Outline

1) Accuracy

Definition:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} = \frac{TP + TN}{TP + TN + FP + FN}.$$

- Simple to understand: "What fraction of samples did we get right?"
- **Fails on imbalanced data:**
  - If you have 99% negatives and 1% positives, a naive model predicting all negatives (‘print("negative")‘) has 99% accuracy but catches **0%** of actual positives.

2) Confusion Matrix

Binary confusion matrix:

	Pred Positive	**Pred Negative**
Actual Positive	TP (True Positive)	FN (False Negative)
Actual Negative	FP (False Positive)	TN (True Negative)

- All evaluation metrics derive from these four counts: TP, FP, TN, FN.
- Example: "TP = 20" means 20 actual positives were correctly predicted as positive.

3) Recall (Sensitivity/TPR), Specificity (TNR), & Precision

From the confusion matrix, we define:

- **Recall (Sensitivity, TPR)**:

$$\frac{TP}{TP + FN}$$

  "Out of actual positives, how many did we catch?"

- **Specificity (TNR)**:

$$\frac{TN}{TN + FP}$$

  "Out of actual negatives, how many did we correctly reject?"

- **Precision**:

$$\frac{TP}{TP + FP}$$

  "Of the predicted positives, how many truly are positive?"

Additional Terms: FPR and Their Relationships

False Positive Rate (FPR):

$$\text{FPR} = \frac{FP}{TN + FP} = 1 - \text{Specificity}.$$

Summary Table:

Metric	**Formula**	**Interpretation**
Recall(Sen,TPR)	$\frac{TP}{TP+FN}$	How many positives found?
Specificity (TNR)	$\frac{TN}{TN+FP}$	How many negatives correctly rejected?
Precision	$\frac{TP}{TP+FP}$	How often a "positive" prediction is corr
FPR	$\frac{FP}{FP+TN}$	Probability of false alarm

4) F1 Score

F1 Score: Harmonic mean of Precision and Recall

$$F1 = 2 \cdot \frac{(\text{Precision}) \cdot (\text{Recall})}{\text{Precision} + \text{Recall}}$$

- Value ranges from 0 to 1; higher is better.
- F1 = 1 only if Precision = 1 and Recall = 1.
- **Useful when you need to balance false positives and false negatives.**

5) Threshold Dependence

- Many models output a probability score (0 to 1).
- Choosing a different threshold changes TP, FP, TN, FN.
- **Lower threshold** $\Rightarrow$ more positives, typically higher Recall but lower Precision.
- **Higher threshold** $\Rightarrow$ fewer positives, typically higher Precision but lower Recall.

Hence, all these metrics can vary greatly with threshold!

- **Overconfident but often wrong person:**
  - Predicts 0.9 for positive class and 0.2 for negative class.
  - High confidence but often incorrect.
- **Always correct but not confident person:**
  - Predicts 0.4 for positive class and 0.1 for negative class.
  - Low confidence but usually correct.

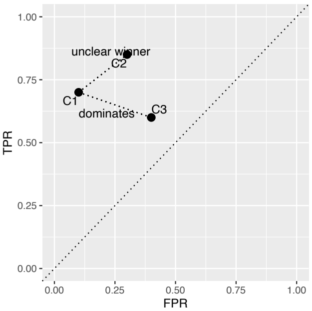
  The second person will predict negative for all samples if the threshold is set to 0.5.

6) ROC Curve & AUC (Conceptual)

- **ROC Curve**: plots TPR (Sensitivity) vs. FPR at various thresholds.
- **AUC**: Area Under the ROC Curve ($0.5 =$ random, $1.0 =$ perfect).
- **Interpretation:**
  - If you vary the threshold from 0 to 1, how do TPR and FPR move?
  - A higher AUC typically means better ability to separate positives from negatives.
- For heavily imbalanced data, sometimes ROC AUC can be overly optimistic.

