Evaluation & Credibility Issues

- What measure should we use?
  - Classification accuracy might not be enough.
- How reliable are the predicted results?
- How much should we believe in what was learned?
  - Error on the training data is not a good indicator of performance on future data.
  - The classifier was computed from the very same training data, any estimate based on that data will be optimistic.

Evaluation Questions

- How to evaluate the performance of a model?
- How to obtain reliable estimates of performance?
- How to compare the relative performance among competing models?
- Given two equally performing models, which one should we prefer?

Metrics for Performance Evaluation

- Focus on the predictive capability of a model.
- Confusion matrix:

		Predicted Class	
		+	-
Actual Class	+	f_{++} (TP)	f_{+-} (FN)
	-	f_{-+} (FP)	f_{--} (TN)

Accuracy

		Predicted Class	
		+	-
Actual Class	+	f_{++} (TP)	f_{+-} (FN)
	-	f_{-+} (FP)	f_{--} (TN)

The most widely-used metric is accuracy:

$$Accuracy = \frac{a+d}{a+b+c+d} = \frac{TP+TN}{TP+TN+FP+FN}$$

Misleading Accuracy

- Consider a two-class problem:
  - Number of class 0 instances = 9990
  - Number of class 1 instances = 10
- Suppose a model predicts everything to be class 0.
  - It's accuracy is 9990/10000=99.9%.
  - It's accuracy is misleading, because the model does not predict any class 1 instance.

Cost Matrix

		Predicted Class	
		+	-
Actual Class	+	$C(+\|+)$	$C(-\|+)$
	-	$C(+\|-)$	$C(-\|-)$

$C(i|j)$ is the cost of misclassifying a class j instance as class i

Computing the Cost of Classification

		Predicted Class	
		+	-
Actual Class	+	-1	100
	-	1	0

		Predicted Class	
		+	-
Actual Class	+	150	40
	-	60	250

$$Accuracy = 80\%$$
$$Cost = 3910$$

		Predicted Class	
		+	-
Actual Class	+	250	45
	-	5	200

$$Accuracy = 90\%$$
$$Cost = 4255$$

Accuracy

		Predicted Class	
		+	-
Actual Class	+	f_{++} (TP)	f_{+-} (FN)
	-	f_{-+} (FP)	f_{--} (TN)

True positive (TP) or f_{++}: positive instances correctly predicted.
False negative (FN) or f_{+-}: positive instances wrongly predicted.
False positive (FP) or f_{-+}: negative instances wrongly predicted.
True negative(TN) or f_{--}: negative instances correctly predicted.

Confusion Matrix

		Predicted Class	
		+	-
Actual Class	+	f_{++} (TP)	f_{+-} (FN)
	-	f_{-+} (FP)	f_{--} (TN)

True positive rate (TPR): fraction of positive instances correctly predicted.
False positive rate (FPR): fraction of positive instances wrongly predicted.
False negative rate (FNR): fraction of negative instances wrongly predicted.
True negative rate (TNR): fraction of negative instances correctly predicted.

Confusion Matrix

		Predicted Class	
		+	-
Actual Class	+	f_{++} (TP)	f_{+-} (FN)
	-	f_{-+} (FP)	f_{--} (TN)

True positive rate (TPR): $TPR = TP/(TP + FN)$.
False positive rate (FPR): $FPR = FP/(TN + FP)$.
False negative rate (FNR): $FNR = FN/(TP + FN)$.
True negative rate (TNR): $TNR = TN/(TN + FP)$.

Cost-Sensitive Measures

$$\text{Precision}(p) = \frac{TP}{TP + FP}$$

$$\text{Recall}(r) = \frac{TP}{TP + FN}$$

As precision ↑, false positives (TN) ↓.

As recall ↑, false negatives (FN) ↓.

F_1 Measure

$$F_1 \text{ measure } = \frac{2rp}{r+p} = \frac{2 \times TP}{2 \times TP + FP + FN}$$

$$F_1 \text{ measure} = \frac{2}{\frac{1}{r} + \frac{1}{p}}$$

As F_1 measure ↑, false positives (FP) and false negatives (FN) ↓.

F_β Measure

$$F_\beta \text{ measure} = \frac{(\beta^2 + 1)rp}{r + \beta^2 p}$$

Both precision and recall are special cases of F_β where $\beta = 0$ and $\beta = \infty$, respectively. Low values of β make F_β closer to precision; high values make it closer to recall.

