CIS501 – Lecture 8

Woon Wei Lee
Fall 2013, 10:00pm – 11:15pm,
Sundays and Wednesdays

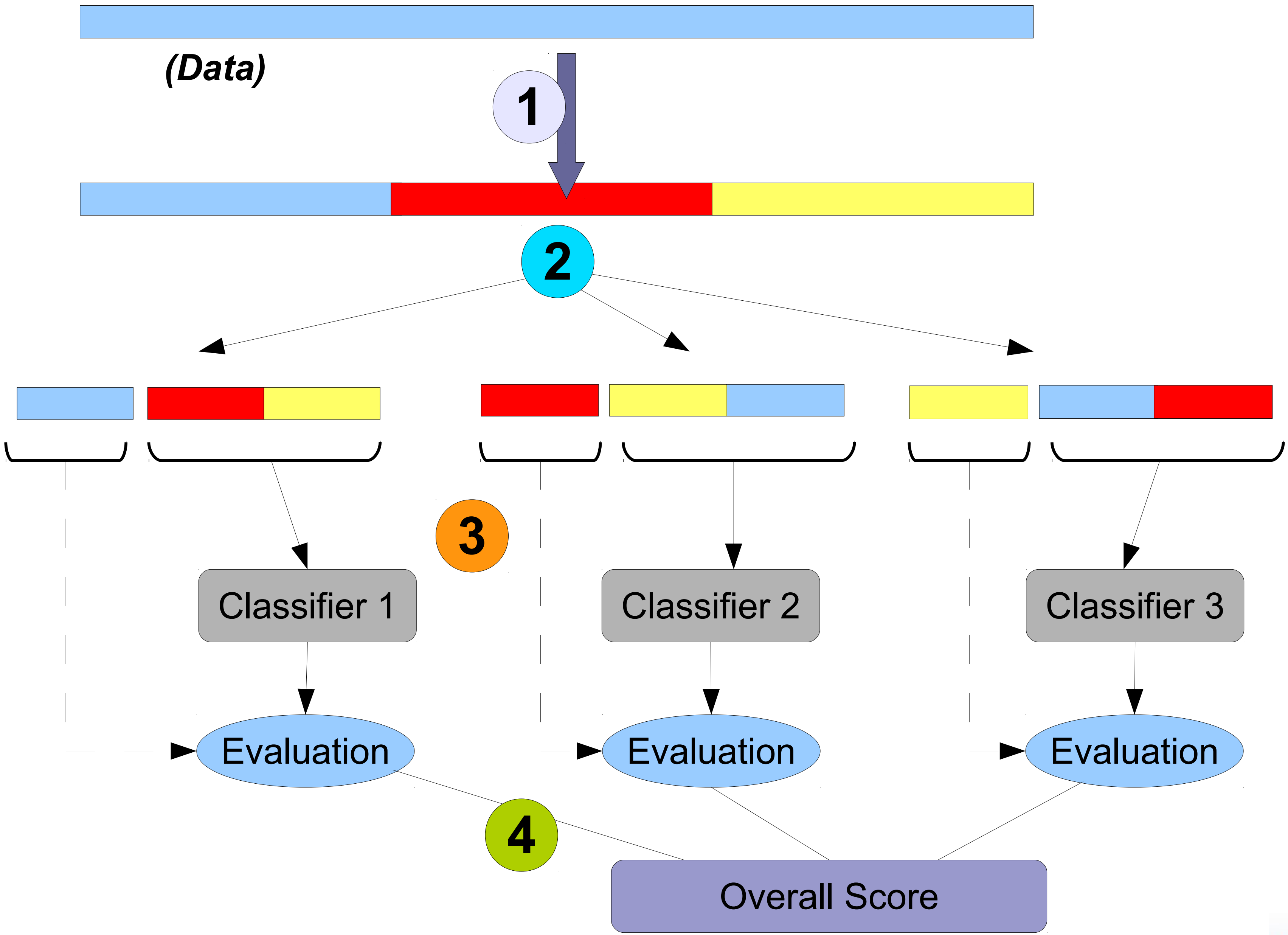
For today:

- Administrative stuff
  - Scheduling arrangements (inc. re midterm quiz)
- Evaluating classifiers
  - Numerical performance indices $\surd$
  - Cross validation
  - ROC (and related) curves

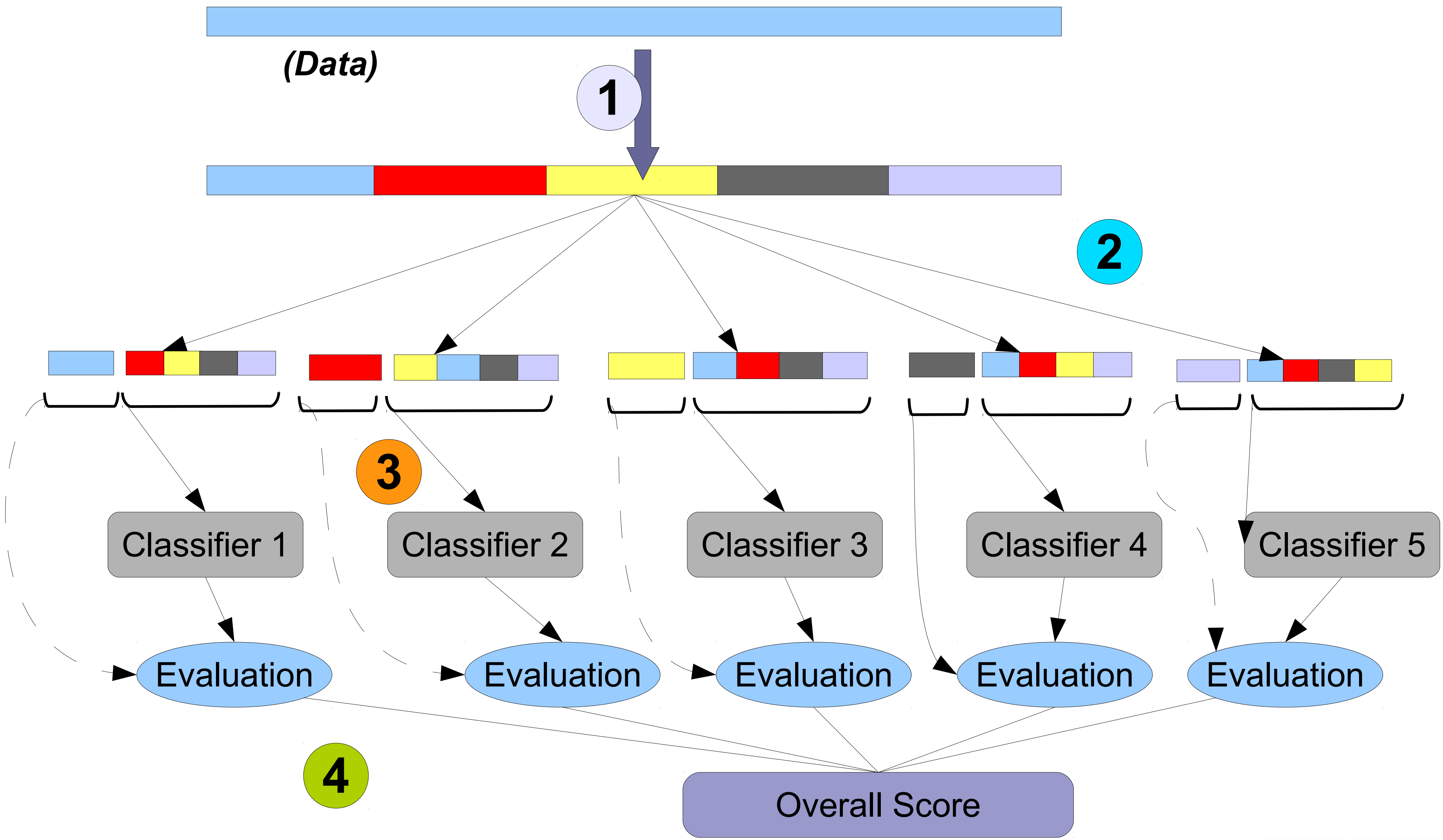
Cross validation

- **Numerical performance metrics good start, but..**
  - Only test the performance of a particular classifier configuration vs. a particular data set
  - Encourages *overfitting*; i.e. reduces generalization capability.
- **Key requirement: train and test on different data**
  - In principle, we could just divide data into separate training and test data sets.
  - However, in practice, data is valuable → dividing into separate sets is a waste!
- **The solution: Cross validation**
  - Rotate between test and training data sets.
  - Allows independent tests without reducing the amount of data that is available.
  - Provide good estimate of the true accuracy of a classifier.

