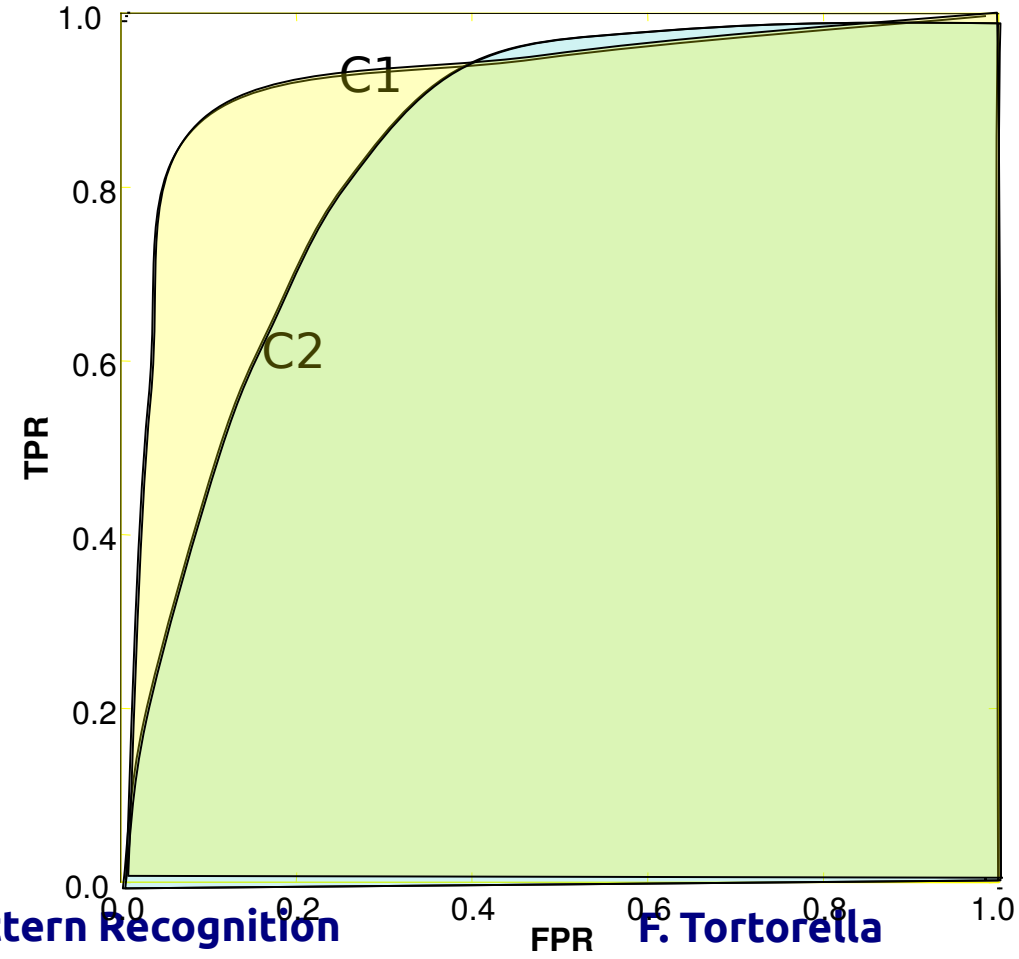
Comparing classifiers



We can say that C2 is definitely better than C1 if the ROC curve of C2 dominates the curve of C1

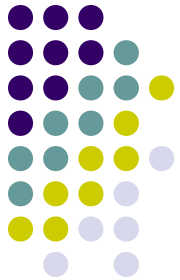
Es. C3 > C2 > C1

Comparing classifiers



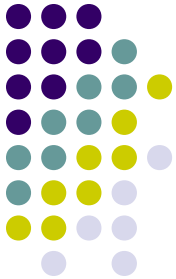
When the two curves intersect and no one clearly dominates, it is possible make a comparison in terms of the **Area under the ROC curve (AUC).**

AUC



- The AUC varies from 0.5 (random classifier) to 1.0 (ideal classifier).
- Its value is independent from the priors of the classes.
- The AUC equates the probability P(f(X)>f(Y)), where f(X) e f(Y) are the outputs of the classifier in correspondence of two samples X and Y randomly extracted from P and N, respectively.

AUC vs. accuracy (error rate)

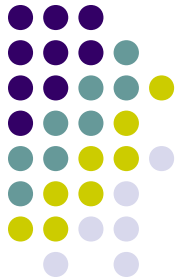


- **In many applications, the overall classification error rate is not the most pertinent performance measure, criteria such as *ordering* or *ranking* seem more appropriate. Consider for example the list of relevant documents returned by a search engine for a specific query.**
  That list may contain several thousand documents, but, in practice, only the top fifty or so are examined by the user. Thus, a search engine's ranking of the documents is more critical than the accuracy of its classification of all documents as relevant or not. More generally, for a binary classifier assigning a real-valued score to each object, a better correlation between output scores and the probability of correct classification is highly desirable.