Precision and Recall

$$\text{Precision}, p = \frac{TP}{TP+FP}$$

$$\text{Recall}, r = \frac{TP}{TP+FN}$$

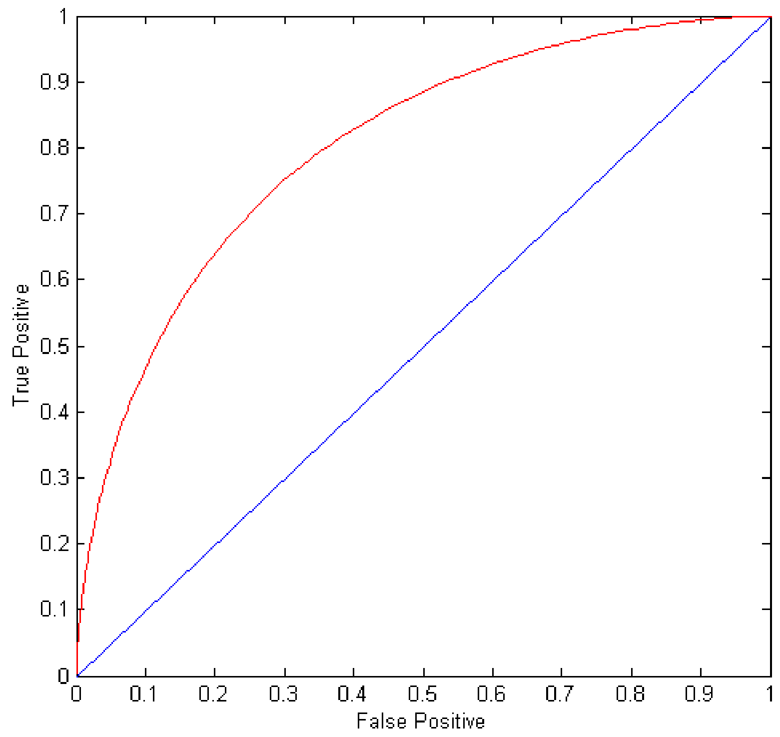
$$F_1\text{measure} = \frac{2rp}{r+p} = \frac{2 \times TP}{2 \times TP + FP + FN}$$

$$F_1\text{measure} = \frac{2}{\frac{1}{r} + \frac{1}{p}}$$

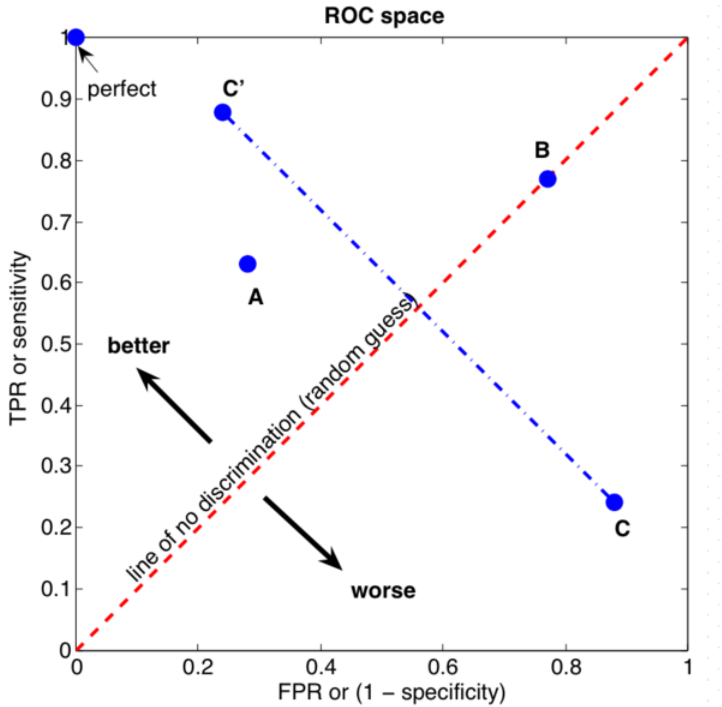
Receiver Operating Characteristic (ROC)

- Developed in the 1950s for signal detection theory to analyze noisy signals.
- ROC curve plots TP (on the y-axis) against FP (on the x-axis).
  - Performance of each classifier represented as a point on the ROC curve.
  - Changing the threshold of algorithm, sample distribution, or cost matrix changes the location of the point.

ROC Curve


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
True Positive
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
False Positive