LABELS: ROC SPACE

- For comparing classifiers, we characterize them by their TPR and FPR values and plot them in a coordinate system.
- We could also use two different ROC metrics which define a trade-off, for instance, TPR and PPV.



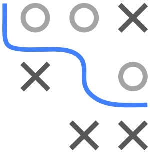
		True Class y	
		$+$	$-$
Pred.	$+$	TP	FP
$\hat{y}$	$-$	FN	TN

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

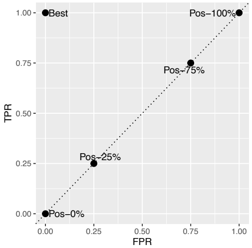
$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

LABELS: ROC SPACE

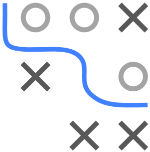
- The best classifier lies on the top-left corner, where FPR equals 0 and TPR is maximal.
- The diagonal is worst as it corresponds to a classifier producing random labels (with different proportions).



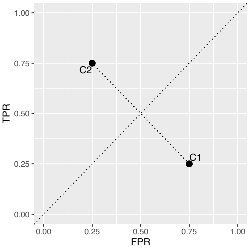
- If each positive x will be randomly classified with 25% as "pos", TPR $= 0.25$.
- If we assign each negative x randomly to "pos", FPR $= 0.25$.



LABELS: ROC SPACE

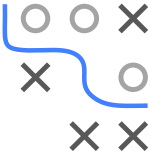


- In practice, we should never obtain a classifier below the diagonal.
- Inverting the predicted labels ($0 \mapsto 1$ and $1 \mapsto 0$) will result in a reflection at the diagonal.
  $\Rightarrow \text{TPR}_{\text{new}} = 1 - \text{TPR}$ and $\text{FPR}_{\text{new}} = 1 - \text{FPR}$.



LABEL DISTRIBUTION IN TPR AND FPR

TPR and FPR (ROC curves) are insensitive to the class distribution in the sense that they are not affected by changes in the ratio n_+/n_- (at prediction).



Example 1:
Proportion $n_+/n_- = 1$

	Actual Positive	Actual Negative
Pred. Positive	40	25
Pred. Negative	10	25

$\text{MCE} = 35/100 = 0.35$
$\text{TPR} = 0.8$
$\text{FPR} = 0.5$

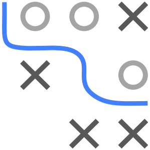
Example 2:
Proportion $n_+/n_- = 2$

	Actual Positive	Actual Negative
Pred. Positive	80	25
Pred. Negative	20	25

$\text{MCE} = 45/150 = 0.3$
$\text{TPR} = 0.8$
$\text{FPR} = 0.5$

Note: If class proportions differ during training, the above is not true. Estimated posterior probabilities can change!

FROM PROBABILITIES TO LABELS: ROC CURVE

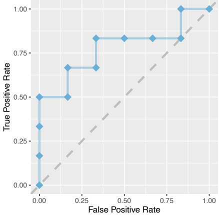


Remember: Both probabilistic and scoring classifiers can output classes by thresholding:

$$h(\mathbf{x}) = [\pi(\mathbf{x}) \geq c] \quad \text{or} \quad h(\mathbf{x}) = [f(\mathbf{x}) \geq c_f].$$

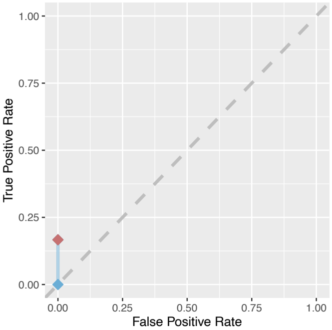
To draw a ROC curve:

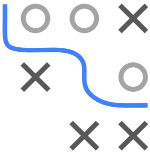
1. Rank test observations on decreasing score.
2. Start with $c = 1$, so we start in $(0, 0)$; we predict everything as negative.
3. Iterate through all possible thresholds c and proceed for each observation x as follows:
   - If x is positive, move TPR $1/n_+$ up, as we have one TP more.
   - If x is negative, move FPR $1/n_-$ right, as we have one FP more.



