CSI 436/536
Introduction to Machine Learning

Evaluating Classifiers

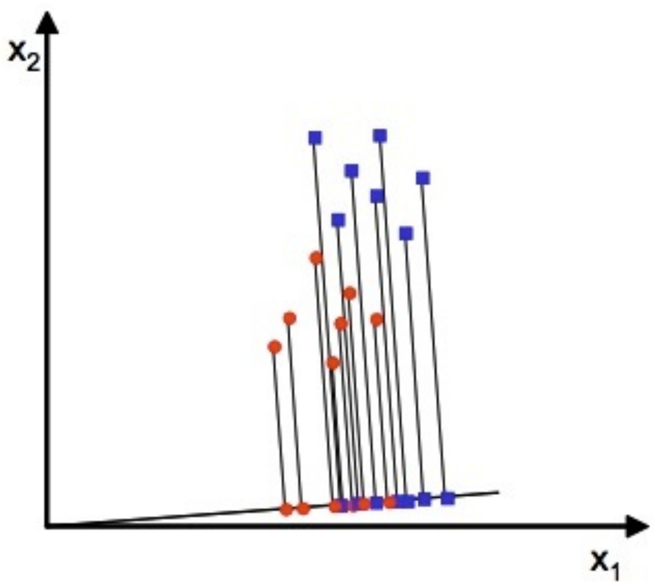
Professor Siwei Lyu
Computer Science
University at Albany, State University of New York

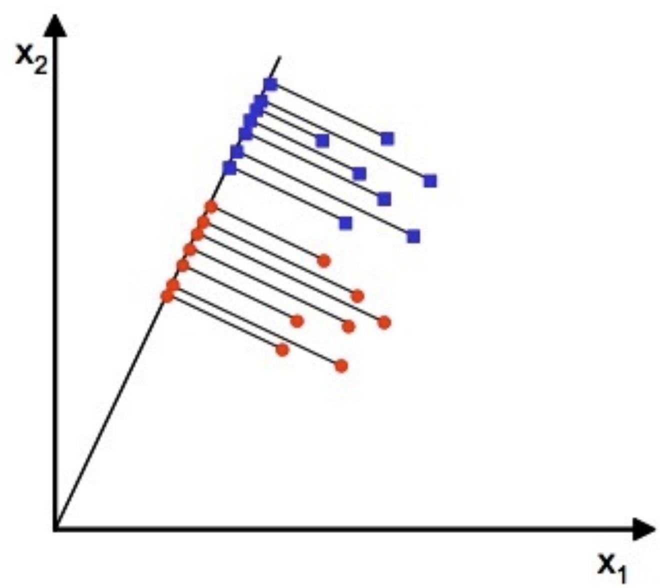
Classification problems

- Classification problem definition: mapping input data to discrete labels using a parametric function (classifier) $f_\theta(x) = y$, x: input data, θ: parameter, y: output label(s)
  - Single-label classification: output is a single label
    - Binary classification: $y \in \{0,1\}$ or $y \in \{-1,1\}$
    - Multi-class classification: $y \in \{1,\cdots,m\}, m > 2$
      - Note: multi-class classification problem can be solved with a series of binary classification problems
  - Multi-label classification: output is a set of labels $Y \subseteq \{1,\cdots,m\}$, i.e., each data can have multiple labels
    - Multi-label classification can be solved with a serious of binary predictors and ranking

Evaluating classifiers

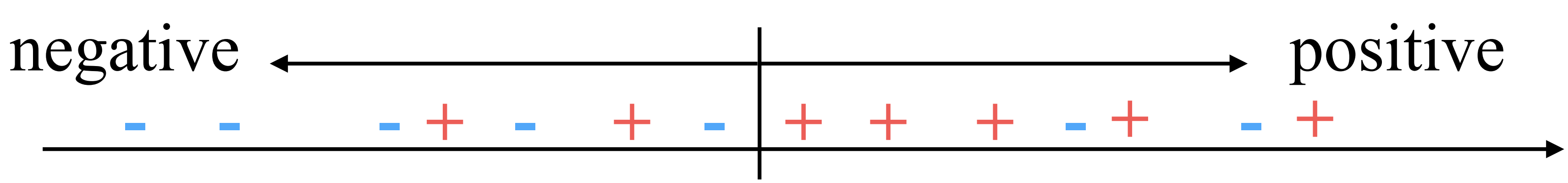
- Given two different classifiers, how do we evaluate and compare them?
  - Note that classification evaluation metric and training objective function (loss function) may not be the same
  - For instance, which of the two binary classifiers is a better one?