ROC Curve



A

TP=63	FP=28	91
FN=37	TN=72	109
100	100	200

TPR = 0.63
FPR = 0.28
ACC = 0.68

B

TP=77	FP=77	154
FN=23	TN=23	46
100	100	200

TPR = 0.77
FPR = 0.77
ACC = 0.50

C

TP=24	FP=88	112
FN=76	TN=12	88
100	100	200

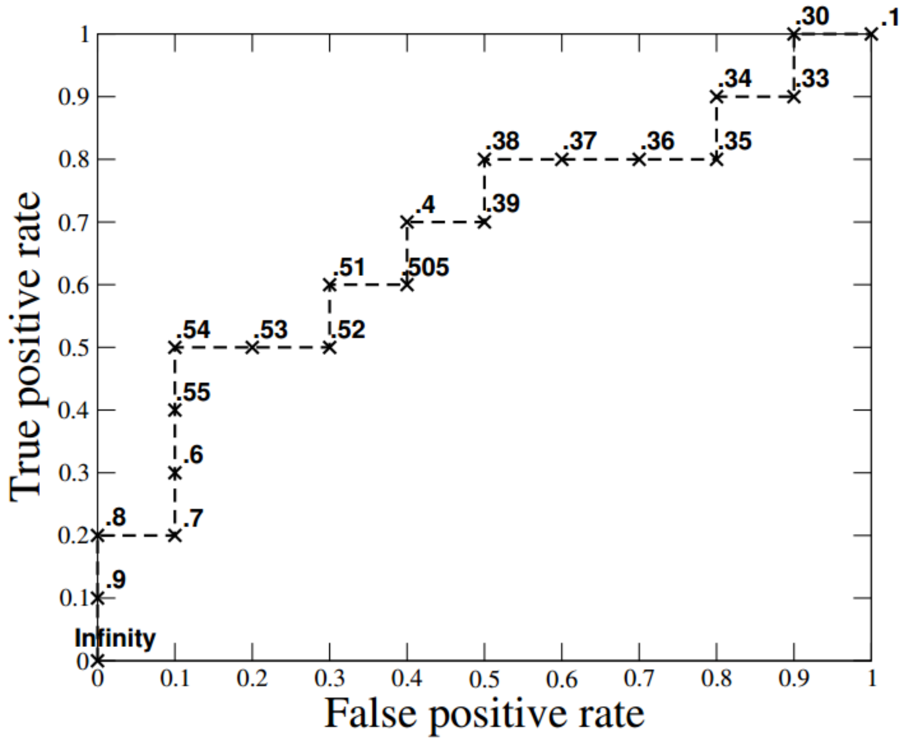
TPR = 0.24
FPR = 0.88
ACC = 0.18

C'

TP=88	FP=24	112
FN=12	TN=76	88
100	100	200

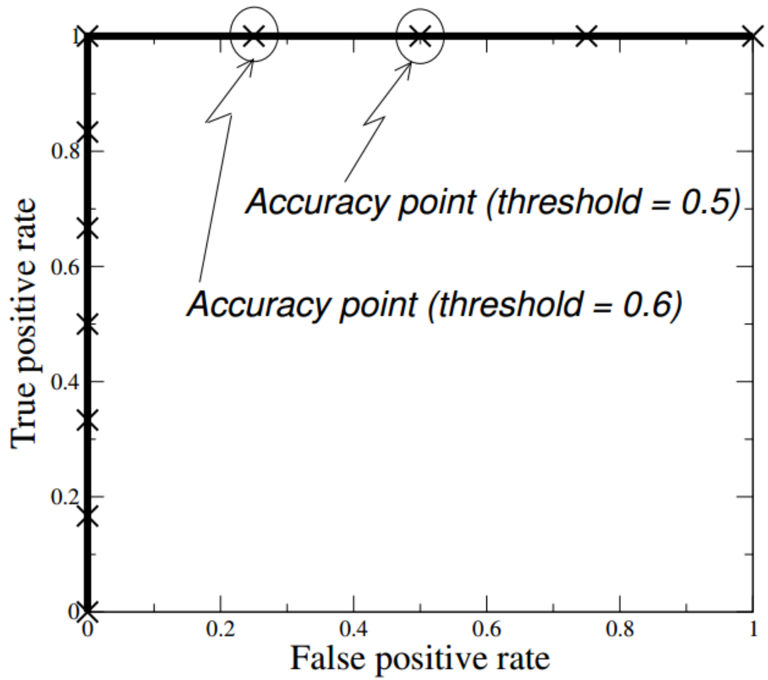
TPR = 0.88
FPR = 0.24
ACC = 0.82

Generating ROC Curves



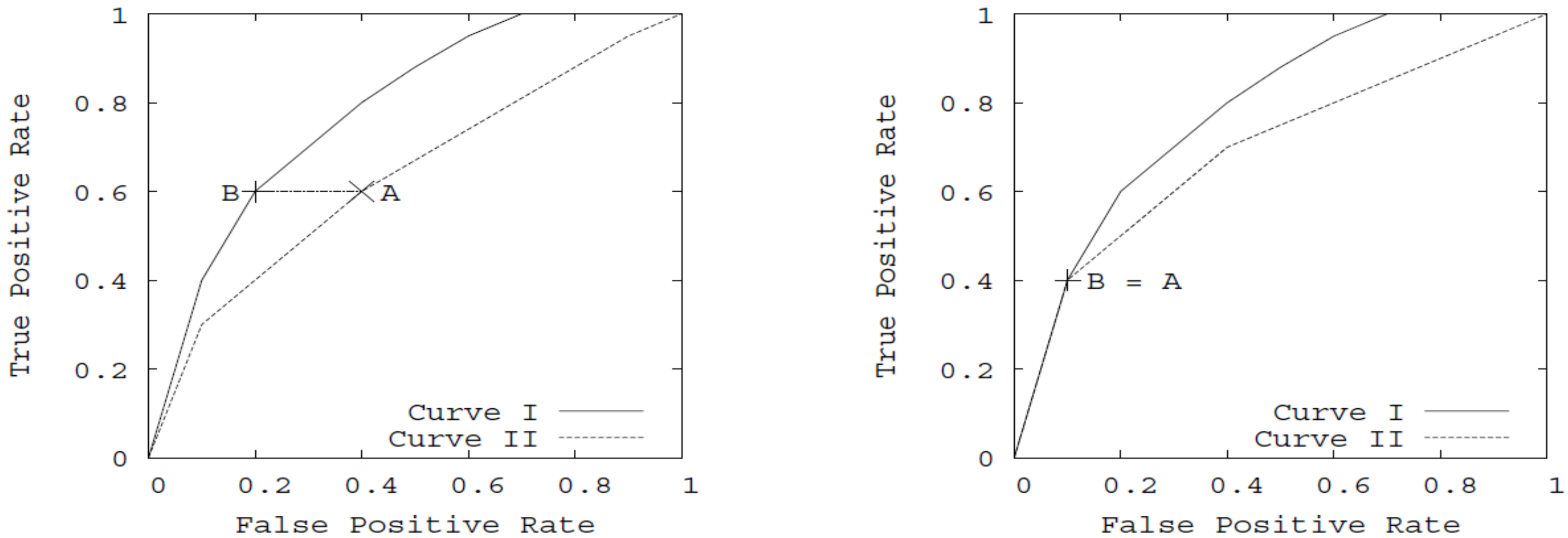
Inst#	Class	Score	Inst#	Class	Score
1	p	.9	11	p	.4
2	p	.8	12	n	.39
3	n	.7	13	p	.38
4	p	.6	14	n	.37
5	p	.55	15	n	.36
6	p	.54	16	n	.35
7	n	.53	17	p	.34
8	n	.52	18	n	.33
9	p	.51	19	p	.30
10	n	.505	20	n	.1

