Performance Metrics

Part II

TPR, FNR, FPR, TNR

- True positive rate is the accuracy on the positive class: $\frac{TP}{TP+FN}$.
  - Same as recall, also called sensitivity
- False negative rate is the error rate on the positive class: $\frac{FN}{TP+FN}$.
  - Also called miss rate
- False positive rate is the error rate on the negative class: $\frac{FP}{FP+TN}$.
  - Also called fall-out or false alarm rate
- True negative rate is the accuracy on the negative class: $\frac{TN}{FP+TN}$.
  - Also called specificity

Balanced Accuracy

Balanced Accuracy = $\frac{1}{2}$ (True Positive Rate + True Negative Rate)

$$= \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right).$$

Always 0.5 for chance predictions.

Equal to accuracy for balanced datasets.

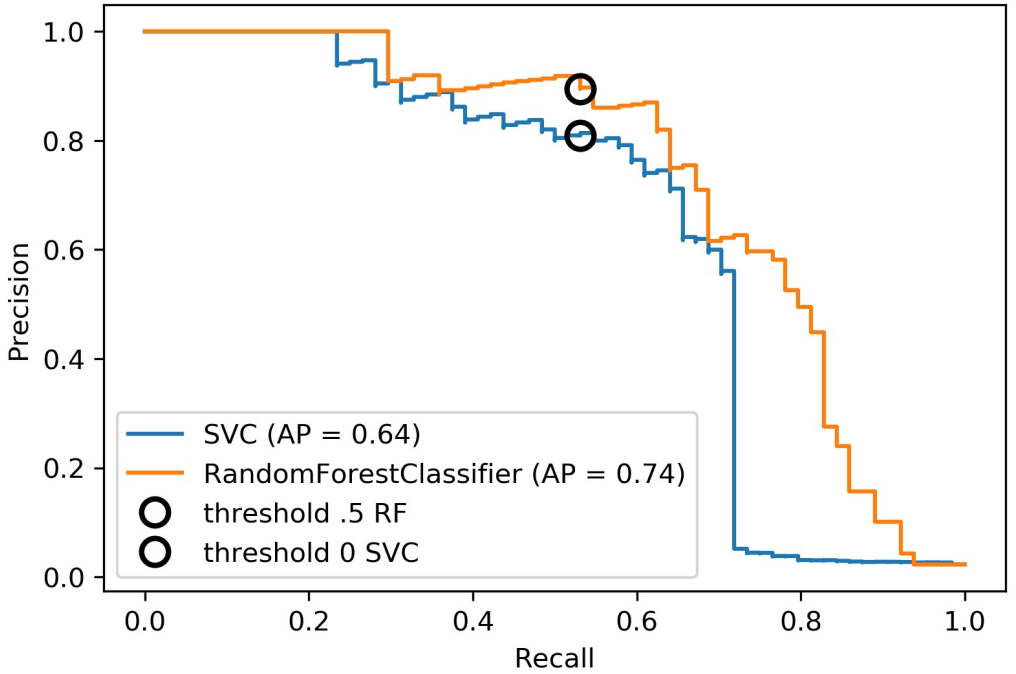
Example: Scores, Predictions, and Labels

ID	Score	Predicted Class	True Class
1	-4.80	-	-
2	-4.43	-	-
3	-2.09	-	-
4	-1.30	-	-
5	-0.53	-	+
6	-0.30	-	+
7	0.49	+	-
8	0.98	+	-
9	2.25	+	+
10	3.37	+	+
11	4.03	+	+
12	4.90	+	+

- Performance measures:
  - Error Rate = 4/12 = .33
  - Precision = 4/6 = .67
  - Recall = 4/6 = .67
  - F1 = 4/6 = .67
- Now predict + iff Score > 2?
  - Error Rate = 2/12 = .17
  - Precision = 4/4 = 1.0
  - Recall = 4/6 = .67
  - F1 = 0.8
- Now predict + iff Score > -1?
  - Error Rate = 2/12 = .17
  - Precision = 6/8 = .75
  - Recall = 6/6 = 1.0
  - F1 = 0.86