Cross Validation (3-Fold)



Cross Validation (5-fold)



(Cont'd)

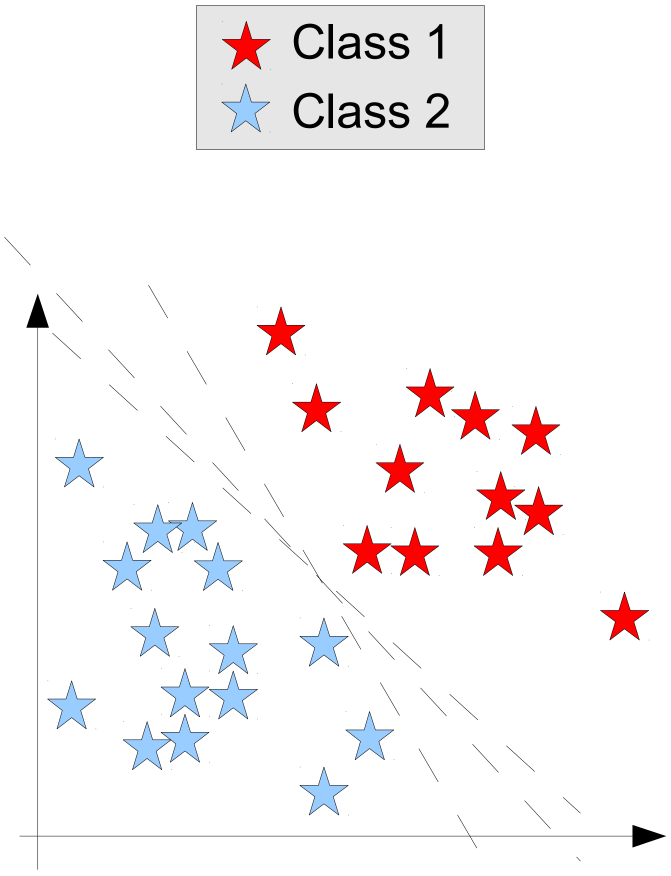
- **In general, the larger "*n*" is the better**
  - Diminishing returns from larger *n*
  - The case where *n* is the size of the data set is known as "leave-one-out" cross validation, AKA "Jack-Knifing"
- **But even when selecting a classifier as shown, there is still a bias**:
  - The reported accuracy value would tend to be better than the "true" performance of the classifier
  - Proper evaluation of the classifiers requires a third set of data, known as the "validation" data.
  - Performance of selected classifier on the validation data would be the one that is reported.
- **A further enhancement to the basic cross validation procedure is the use of "stratified sampling"**

ROC curve

- In general, accuracy measures only describe the performance of a classifier at a particular threshold value.
- They do not give a very good representation of the overall “quality” of a classifier.
- See for example, the plot on the right

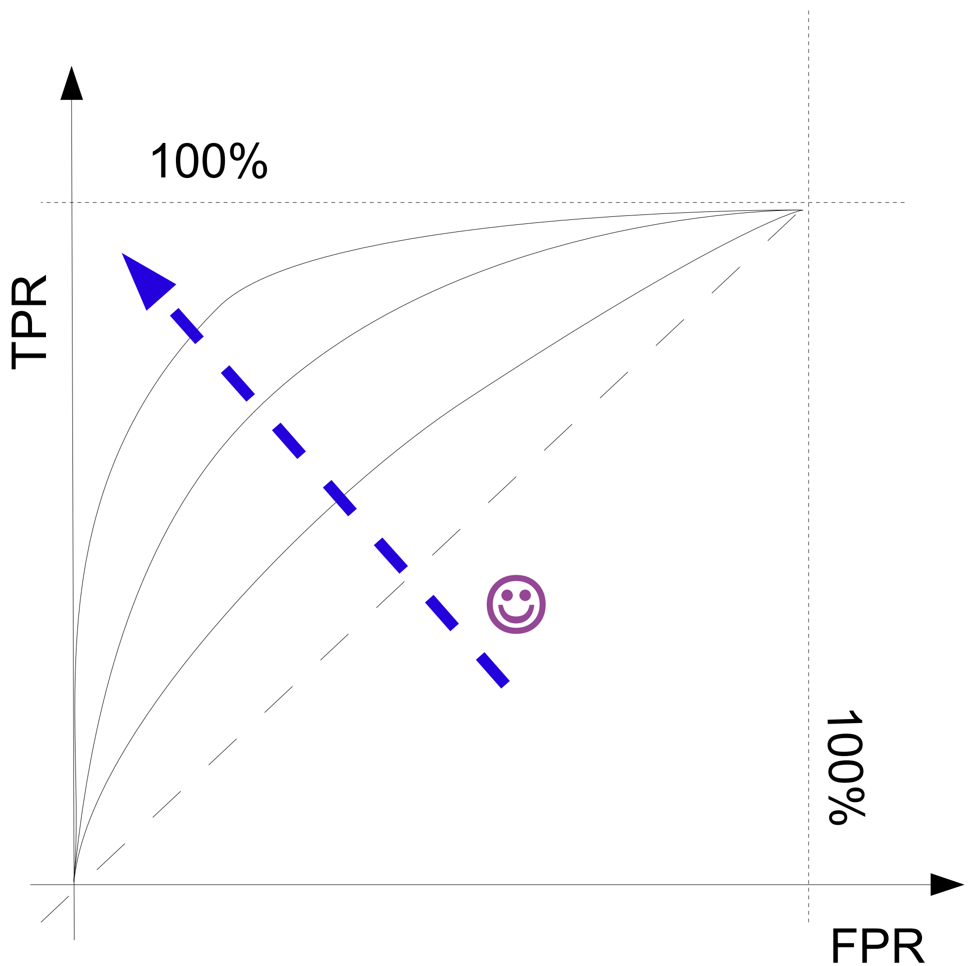
  → each of the dotted lines correspond with one classifier

  - All “100%” accurate, but clearly there is a difference in the quality of the classification and generalizability
- ROC curves provide an alternative way of evaluating classifier “quality”