Generating ROC Curves



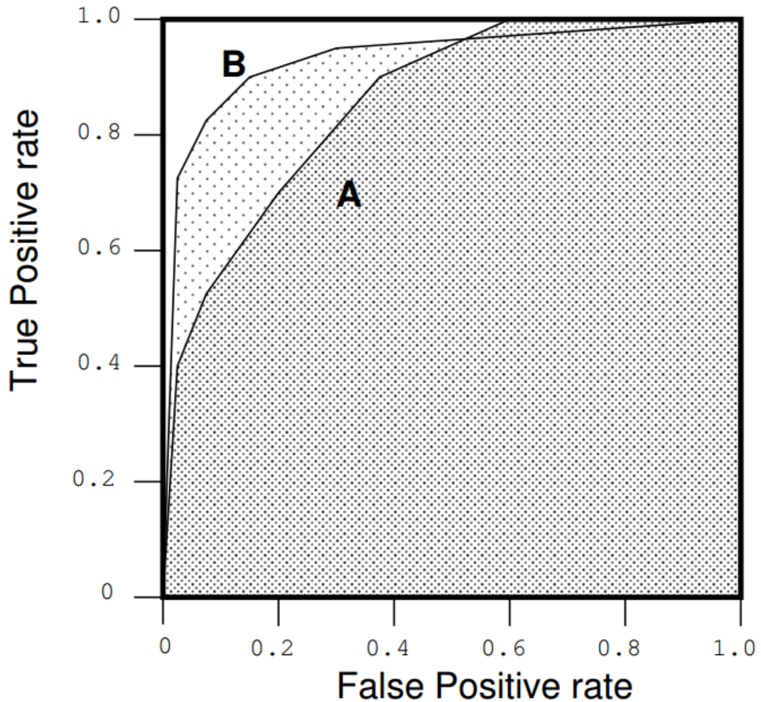
Inst no.	Class True	Class Hyp	Score
1	p	Y	0.99999
2	p	Y	0.99999
3	p	Y	0.99993
4	p	Y	0.99986
5	p	Y	0.99964
6	p	Y	0.99955
7	n	Y	0.68139
8	n	Y	0.50961
9	n	N	0.48880
10	n	N	0.44951