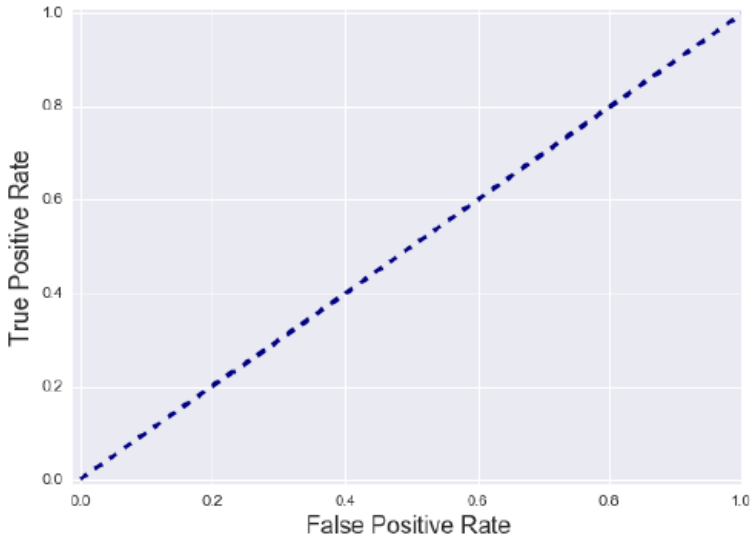
Precision-Recall Curve

- What happens to recall as we decrease threshold?
  - Recall increases (or at least never decreases)
- What happens to precision as we decrease threshold?
  - We expect higher threshold to have higher precision.
  - But no guarantees in general.
- Average Precision (AP) summarizes precision-recall plot as the weighted mean of precisions achieved at each threshold, with the increase in recall from the previous threshold used as the weight.