Masdar INSTITUTE

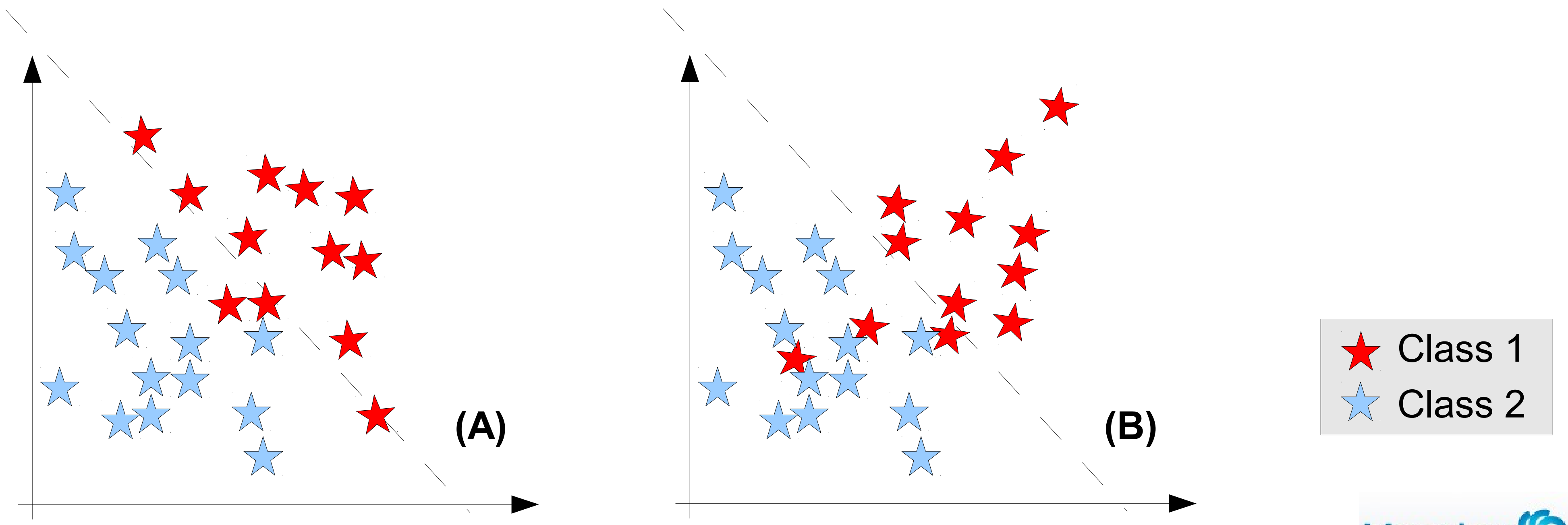
ROC curve (Cont'd)



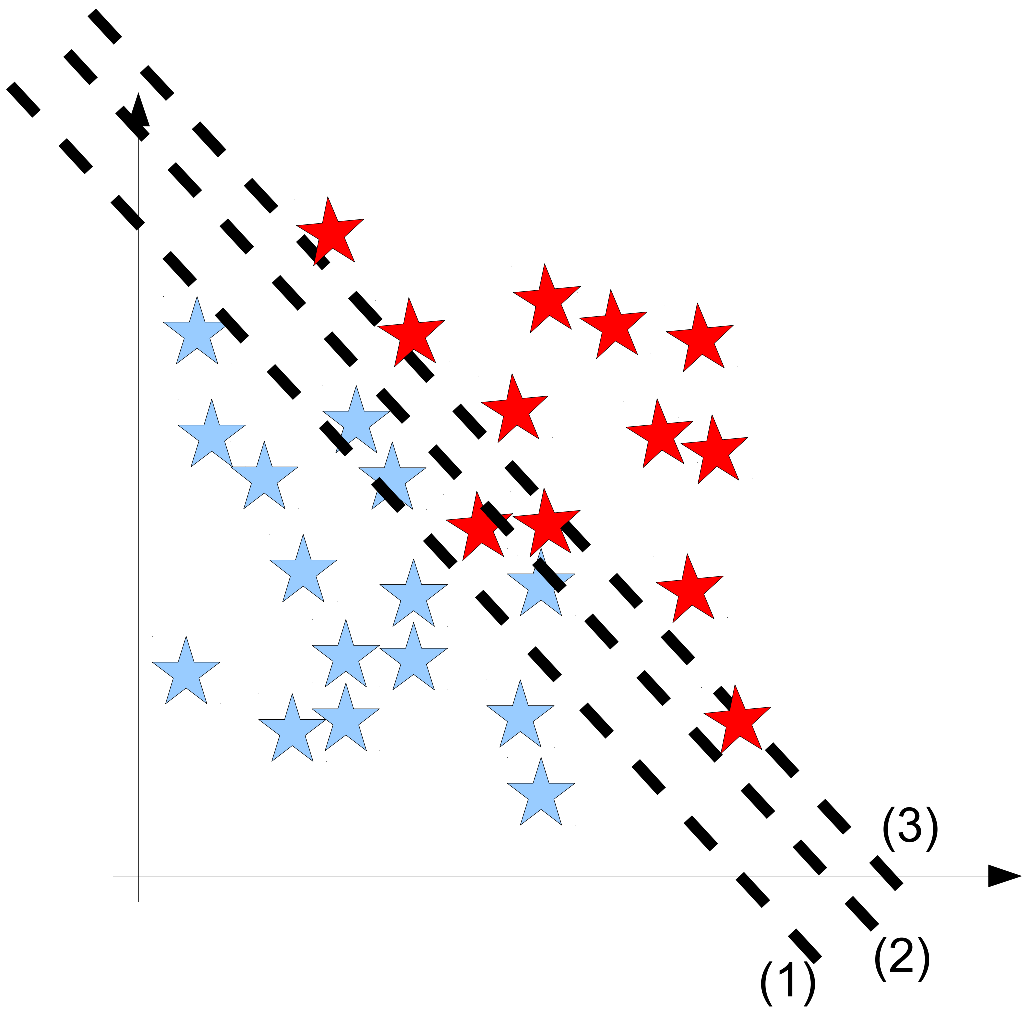
- Stands for “Receiver Operator Characteristic” curve
- Origins in signal detection
- Presents the trade-off between *Precision* (affected by false positive rate) and *Recall* (affected by true positive rate)
- In general, we want:
  - High true positive rates (TPR)
  - Low false positive rates (FPR)
- Two are directly antagonistic:
  - Trivial to have 100% TPR by always returning “1”
  - Similarly for FPR
- **Question: What is the diagonal line?**