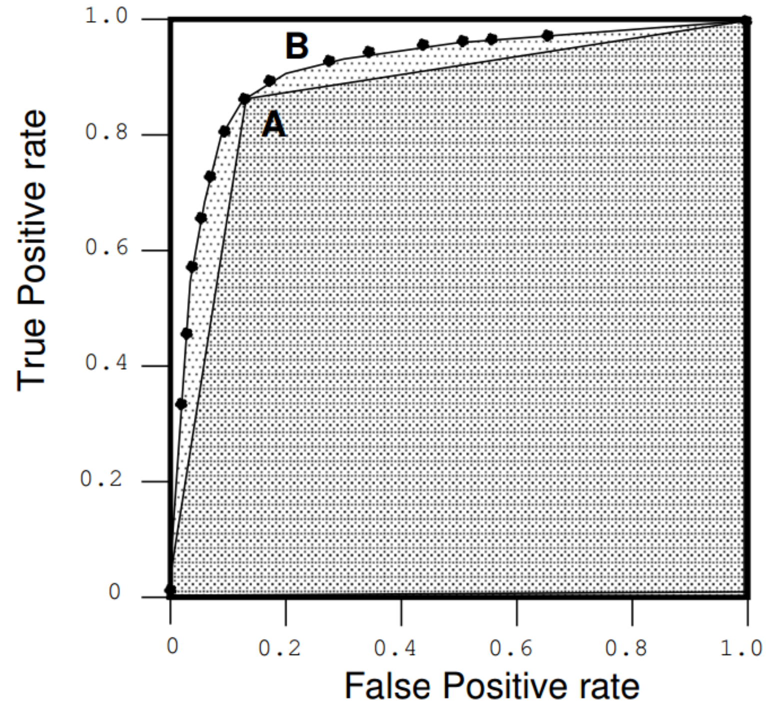
Dominating Classifiers in ROC Space



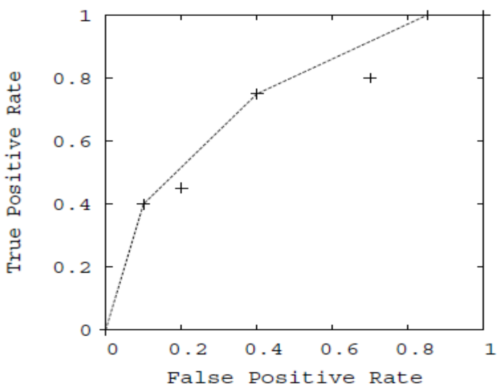
(a) Case 1: $FPR(A) > FPR(B)$ (b) Case 2: $FPR(A) = FPR(B)$