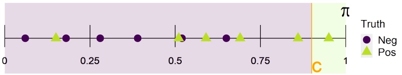
DRAWING ROC CURVES

#	Truth	Score
1	Pos	0.95
2	Pos	0.86
3	Pos	0.69
4	Neg	0.65
5	Pos	0.59
6	Neg	0.52
7	Pos	0.51
8	Neg	0.39
9	Neg	0.28
10	Neg	0.18
11	Pos	0.15
12	Neg	0.06



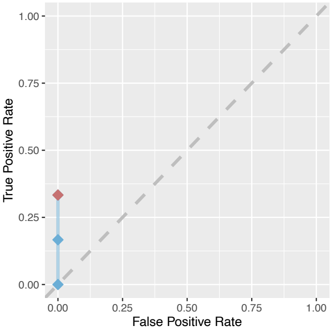


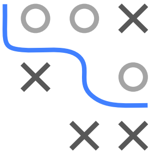
$c = 0.9$
$\rightarrow$ TPR = 0.167
$\rightarrow$ FPR = 0



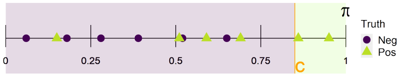
DRAWING ROC CURVES

#	Truth	Score
1	Pos	0.95
2	Pos	0.86
3	Pos	0.69
4	Neg	0.65
5	Pos	0.59
6	Neg	0.52
7	Pos	0.51
8	Neg	0.39
9	Neg	0.28
10	Neg	0.18
11	Pos	0.15
12	Neg	0.06



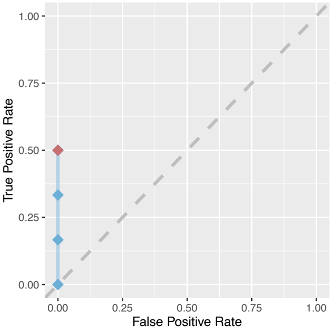


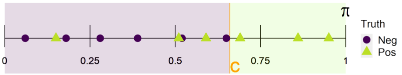
$c = 0.85$
$\rightarrow$ TPR = 0.333
$\rightarrow$ FPR = 0



DRAWING ROC CURVES

#	Truth	Score
1	Pos	0.95
2	Pos	0.86
3	Pos	0.69
4	Neg	0.65
5	Pos	0.59
6	Neg	0.52
7	Pos	0.51
8	Neg	0.39
9	Neg	0.28
10	Neg	0.18
11	Pos	0.15
12	Neg	0.06





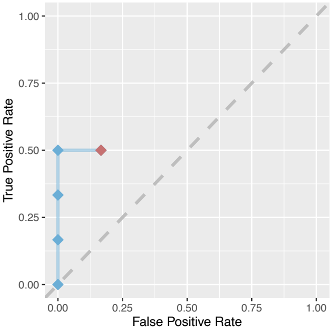
$c = 0.66$

$\rightarrow$ TPR = 0.5

$\rightarrow$ FPR = 0

DRAWING ROC CURVES

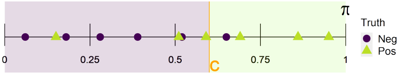
#	Truth	Score
1	Pos	0.95
2	Pos	0.86
3	Pos	0.69
4	Neg	0.65
5	Pos	0.59
6	Neg	0.52
7	Pos	0.51
8	Neg	0.39
9	Neg	0.28
10	Neg	0.18
11	Pos	0.15
12	Neg	0.06