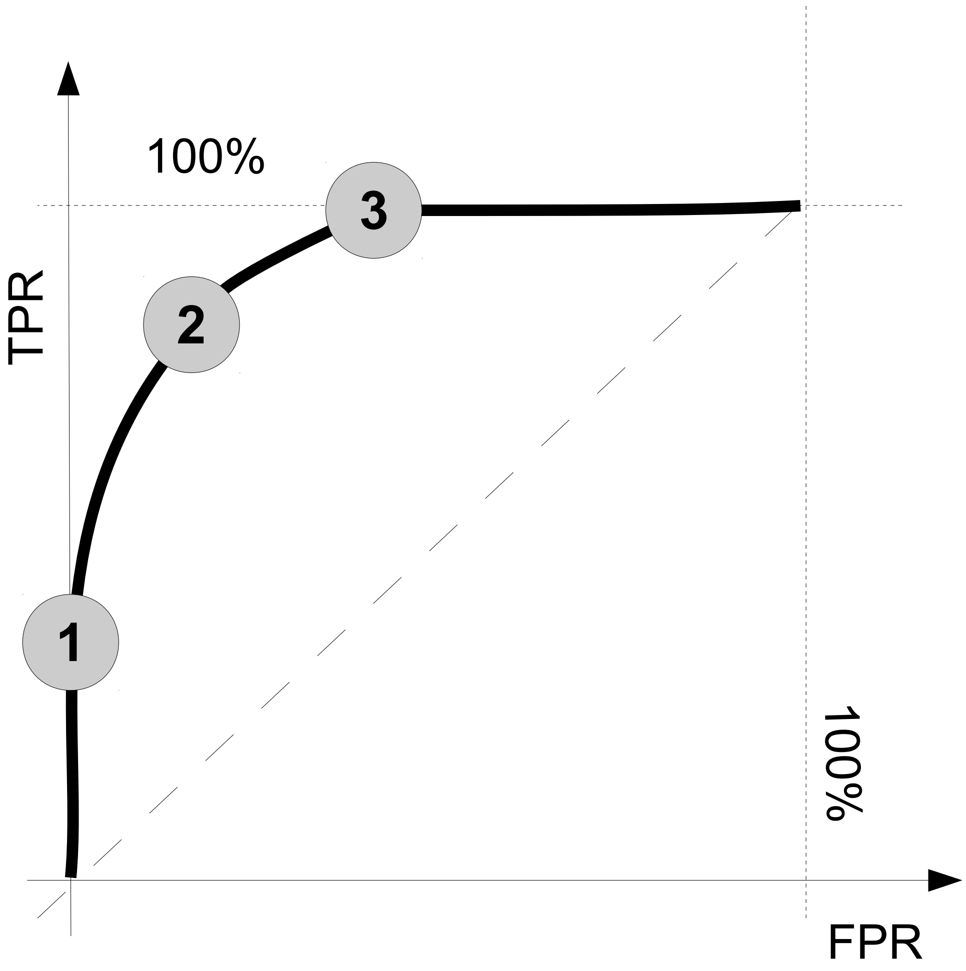
ROC curve (Cont'd)

- A simplified (but demonstrative) example:
- Two separate instances, (A) and (B)
  - In both cases, we have 3 misclassified cases
  - However, there is clearly a difference between cases (A) and (B)
- We can say that the classifier built for case (B) is somehow “better”