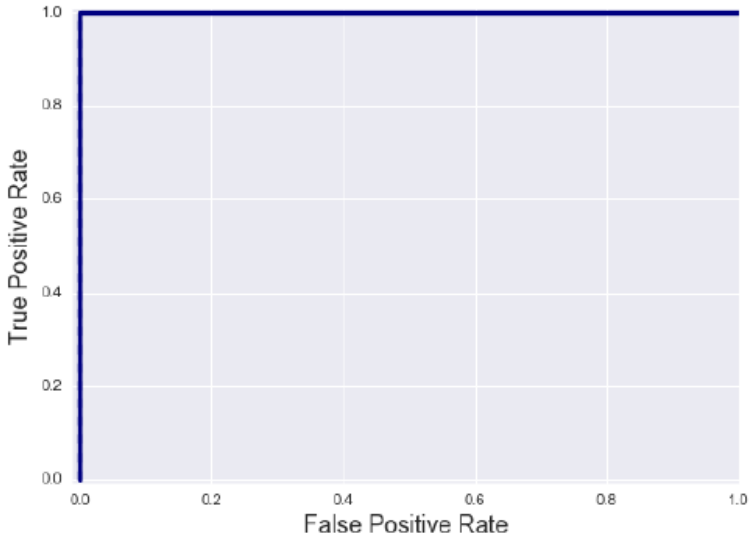


ROC (Receiver Operating Characteristic) Curve

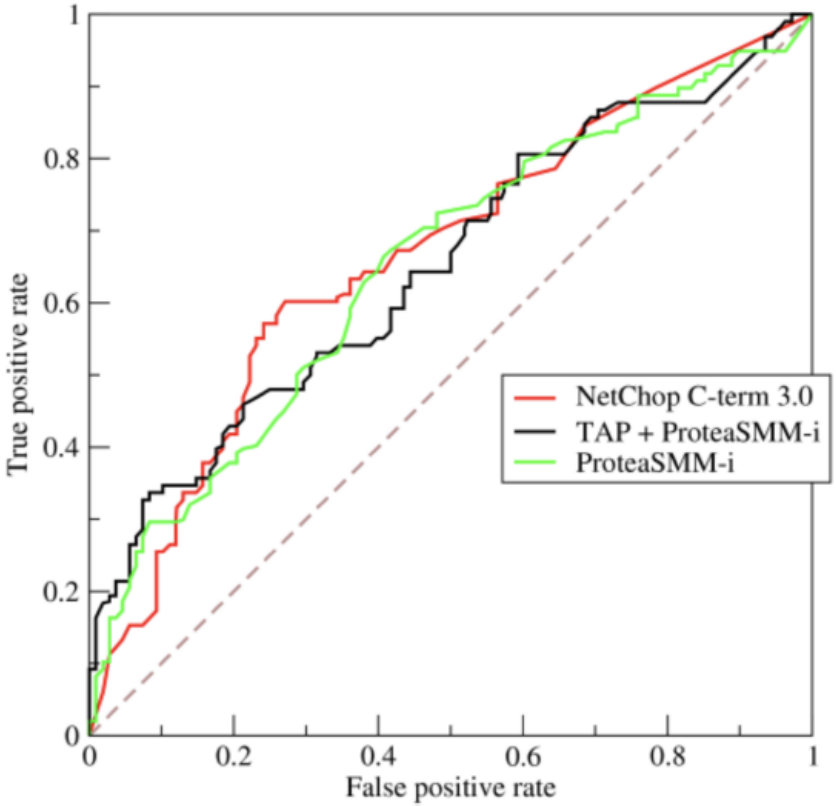
Random Guessing



Perfect Classifier



Comparing ROC Curves



- Here we have ROC curves for 3 score functions.
- For different FPRs, different score functions give better TPRs.
- No score function dominates another at every FPR.
- Can we produce an overall performance measure for a score function?

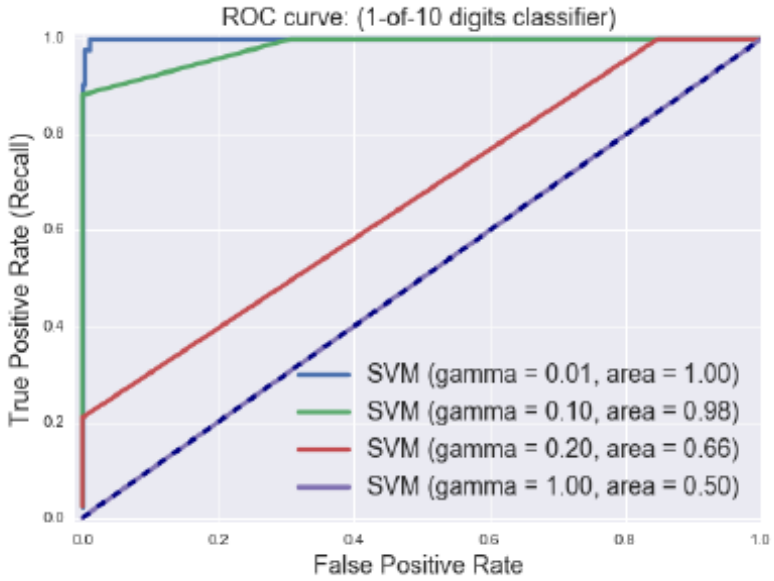
AUC Score

AUC = Area Under the Curve (i.e., the ROC Curve)

It measures how good a classifier is to rank the elements from the most likely to the least likely to be positive.

- AUC = 0.50: random prediction
- AUC = 1.00: perfect prediction

This is a commonly used metric for imbalanced data



The Zoo

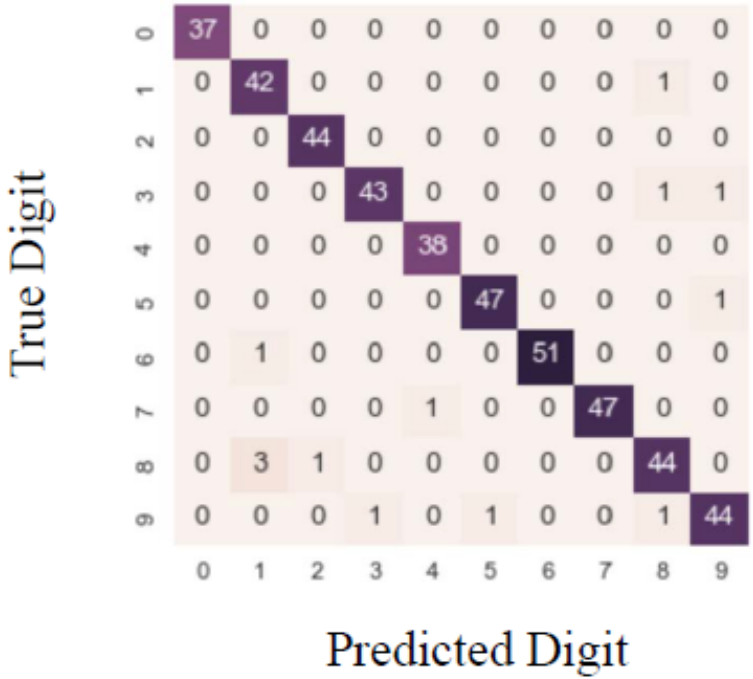
Source: Wikipedia – Precision and Recall