Definitions

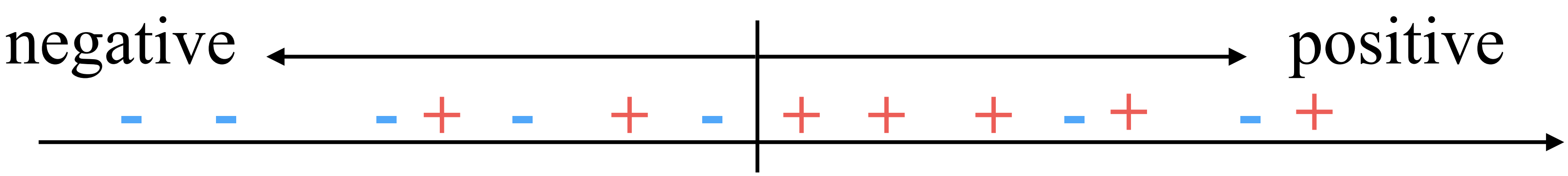
- A binary classifier with a threshold (positive on right)
  - actual = positive, classification = positive (TP)
    - 5 + examples > threshold
  - actual = negative, classification = negative (TN)
    - 5 - examples < threshold
  - actual = negative, classification = positive (FP)
    - 2 - examples > threshold
  - actual = positive, classification = negative (FN)
    - 2 + examples < threshold



Confusion matrix

- Confusion matrix

TP = 5	FN = 2	Actual positives = 7
FP = 2	TN = 5	Actual negatives = 7
Classified positives = 7	Classified negatives = 7	All examples = 14

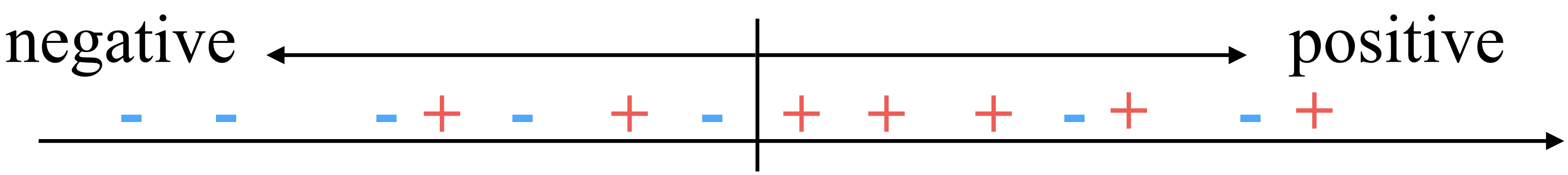


Confusion matrix

- Confusion matrix

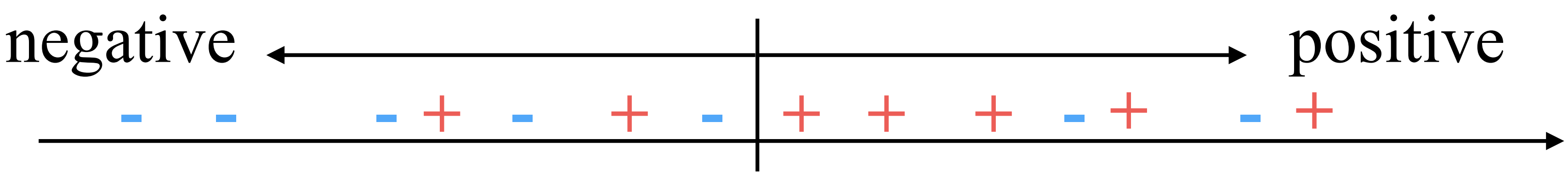
TP	FN	Actual positives
FP	TN	Actual negatives
Classified positives	Classified negatives	All examples

- Diagonal is correct classification
- Anti-diagonal is incorrect classification



Accuracy based metrics

- accuracy = (TP + TN)/(TP + TN + FP + FN) ≈ 71.4%
- error rate = (FP+FN)/(TP + TN + FP + FN) ≈ 28.6%
- true positive rate (TPR) = TP/(TP+FN) = 5/7 ≈ 71.4%
- false positive rate (FPR) = FP/(TN+FP) = 2/7 ≈ 28.6%
- precision (PR) = TP/(TP+FN) = TPR = 5/7 ≈ 71.4%
- recall (RC) = TP/(TP+FP) = 5/7 ≈ 71.4%
- $$\text{F1 score} = \frac{1}{\frac{1}{2}\left(\frac{1}{\text{TP}} + \frac{1}{\text{TN}}\right)} = \frac{2\text{TP} \times \text{TN}}{\text{TP} + \text{TN}} \approx 71.4\%$$