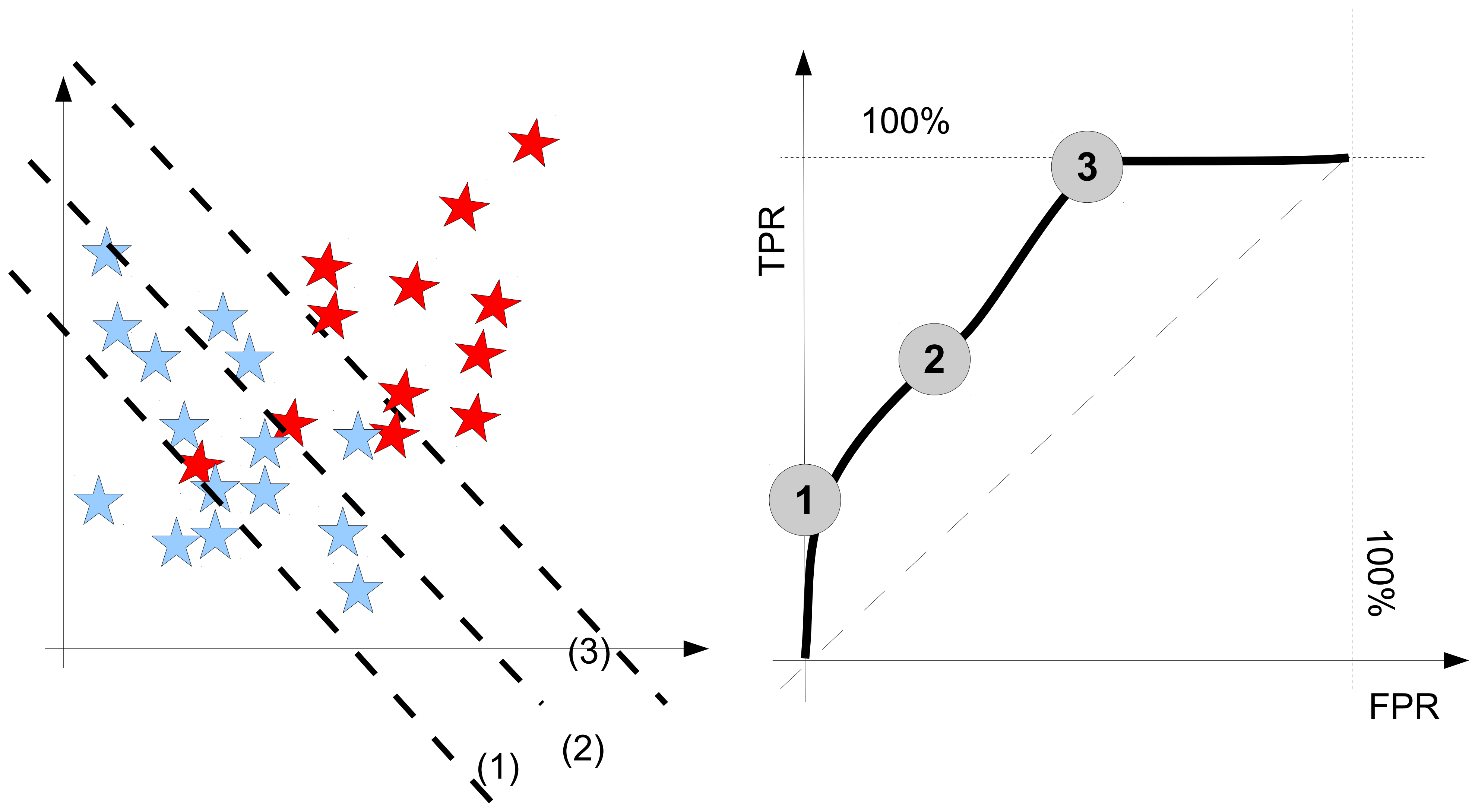


Case A

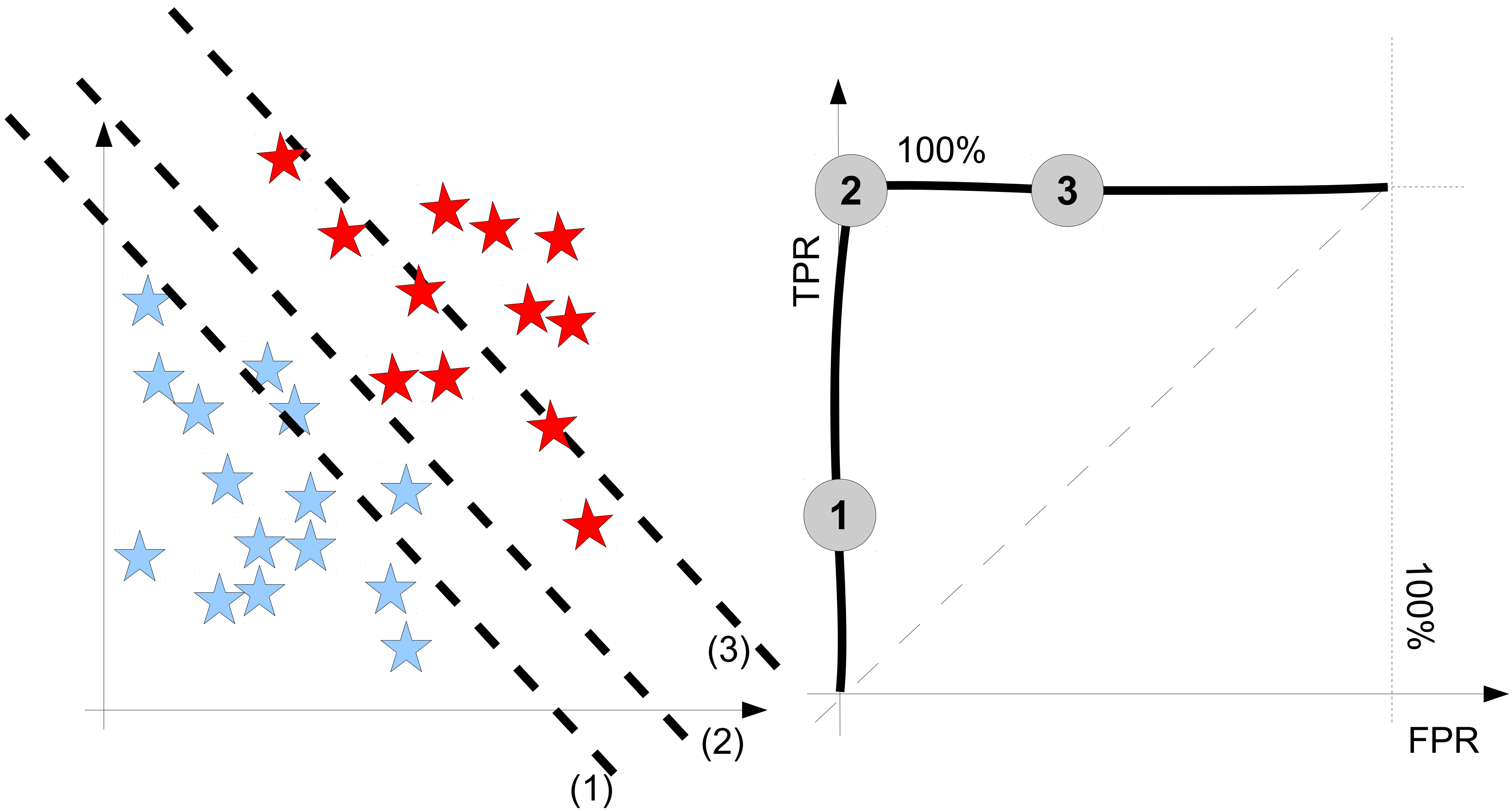




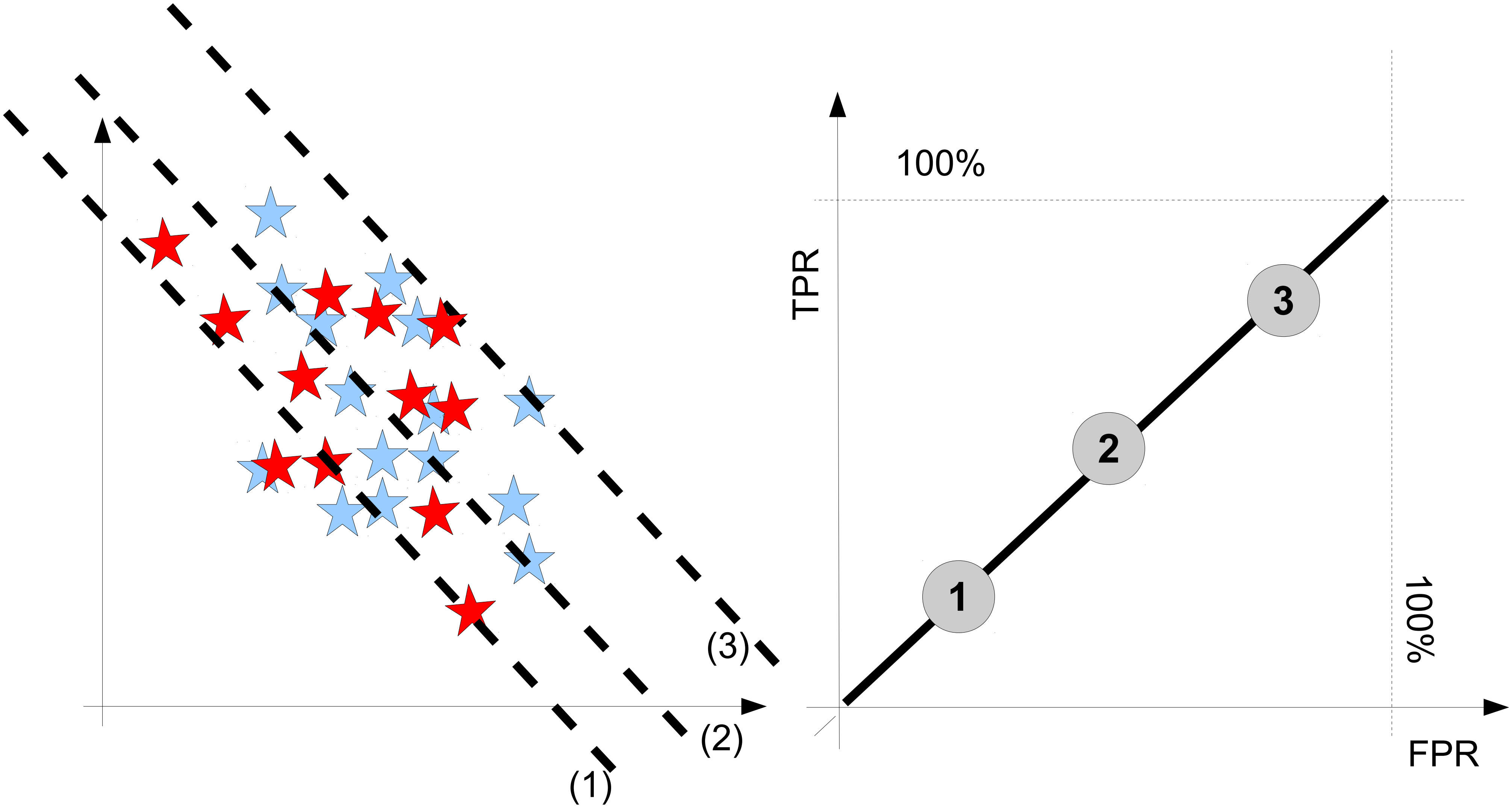
Case B



“Perfect” curve



Worse possible curve



Question: Can the curve dip below the diagonal line?

Cont'd

- Evaluation of ROC curve is often done in terms of the
  - “AUC” - **A**rea **U**nder the **C**urve
  - The distance between the no-discrimination line and the intercept of the curve and the line perpendicular to the no-discrimination line
- “Grading” can be done in a number of ways but a simple system would be along the lines of:

  >90% *excellent* (A)

  80%-90% *good* (B)

  70%-80% *fair* (C)

  60%-70% *poor* (D)

  <60% *fail* (F)

- Statistically, the AUC is the probability that a randomly chosen positive point is scored higher than a randomly chosen negative one