$c = 0.6$

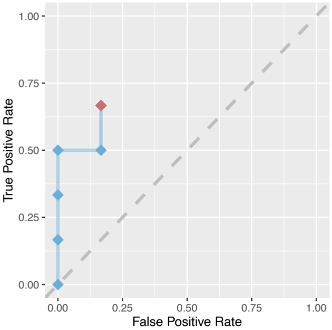
$\rightarrow$ TPR = 0.5

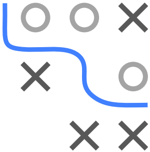
$\rightarrow$ FPR = 0.167



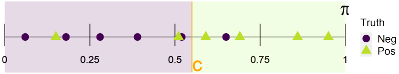
DRAWING ROC CURVES

#	Truth	Score
1	Pos	0.95
2	Pos	0.86
3	Pos	0.69
4	Neg	0.65
5	Pos	0.59
6	Neg	0.52
7	Pos	0.51
8	Neg	0.39
9	Neg	0.28
10	Neg	0.18
11	Pos	0.15
12	Neg	0.06



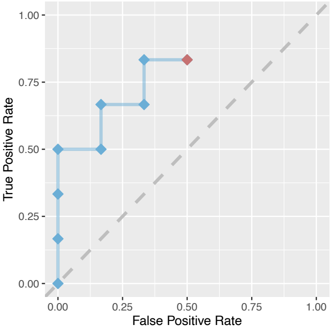


$c = 0.55$
$\rightarrow$ TPR = 0.667
$\rightarrow$ FPR = 0.167

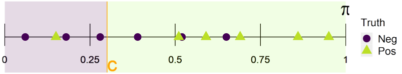


DRAWING ROC CURVES

#	Truth	Score
1	Pos	0.95
2	Pos	0.86
3	Pos	0.69
4	Neg	0.65
5	Pos	0.59
6	Neg	0.52
7	Pos	0.51
8	Neg	0.39
9	Neg	0.28
10	Neg	0.18
11	Pos	0.15
12	Neg	0.06

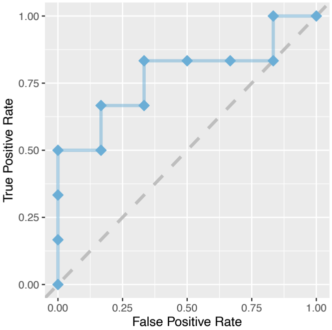


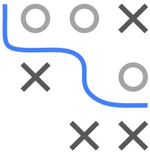
$c = 0.3$
$\rightarrow$ TPR = 0.833
$\rightarrow$ FPR = 0.5



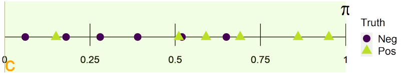
DRAWING ROC CURVES

#	Truth	Score
1	Pos	0.95
2	Pos	0.86
3	Pos	0.69
4	Neg	0.65
5	Pos	0.59
6	Neg	0.52
7	Pos	0.51
8	Neg	0.39
9	Neg	0.28
10	Neg	0.18
11	Pos	0.15
12	Neg	0.06



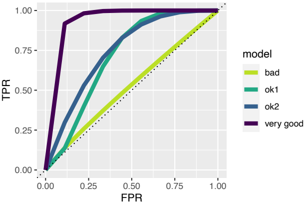


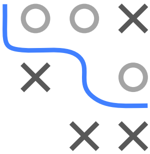
$c = 0$
$\rightarrow$ TPR = 1
$\rightarrow$ FPR = 1



ROC CURVE PROPERTIES

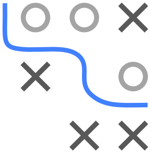
- The closer the curve to the top-left corner, the better.
- If ROC curves cross, a different model might be better in different parts of the ROC space.





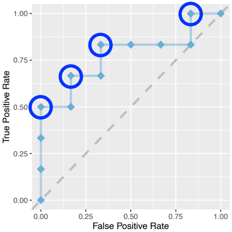
- Small thresholds will very liberally predict the positive class, and result in a potentially higher FPR, but also higher TPR.
- High thresholds will very conservatively predict the positive class, and result in a lower FPR and TPR.
- As we have not defined the trade-off between false positive and false negative costs, we cannot easily select the "best" threshold.
  $\rightarrow$ Visual inspection of all possible results seems useful.