C. Cortes* and M. Mohri, *AUC Optimization vs. Error Rate Minimization*, Advances in Neural Information Processing Systems (NIPS 2003)

(*) Google Labs

Estimating the AUC

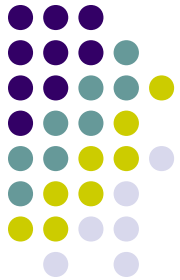


- According to the definition, the AUC can be estimated by numerically evaluating the integral of the empirical ROC curve.
- An alternative method can be considered if we recall that the AUC equals the probability P(f(X)>f(Y)). The Wilcoxon-Mann-Whitney statistics provides an estimate of such probability:

$$\frac{\sum_{i=1}^{P}\sum_{j=1}^{N} I(X_i, Y_j)}{N \cdot P} \qquad I(x,y) = \begin{cases} 1 & \text{if } x>y \\ 0.5 & \text{if } x=y \\ 0 & \text{if } x<y \end{cases}$$

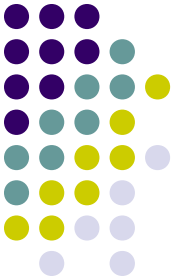
Wilcoxon-Mann-Whitney statistics

Neyman-Pearson decision rule



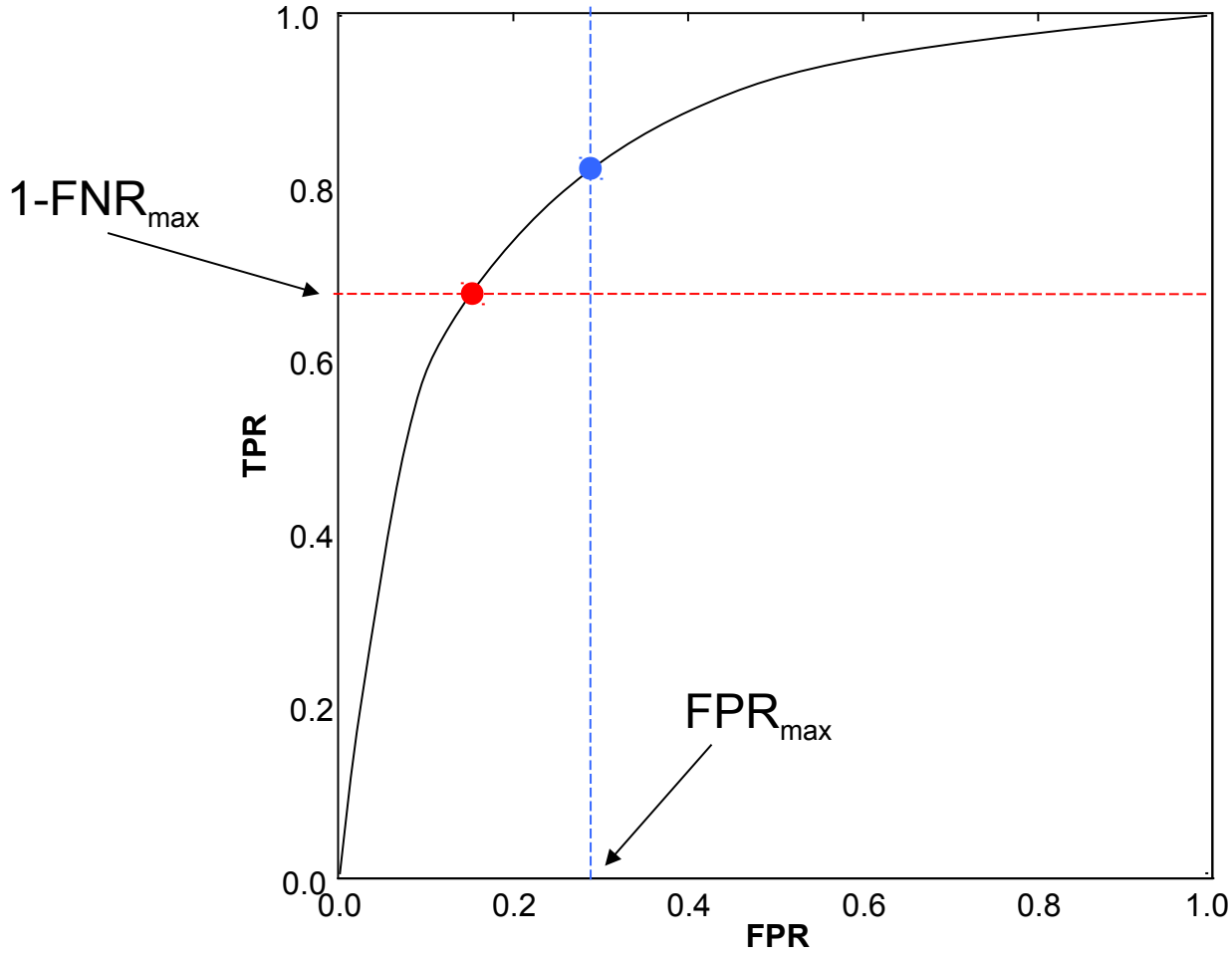
- Let us recall that in some cases, instead of minimizing an overall penalty (risk or error), we need to fix a bound on the error on one class while minimizing the error on the other class.
- For example we want the probability of error ε_2 on the class ω_2 is lower than α and that the probability of error ε_1 on the class ω_1 is minimum.
- This is the *Neyman-Pearson decision rule*

Neyman-Pearson decision rule on the ROC curve

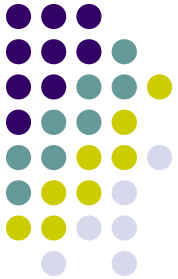


- In the ROC space the Neyman-Pearson decision rule specifies a maximum FPR (or FNR) that can be accepted.
- Thus the optimal operating point satisfying the NP criterion is readily è facilmente identified by the intersection between the ROC curve and the line $FPR=FPR_{max}$ ($FNR=FNR_{max}$).

Neyman-Pearson decision rule on the ROC curve



Partial AUC



- Also in this case we can consider the area under the curve limited by FPR_{max}.
- This is the *partial area under the ROC curve* (pAUC).
- It is frequently preferred as an index of diagnostic performance because the AUC summarizes the entire ROC curve, including regions that frequently are not relevant to practical applications (e.g., regions with low levels of specificity).
- The pAUC summarizes a portion the curve over the pre-specified range of interest.