Sources: [5][6][7][8][9][10][11][12] view · talk · edit

		Predicted condition			
	Total population $= P + N$	**Positive (PP)**	**Negative (PN)**	Informedness, bookmaker informedness (BM) $= TPR + TNR - 1$	Prevalence threshold (PT) $= \frac{\sqrt{TPR \times FPR} - FPR}{TPR - FPR}$
Actual condition	**Positive (P)**	**True positive (TP),** **hit**	**False negative (FN),** **type II error, miss,** **underestimation**	True positive rate (TPR), recall, sensitivity (SEN), probability of detection, hit rate, power $= \frac{TP}{P} = 1 - FNR$	False negative rate (FNR), miss rate $= \frac{FN}{P} = 1 - TPR$
	Negative (N)	**False positive (FP),** **type I error, false alarm,** **overestimation**	**True negative (TN),** **correct rejection**	False positive rate (FPR), probability of false alarm, fall-out $= \frac{FP}{N} = 1 - TNR$	True negative rate (TNR), specificity (SPC), selectivity $= \frac{TN}{N} = 1 - FPR$
	Prevalence $= \frac{P}{P + N}$	Positive predictive value (PPV), precision $= \frac{TP}{PP} = 1 - FDR$	False omission rate (FOR) $= \frac{FN}{PN} = 1 - NPV$	Positive likelihood ratio (LR+) $= \frac{TPR}{FPR}$	Negative likelihood ratio (LR−) $= \frac{FNR}{TNR}$
	Accuracy (ACC) $= \frac{TP + TN}{P + N}$	False discovery rate (FDR) $= \frac{FP}{PP} = 1 - PPV$	Negative predictive value (NPV) $= \frac{TN}{PN} = 1 - FOR$	Markedness (MK), deltaP (Δp) $= PPV + NPV - 1$	Diagnostic odds ratio (DOR) $= \frac{LR+}{LR-}$
	Balanced accuracy (BA) $= \frac{TPR + TNR}{2}$	F_1 score $= \frac{2PPV \times TPR}{PPV + TPR} = \frac{2TP}{2TP + FP + FN}$	Fowlkes–Mallows index (FM) $= \sqrt{PPV \times TPR}$	Matthews correlation coefficient (MCC) $= \sqrt{TPR \times TNR \times PPV \times NPV} - \sqrt{FNR \times FPR \times FOR \times FDR}$	Threat score (TS), critical success index (CSI), Jaccard index $= \frac{TP}{TP + FN + FP}$

Multi-Class Evaluation

Multi-Class Confusion Matrix:



Micro Versus Macro Average: Part I

Class	Predicted Class	Correct?
orange	lemon	0
orange	lemon	0
orange	apple	0
orange	orange	1
orange	apple	0
lemon	lemon	1
lemon	apple	0
apple	apple	1
apple	apple	1

Macro-average:

- Each class has equal weight.

1. Compute metric within each class
2. Average resulting metrics across classes

Class	Precision
orange	1/5 = 0.20
lemon	1/2 = 0.50
apple	2/2 = 1.00

Macro-average precision:
(0.20 + 0.50 + 1.00) / 3 = **0.57**

Micro Versus Macro Average: Part II

Class	Predicted Class	Correct?
orange	lemon	0
orange	lemon	0
orange	apple	0
orange	orange	1
orange	apple	0
lemon	lemon	1
lemon	apple	0
apple	apple	1
apple	apple	1

Micro-average:

- Each instance has equal weight.
- Largest classes have most influence

1. Aggregate outcomes across all classes
2. Compute metric with aggregate outcomes

Micro-average precision:
4 / 9 = **0.44**

Micro Versus Macro Average: Part III

- If the classes have about the same number of instances, macro-and micro-average will be about the same.
- If some classes are much larger (more instances) than others, and you want to:
  - Weight your metric toward the largest ones, use micro-averaging.
  - Weight your metric toward the smallest ones, use macro-averaging.
- If the micro-average is much lower than the macro-average then examine the larger classes for poor metric performance.
- If the macro-average is much lower than the micro-average then examine the smaller classes for poor metric performance.