TPR-FPR vs. PR-RC

TPR

FPR

TP	FN	Actual positives
FP	TN	Actual negatives
Classified positives	Classified negatives	All examples

PR

RC

TP	FN	Actual positives
FP	TN	Actual negatives
Classified positives	Classified negatives	All examples

Why single rates are not good

- consider a data set with 10 positive and 90 negative examples, and two binary classifiers
  - Classifier 1

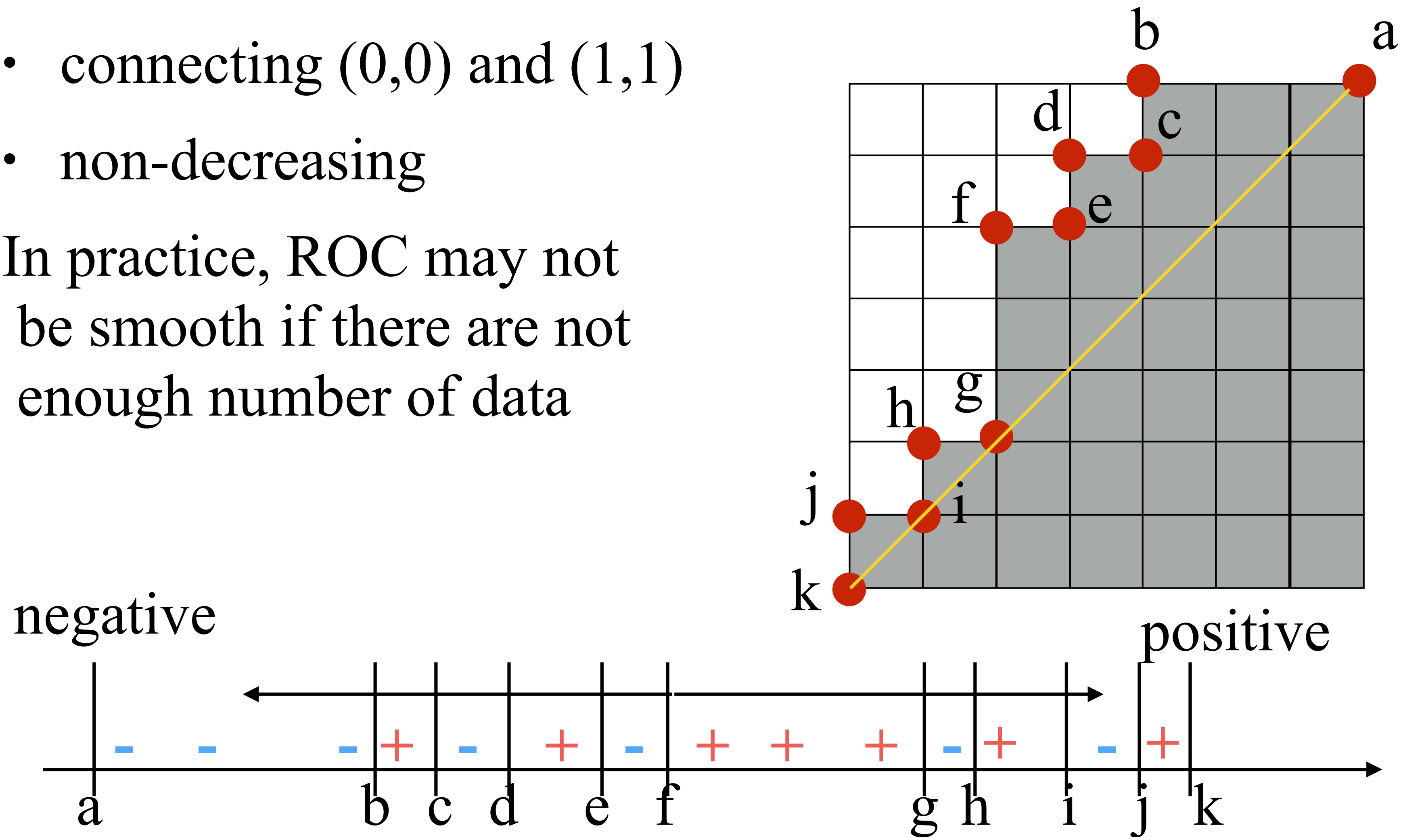
10	0	10
8	82	90
18	82	100

  - Classifier 2

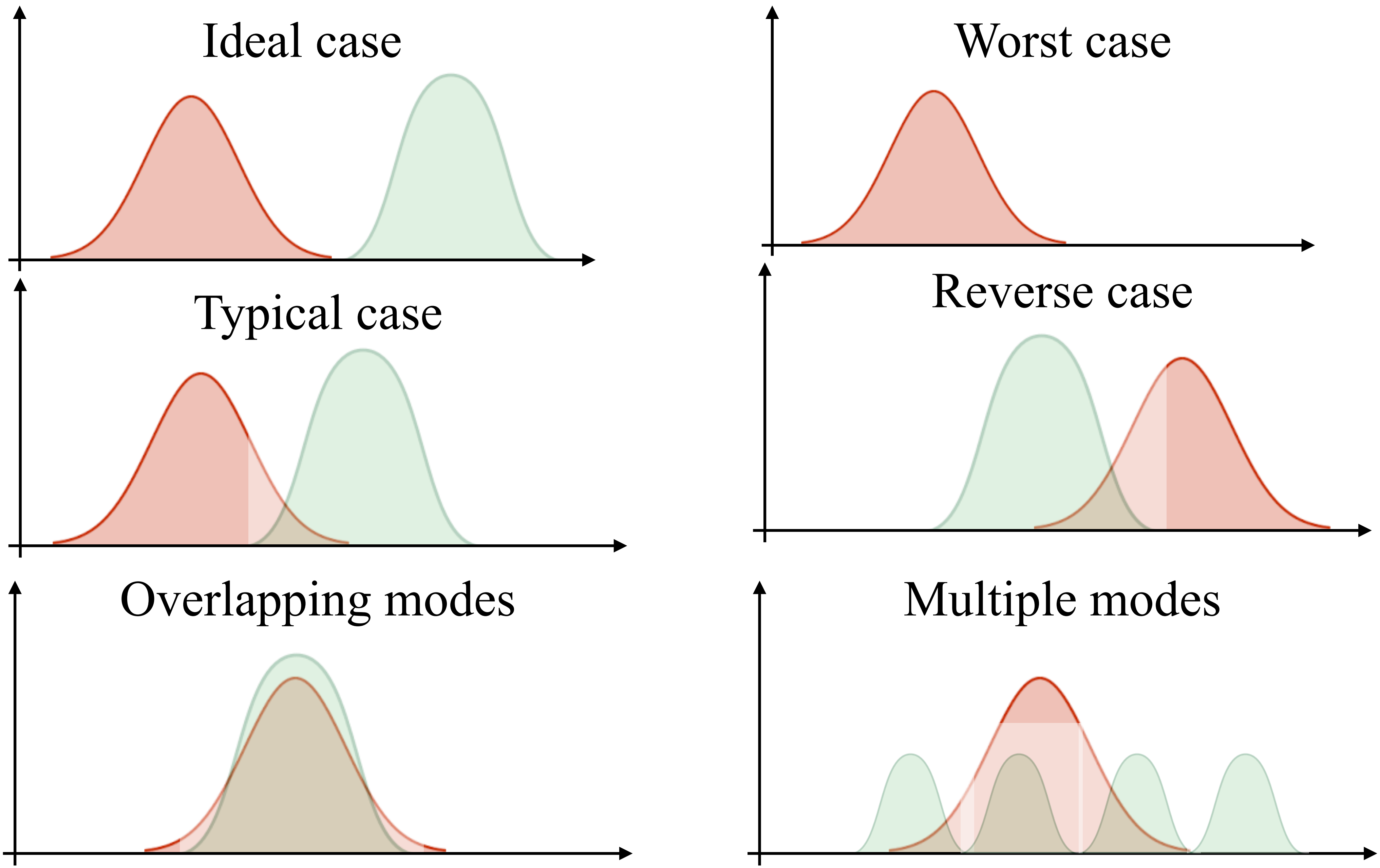
2	8	10
0	90	90
2	98	100

ROC curve

- Receiver Operator Characteristics (ROC) curve
  - tracing the curve of (FPR, TPR) with varying classification threshold
  - connecting (0,0) and (1,1)
  - non-decreasing
- In practice, ROC may not be smooth if there are not enough number of data