CHOOSING THRESHOLD / OPERATING POINT



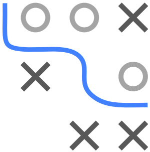
Often done visually and post-hoc, as class imbalances or costs are unknown a-priori.

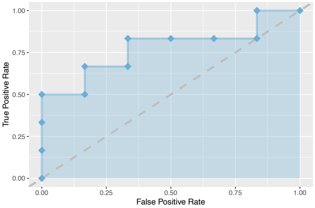
- Identify non-dominated points
- Assess TPR / FPR
- Decide which combo is best for task
- Pick associated threshold



AUC: AREA UNDER ROC CURVE

- AUC $\in [0, 1]$ is a single metric to evaluate scoring classifiers – independent of the chosen threshold.
  - AUC = 1: perfect classifier
  - AUC = 0.5: random, non-discriminant classifier
  - AUC = 0: perfect, with inverted labels





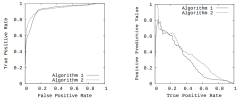
7) Precision-Recall Curve (Conceptual)

Good article

- **Precision-Recall Curve**: Plots Precision vs. Recall across thresholds.
- Especially informative in **imbalanced** datasets (e.g., disease detection).
- Summary metric: **Average Precision (AP)** or the area under the P-R curve.
- If positives are rare, even small changes in FP can significantly affect Precision.

Introduction to Machine Learning

Evaluation
Precision-Recall Curves



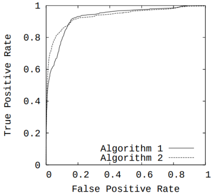
(a) Comparison in ROC space (b) Comparison in PR space

Learning goals

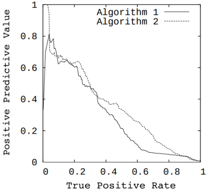
- Understand PR curves
- Same as PPV-TPR curve
- Compare to standard TPR-FPR ROC curve

PRECISION-RECALL CURVES

- Slightly changed ROC plot
- Simply plot precision and recall, instead of TPR-FPR
- Precision = $\rho_{PPV} = \frac{TP}{TP+FP}$, recall = $\rho_{TPR} = \frac{TP}{TP+FN}$
- Might call them TPR-PPV curve
- NB: Both metrics don't depend on TNs



(a) Comparison in ROC space

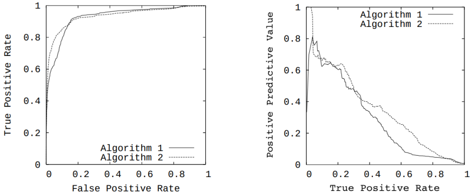


(b) Comparison in PR space

Davis and Goadrich (2006): The Relationship Between Precision-Recall and ROC Curves (URL).

PRECISION-RECALL CURVES

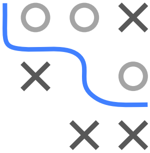
- Might be better for highly imbal data ($n_- \gg n_+$) than TPR-FPR
- Figure (a): ROC; both learners seem to perform well
- Figure (b): PR; visible room for improvement (top-right=best)
- PR reveals better that algo 2 has advantage over 1



(a) Comparison in ROC space
(b) Comparison in PR space

Davis and Goadrich (2006): The Relationship Between Precision-Recall and ROC Curves (URL).

IMBALANCED DATA



- Assume imbalanced classes with $n_- \gg n_+$
- If neg class large, typically less interested in high TNR = low FPR, but more in PPV
- Large (abs) change in FP yields small change in FPR
- PPV likely more informative

<u>FP=10</u>:

	True +1	True -1
Pred. Pos	100	10
Pred. Neg	10	9990
Total	110	10000

$\text{TPR} = 10/11$
$\text{FPR} = 0.001$
$\text{PPV} = 10/11$

<u>FP=100</u>:

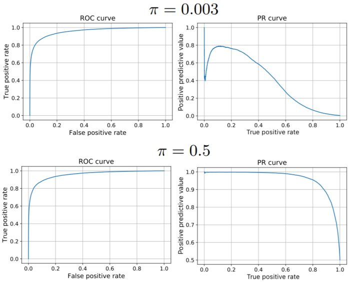
	True +1	True -1
Pred. +1	100	100
Pred. -1	10	9900
Total	110	10000

$\text{TPR} = 10/11$
$\text{FPR} = 0.01$
$\text{PPV} = 1/2$

RHS: Given test says +1, it's now a coin flip that this is correct.