Area Under the ROC Curve

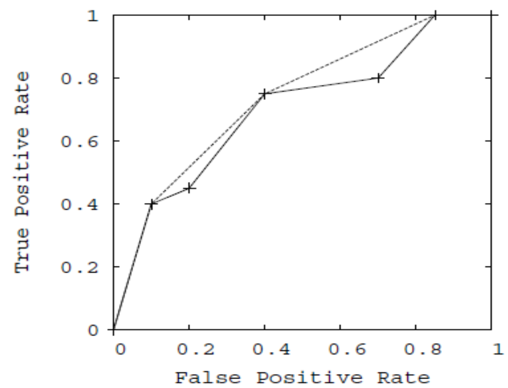




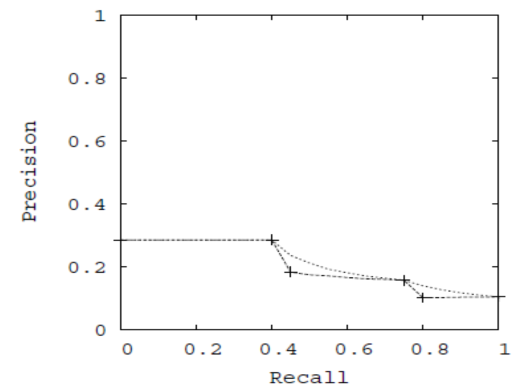
Precision-Recall Curves



(a) Convex hull in ROC space

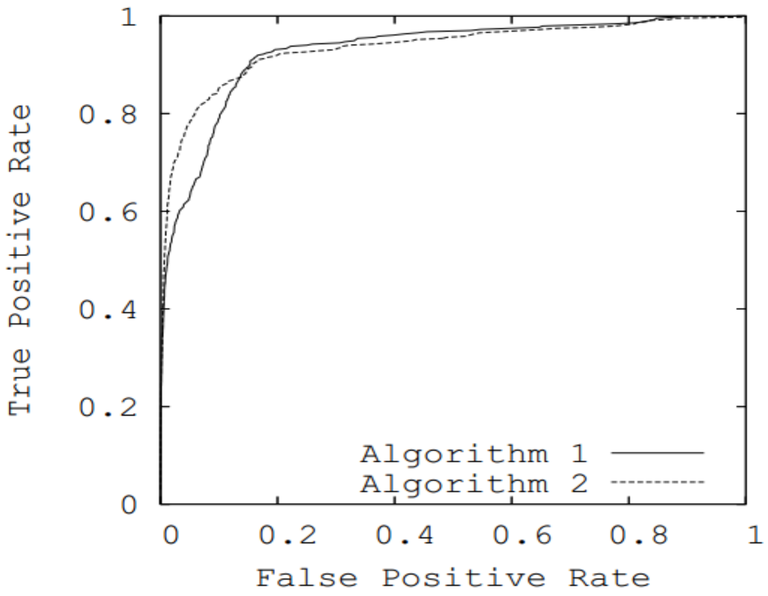


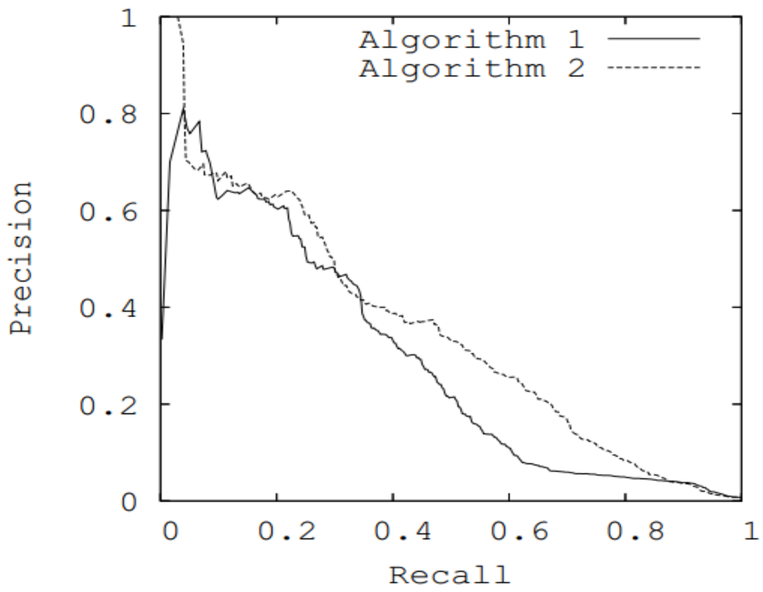
(b) Curves in ROC space



(c) Equivalent curves in PR space

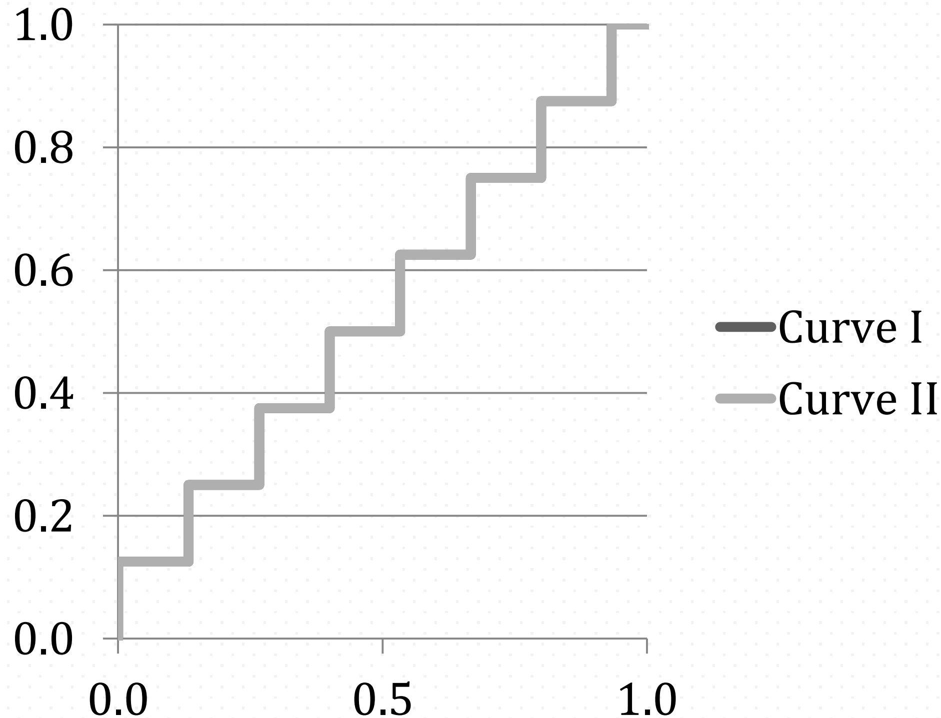
ROC and PR Curves





ROC is Skew Insensitive

Prediction	Class I	Class II
$\mu_0 = max$	1	1
$\mu_1 = \mu_0 - \epsilon_1$	1	0
$\mu_2 = \mu_1 - \epsilon_2$	0	0
$\mu_3 = \mu_2 - \epsilon_3$	0	0
$\mu_4 = \mu_3 - \epsilon_4$	1	1
$\mu_5 = \mu_4 - \epsilon_5$	1	0
$\mu_6 = \mu_5 - \epsilon_6$	0	0
$\mu_7 = \mu_6 - \epsilon_7$	0	0
.	.	.
.	.	.
$\mu_n = \mu_{n-1} - \epsilon_n$	.	.