Reading ROC curves

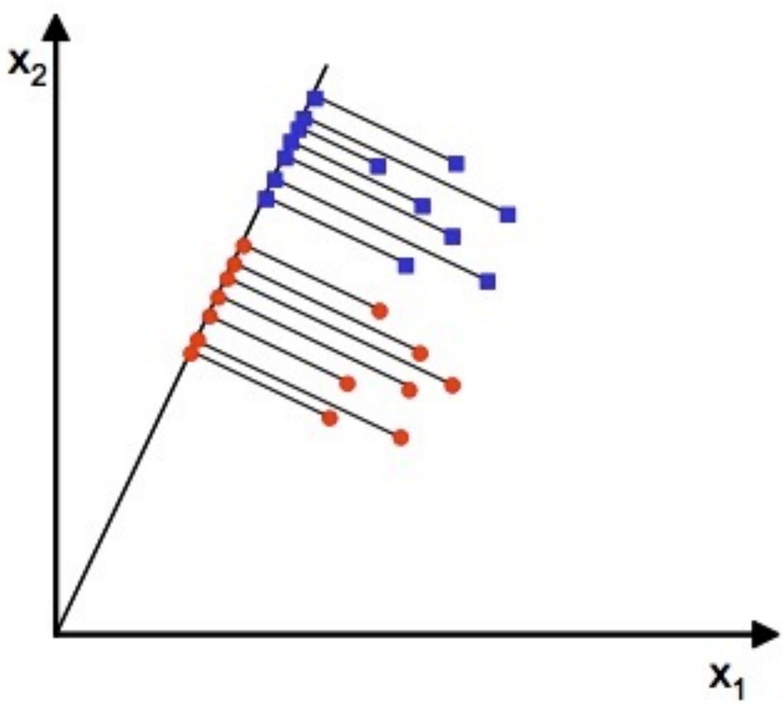


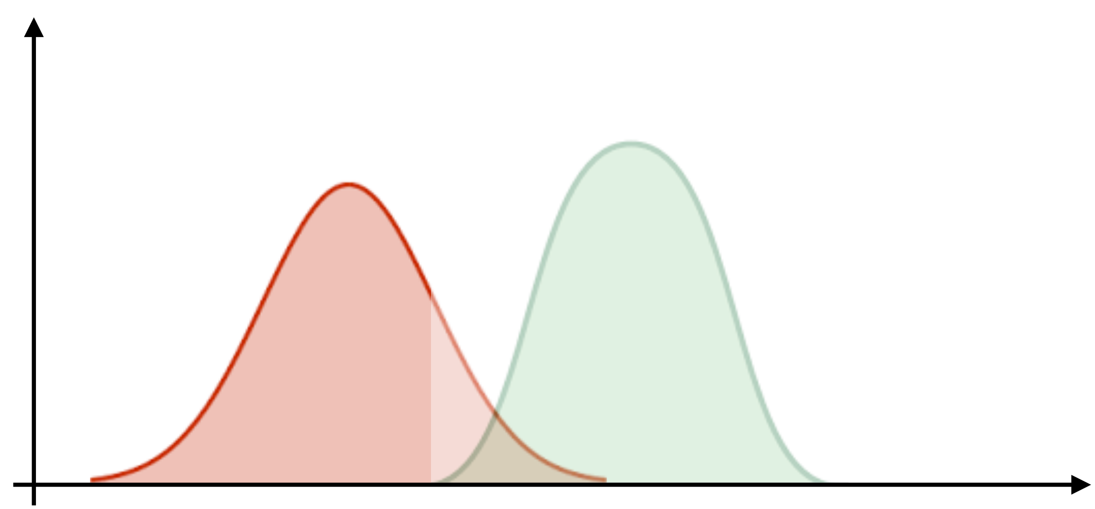
how to read ROC: summary

- ROC curve is trickier to understand
  - ideal ROC is line segment connecting (0,0) to (0,1) to (1,1)
  - worst ROC (random classifier) is the diagonal line connecting (0,0) and (1,1)
  - a good ROC is a convex curve connecting (0,0) and (1,1)
  - a ROC symmetric to the diagonal line can be obtained by flipping the class labels
  - support overlapping causes inflective ROC curve
  - multi-modal distribution causes multi-turn ROC curve

