

Evaluation Metrics

CS229

Nandita Bhaskhar

(Adapted from past CS229 teams)

April 22nd, 2022

Topics

- Why are metrics important?
- Binary classifiers
 - Rank view, Thresholding
- Metrics
 - Confusion Matrix
 - Point metrics: Accuracy, Precision, Recall / Sensitivity, Specificity, F-score
 - Summary metrics: AU-ROC, AU-PRC, Log-loss.
- Class Imbalance
 - Failure scenarios for each metric
- Multi-class
- Choosing Metrics
- Unsupervised, Regression, Segmentation Metrics

Why are metrics important?

- Training objective (cost function) is only a proxy for real world objectives.
- Metrics help capture a business goal into a quantitative target (not all errors are equal).
- Helps organize ML team effort towards that target.
 - Generally in the form of improving that metric on the dev set.
- Useful to quantify the “gap” between:
 - Desired performance and baseline (estimate effort initially).
 - Desired performance and current performance.
 - Measure progress over time.
- Useful for lower level tasks and debugging (e.g. diagnosing bias vs variance).
- Ideally training objective should be the metric, but not always possible. Still, metrics are useful and important for evaluation.

Binary Classification

Binary Classification

- x is input
- y is binary output (0/1)
- Model is $\hat{y} = h(x)$
- Two types of models
 - Models that output a categorical class directly (K-nearest neighbor, Decision tree)
 - Models that output a real valued score (SVM, Logistic Regression)
 - Score could be margin (SVM), probability (LR, NN)
 - Need to pick a threshold
 - We focus on this type (the other type can be interpreted as an instance)

Score based models

Score = 1



Score = 0

●	Positive example
○	Negative example

Example of Score: Output of logistic regression.
For most metrics: Only ranking matters.
If too many examples: Plot class-wise histogram.

$$\text{Prevalence} = \frac{\# \text{ positive examples}}{\# \text{ positive examples} + \# \text{ negatives examples}}$$

Threshold \rightarrow Classifier \rightarrow Point Metrics



Point metrics: Confusion Matrix

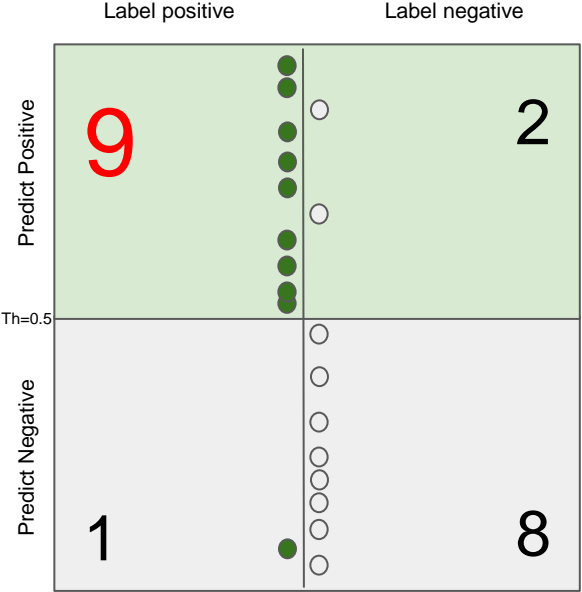


Properties:

- Total sum is fixed (population).
- Column sums are fixed (class-wise population).
- Quality of model & threshold decide how columns are split into rows.
- We want diagonals to be “heavy”, off diagonals to be “light”.

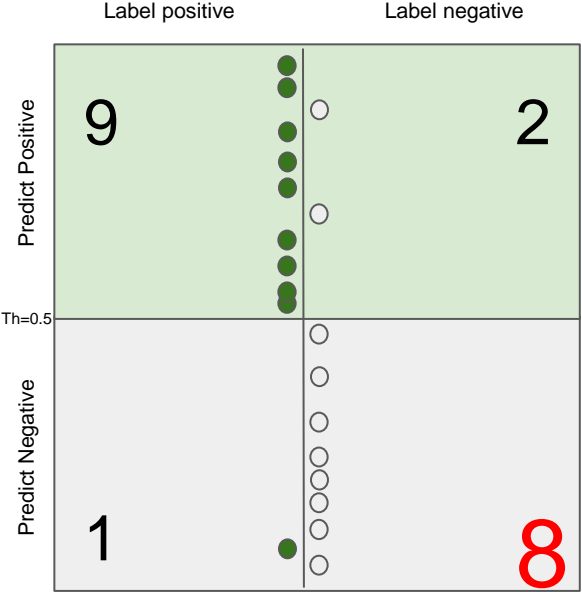
		Predicted condition			
		Positive (PP)	Negative (PN)	Informedness, bookmaker informedness (BM) = TPR + TNR - 1	Prevalence threshold (PT) = $\frac{\sqrt{TPR \times FPR} - FPR}{TPR - FPR}$
Actual condition	Total population = P + N				
	Positive (P)	True positive (TP), hit	False negative (FN), type II error, 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) = $\frac{\sqrt{TPR \times TNR \times PPV \times NPV} - \sqrt{FNR \times FPR \times FOR \times FDR}}$	Threat score (TS), critical success index (CSI), Jaccard index = $\frac{TP}{TP + FN + FP}$