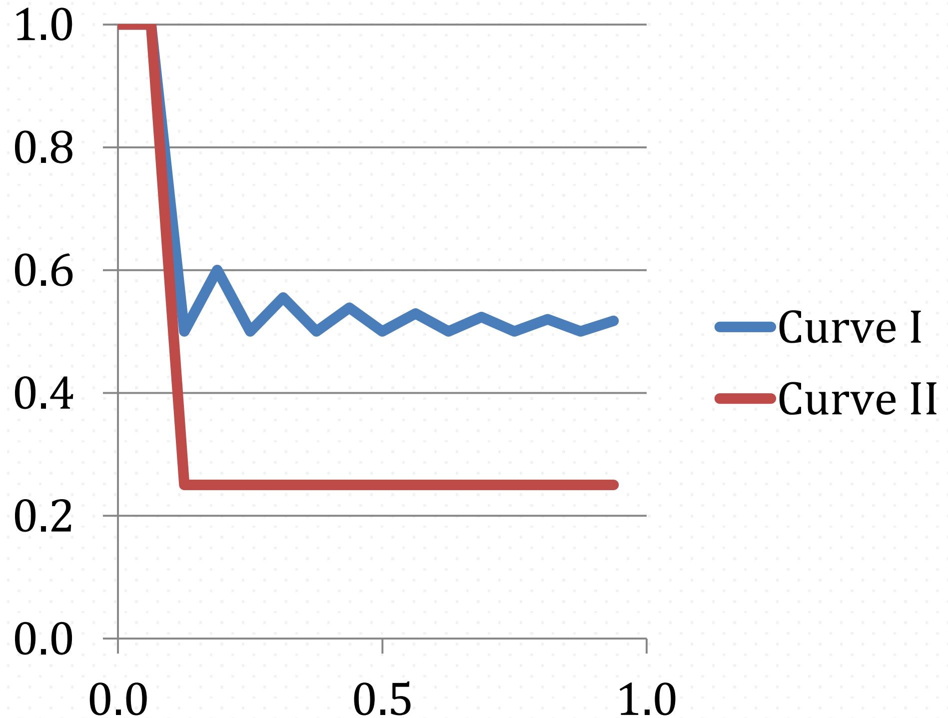


Curve I AUROC = 0.563

Curve II AUROC = 0.563

Precision-Recall is Skew Sensitive

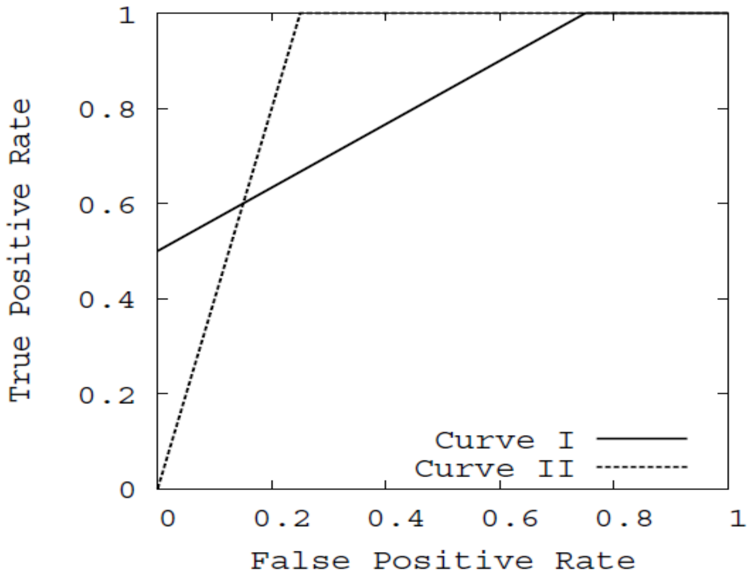
Prediction	Class I	Class II
$\mu_0 = max$	1	1
$\mu_1 = \mu_0 - \epsilon_1$	1	0
$\mu_2 = \mu_1 - \epsilon_2$	0	0
$\mu_3 = \mu_2 - \epsilon_3$	0	0
$\mu_4 = \mu_3 - \epsilon_4$	1	1
$\mu_5 = \mu_4 - \epsilon_5$	1	0
$\mu_6 = \mu_5 - \epsilon_6$	0	0
$\mu_7 = \mu_6 - \epsilon_7$	0	0
.	.	.
.	.	.
$\mu_n = \mu_{n-1} - \epsilon_n$	.	.