Lift charts

- Alternative graph with identical information content

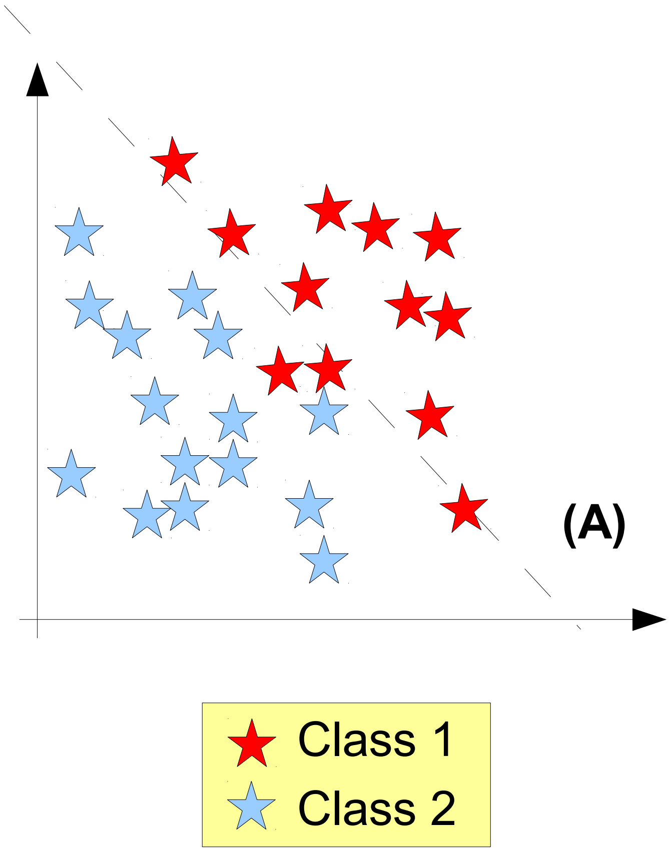
- Procedure is as follows:

  1. For all examples in training set, tabulate the posterior distributions for (say) “Class 1” vs actual label of the example

  2. Sort the rows of the table with respect to $P(C_1|x)$

  3. Calculate the cumulative values for the actual label column.

  4. Plot cumulative value vs. instance #.



Predicted Prob. of Success	Actual Value of HICLASS
0.9734	1
0.0015	0
0.6002	0
0.0000	0
0.9893	1
0.2156	0
0.0000	0
0.2468	0
0.0130	0
0.0000	0
0.0000	0
0.0000	[illegible]
0.0000	0
0.9884	1
0.9715	1
0.9744	1
0.0641	0
0.4900	0
0.0000	0
0.0000	0
0.0000	0
0.9999	1
0.5218	0