IMBALANCED DATA

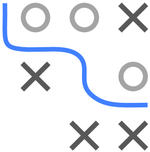
- Top row: Imbal classes with $\pi = 0.003$
- Bottom: balanced with $\pi = 0.5$
- ROC curves (LHS) are similar
- PR curve (RHS) changes strongly from imbal to bal classes

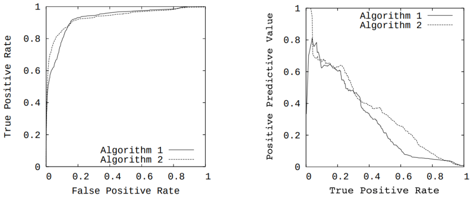


Wissam Siblini et. al. (2004): Master your Metrics with Calibration (URL).

CONCLUSIONS

- Curve fully dominates in ROC space iff dominates in PR-space
- In imbalanced situations rather use PR than standard TPR-FPR
- If comparing few models on a single task, probably plot both. Then observe and think.
- For tuning: can also use PR-AUC (or partial versions)





(a) Comparison in ROC space (b) Comparison in PR space

Davis and Goadrich (2006): The Relationship Between Precision-Recall and ROC Curves (URL).

8) Real-World Examples

Medicine (e.g., Cancer Screening):

- High **Sensitivity** (TPR) is crucial: we do not want to miss actual positives (FN).
- Accepting more FP might be okay, because a false positive leads to follow-up tests rather than missed diagnoses.
- Specificity also matters in large-scale screenings (to avoid overloading the system with false alarms), but typically secondary to not missing positives.

Spam Detection:

- **Precision** often matters more: wrongly classifying an important email as spam has big consequences.
- Recall is still relevant, but missing some spam is often less critical than losing genuine mail.

9) Example Confusion Matrix

Dataset: 50 patients tested for a disease

- 10 are actually positive
- 40 are actually negative

Suppose the model's confusion matrix is:

	Pred Pos	**Pred Neg**	**Total**
Actual Pos (10)	$TP = 8$	$FN = 2$	10
Actual Neg (40)	$FP = 5$	$TN = 35$	40
Total	13	37	50

Metrics Computation

From this table:

- **Accuracy**: $(8 + 35)/50 = 43/50 = 0.86$ (86%).
- **Sensitivity (Recall, TPR)**: $8/(8 + 2) = 0.80$ (80%).
- **Specificity (TNR)**: $35/(35 + 5) = 0.875$ (87.5%).
- **Precision**: $8/(8 + 5) = 0.615$ (61.5%).
- **F1 Score**: $2 \times \frac{0.615 \times 0.80}{0.615 + 0.80} \approx 0.70$.

Notice that while Accuracy is 86%, Precision is only about 62%. Meanwhile, Recall is 80%.

10) Summary

- **Accuracy** can be misleading for imbalanced datasets.
- **Confusion Matrix** reveals TP, FP, TN, FN — the basis for all metrics.
- **Sensitivity (TPR)** & **Specificity (TNR)** show how well we catch positives or avoid false alarms.
- **Precision** checks how reliable a positive prediction is.
- **F1** balances Precision & Recall in a single measure.
- Metrics are **threshold-dependent**; we can analyze performance across thresholds with ROC (TPR vs. FPR) or Precision-Recall curves.
- **Medicine** often demands high Sensitivity; **spam detection** might demand high Precision.

Threshold Tuning

- Sklearn
- ML Mastery