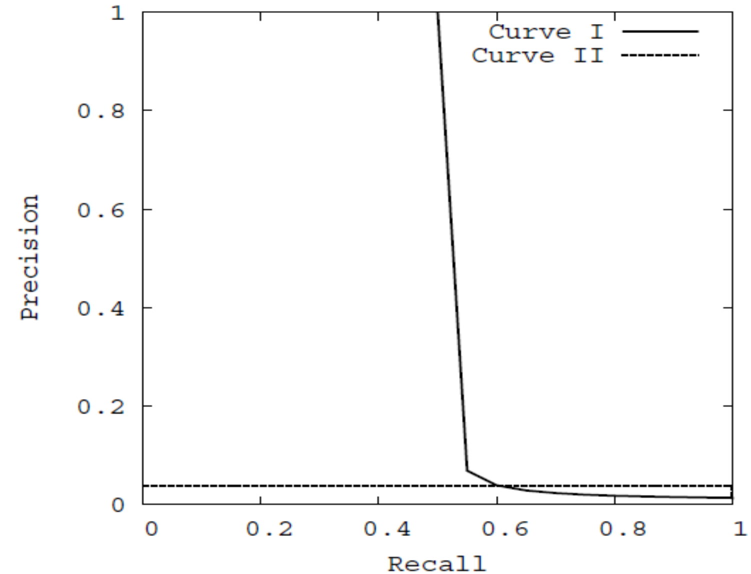
Curve I AUPR = 0.596

Curve II AUPR = 0.370

Optimizing the AUROC vs. AUPR



(a) Comparing AUC-ROC for two algorithms



(b) Comparing AUC-PR for two algorithms

Curve I

AUROC: 0.813

AUPR: **0.514**

Curve II

AUROC: **0.875**

AUPR: 0.038

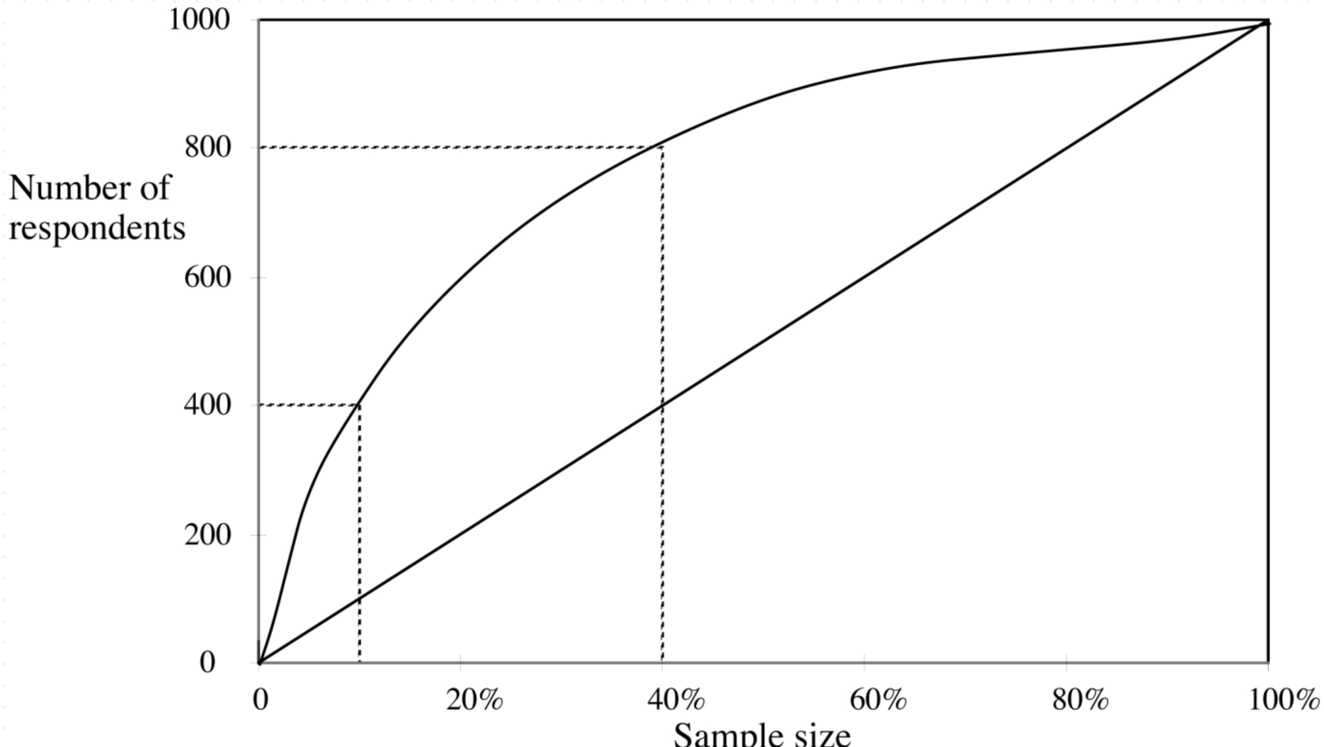
Lift Charts

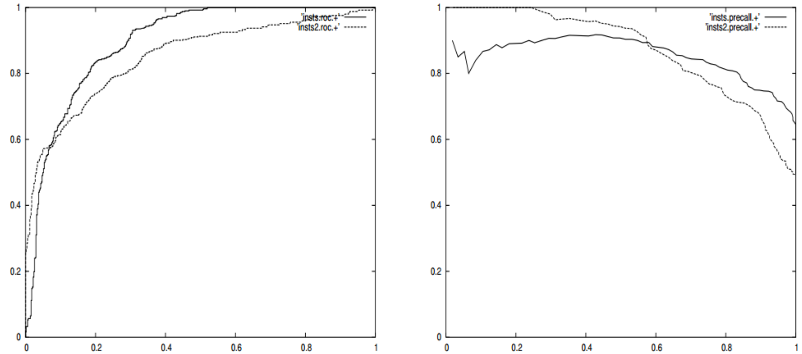
- *Lift* is a measure of the effectiveness of a predictive model calculated as the ratio between the results obtained with and without the predictive model.
- The greater the area between the lift curve and the baseline, the better the model.

Example: Direct Marketing

- Mass mailout of promotional offers (1,000,000).
- The proportion who normally respond is 0.1% (1,000).
- A data mining tool can identify a subset of a 100,000 for which the response rate is 0.4% (400).
- In marketing terminology, the increase of response rate is known as the *lift factor* yielded by the model.
- The same data mining tool may be able to identify 400,000 households for which the response rate is 0.2% (800).
- The overall goal is to find subsets of test instances that have a high proportion of true positives.

Example: Direct Marketing





(a) ROC curves, 1:1 (b) Precision-recall curves, 1:1