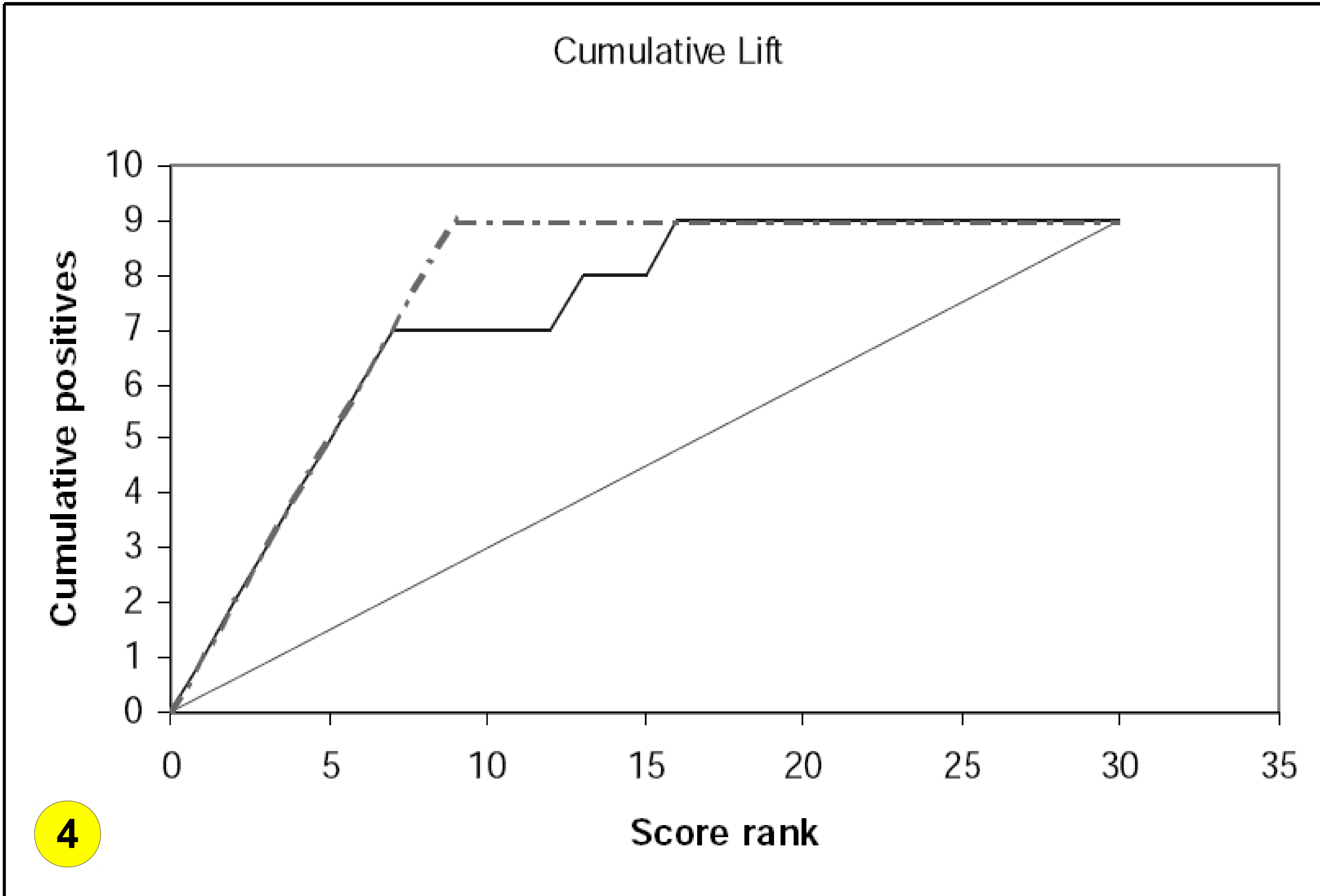
1

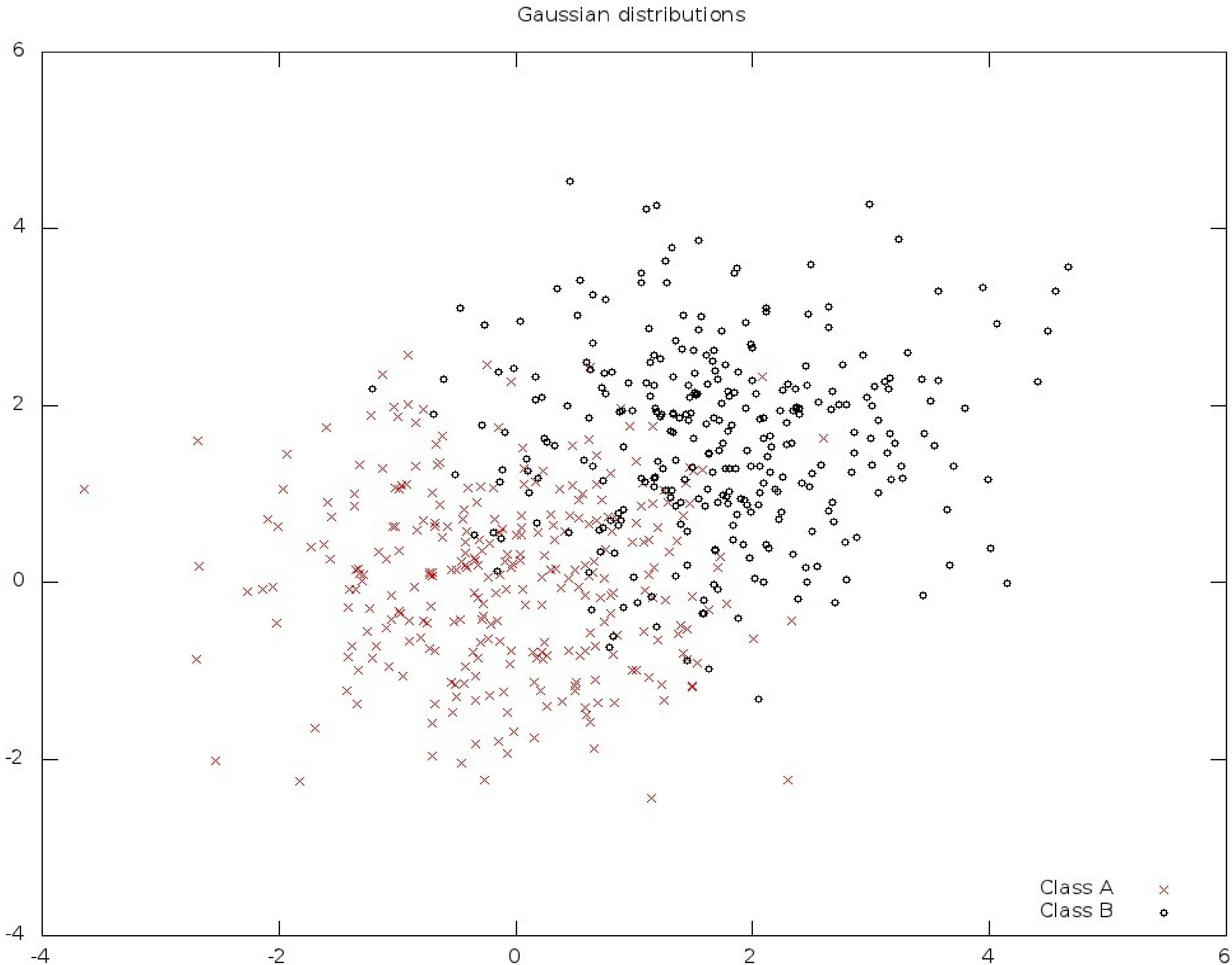
Predicted Prob. of Success	Actual Value of HICLASS
0.9999	1
0.9893	1
0.9884	1
0.9744	1
0.9734	1
0.9715	1
0.8489	1
0.6002	0
0.5218	0
0.4900	0
0.2468	0
0.2156	[illegible]
0.1281	1
0.0641	0
0.0130	0
0.0023	1
0.0015	0
0.0001	0
0.0000	0
0.0000	0
0.0000	0
0.0000	0
0.0000	0

2

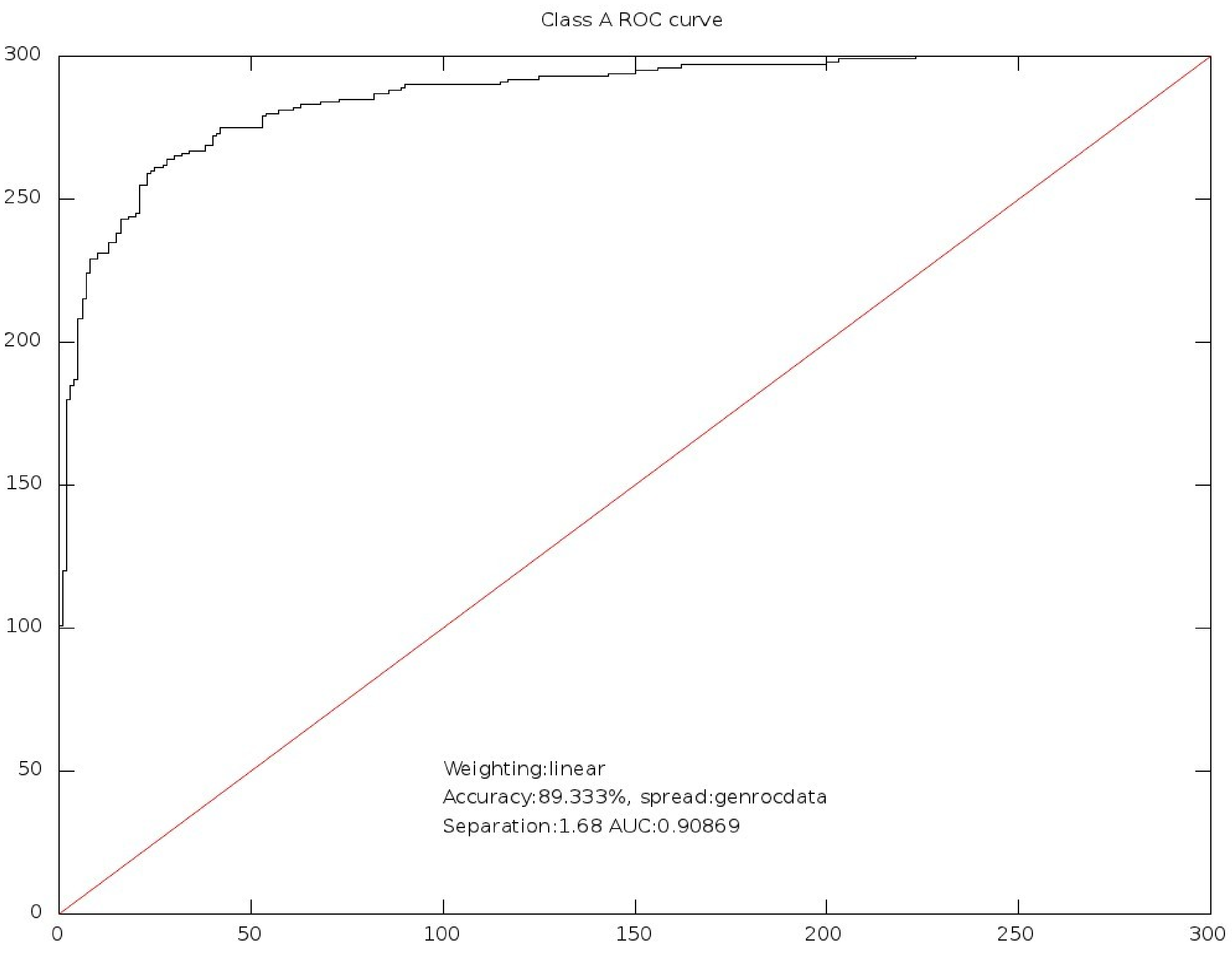
Predicted Prob. of Success	Actual Value of HICLASS	cumulative Actual Value
0.9999	1	1
0.9893	1	2
0.9884	1	3
0.9744	1	4
0.9734	1	5
0.9715	1	6
0.8489	1	7
0.6002	0	7
0.5218	0	7
0.4900	0	7
0.2468	0	7
0.2156	0	7
0.1281	1	8
0.0641	0	8
0.0130	0	8
0.0023	1	9
0.0015	0	9
0.0001	0	9
0.0000	0	9
0.0000	[illegible]	9
0.0000	0	9
0.0000	0	9