Choosing threshold

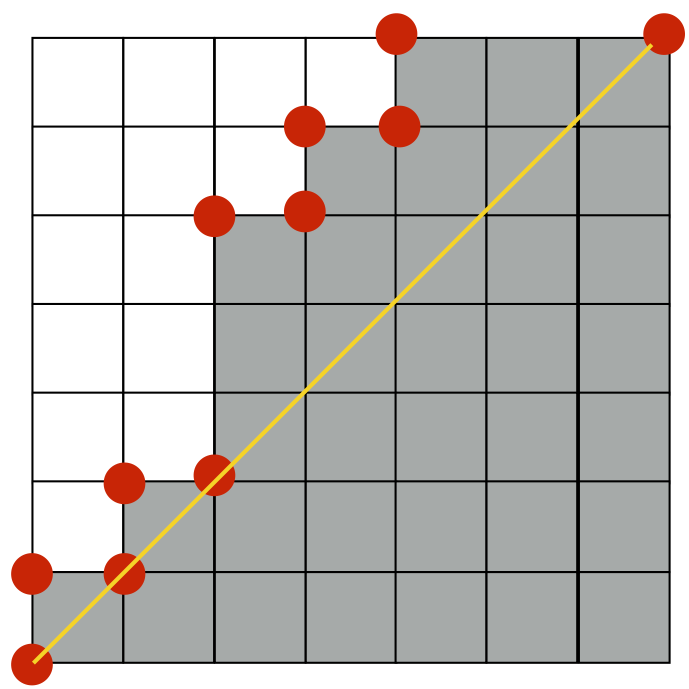
- using maximum classification separation rule we get the intersection of the two distributions
  - The point with the minimum overall classification error
  - The point on ROC with (1,1) derivative

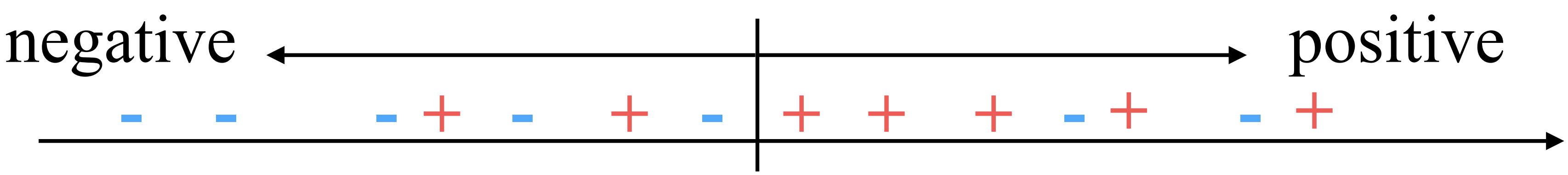




Area under ROC curve (AUR)

- AUC ∈ [0.5,1.0], 0.5 is the random binary classifier, 1.0 is the perfect binary classifier
  - In this case, AUC is 35/49 ≈ 71.4%





Computing AUC on finite data

- AUC on finite data set can be computed using the Mann-Whitney-Wilcoxon (MWW) statistics

$$\frac{1}{n_+ n_-} \sum_{i=1}^{n_+} \sum_{j=1}^{n_-} 1_{x_i^- \le x_j^+}$$

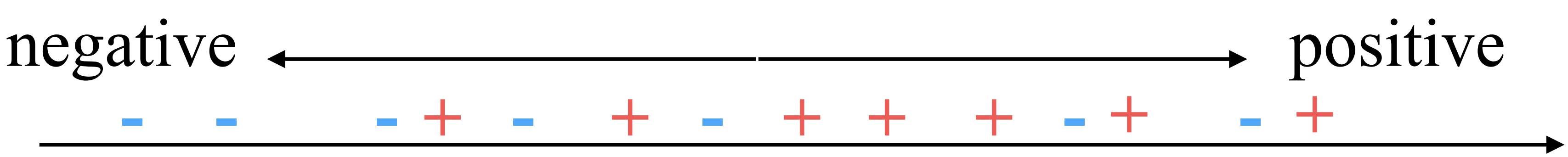
  - Proof:

    $\Pr(X>Y) = \iint_{x>y} f(x)g(y)dxdy = \iint \delta(x>y)f(x)g(y)dxdy$

    $= \int f(x)dx \int \delta(x>y)g(y)dy = \int G(x)\, f(x)dx$

    $= \int G(x)\, dF(x) = \int \mathrm{ROC}(F)dF = \mathrm{AUC}$

  - MWW statistics is the fraction of pairs with correct orders $\dfrac{7+7+7+6+5+2+1}{7\times 7} \approx 71.4\,\%$



Multi-class classifier

- confusion matrix
  - Diagonal: correct examples, Off-diagonal: errors
  - Accuracy: sum of diagonal dividing the total sum

	A	B	C	D
A	10	2	3	0
B	4	12	1	5
C	3	0	7	2
D	1	3	0	5