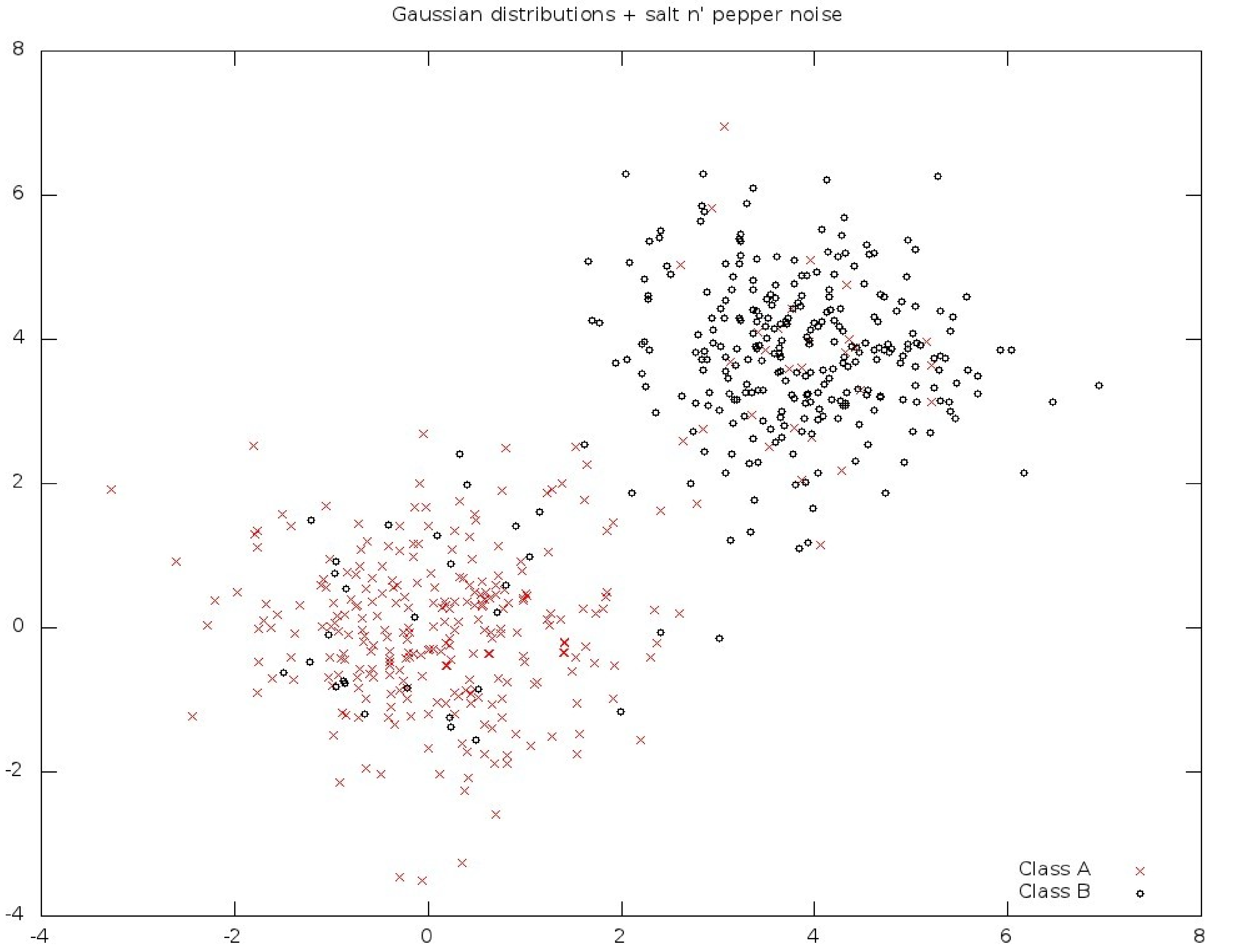
3

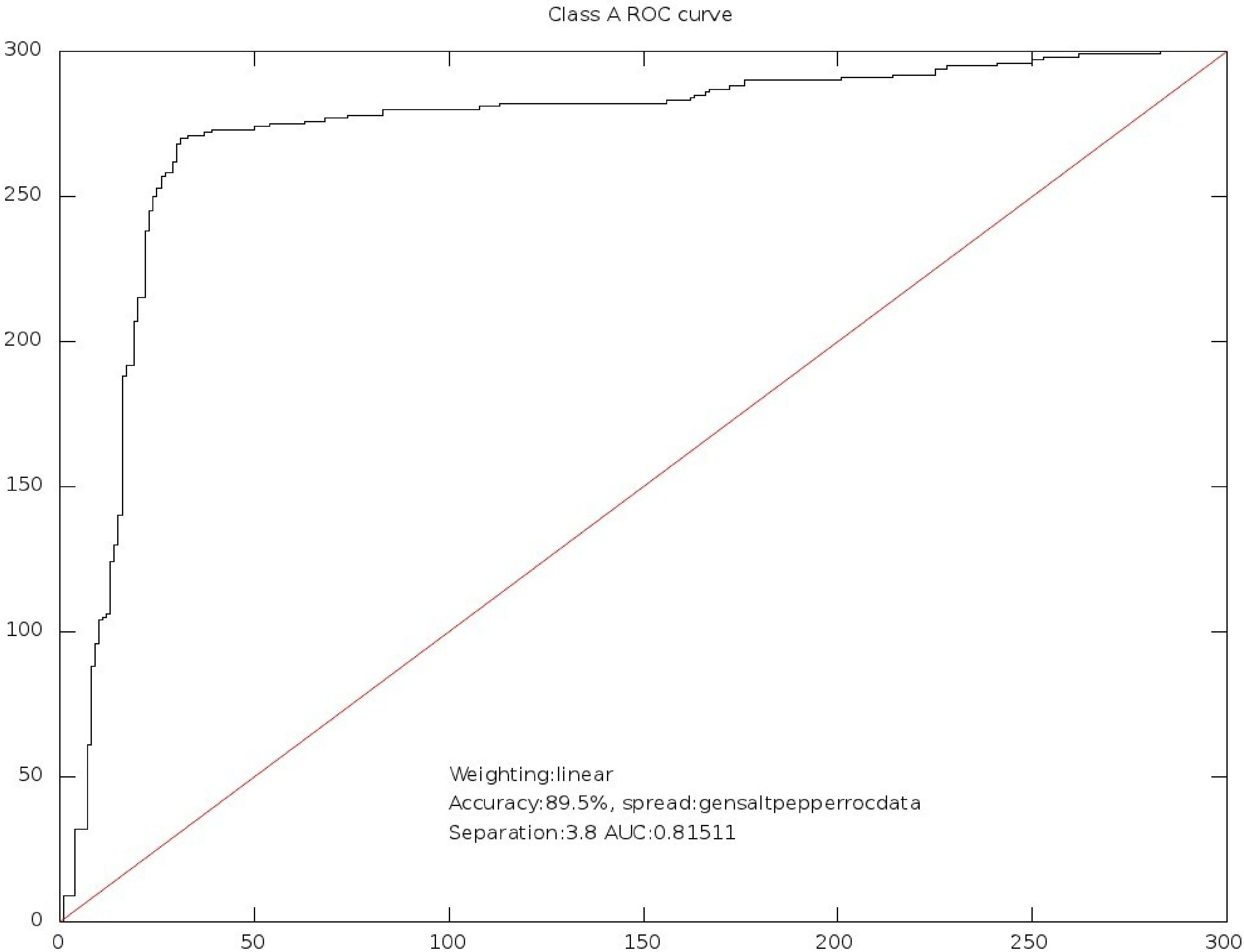

Cumulative Lift
Cumulative positives
10
9
8
7
6
5
4
3
2
1
0
0
5
10
15
20
25
30
35
Score rank
4



Gaussian distributions
6
4
2
0
-2
-4
-4
-2
0
2
4
6
Class A
Class B


Class A ROC curve
300
250
200
150
100
50
0
0
50
100
150
200
250
300
Weighting:linear
Accuracy:89.333%, spread:genrocdata
Separation:1.68 AUC:0.90869


Gaussian distributions + salt n' pepper noise
8
6
4
2
0
-2
-4
-4
-2
0
2
4
6
8
Class A
Class B


Class A ROC curve
Weighting:linear
Accuracy:89.5%, spread:gensaltpepperrocdata
Separation:3.8 AUC:0.81511
0
50
100
150
200
250
300