Point metrics: True Positives



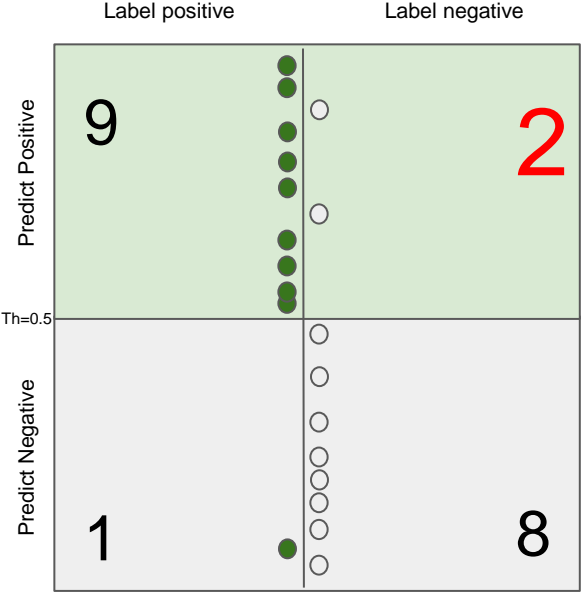
Th	TP
0.5	9

Point metrics: True Negatives



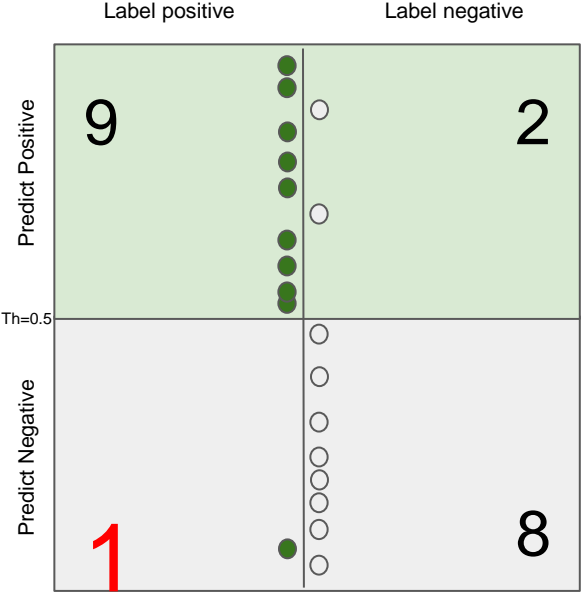
Th	TP	TN
0.5	9	8

Point metrics: False Positives



Th	TP	TN	FP
0.5	9	8	2

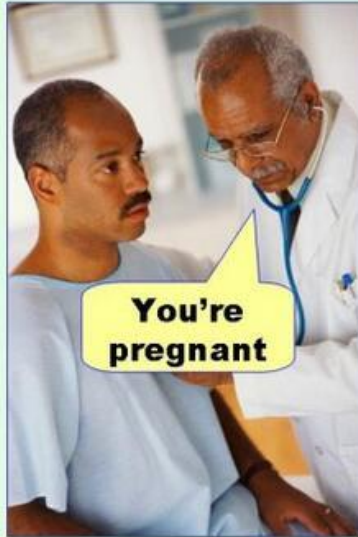
Point metrics: False Negatives



Th	TP	TN	FP	FN
0.5	9	8	2	1

FP and FN also called Type-1 and Type-2 errors

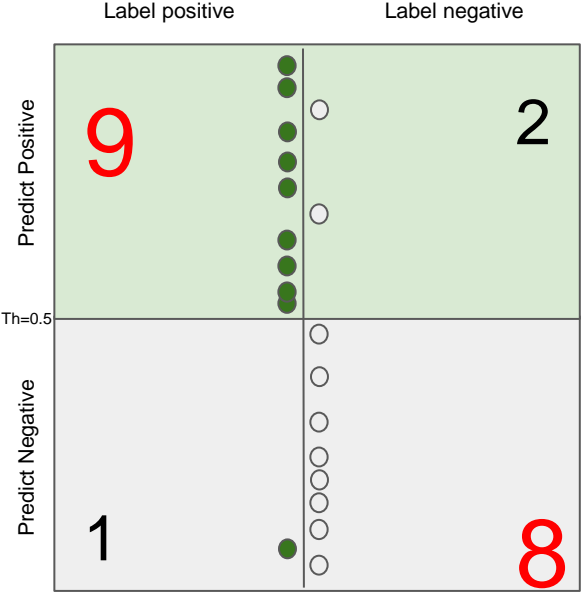
Type I error
(false positive)



Type II error
(false negative)



Point metrics: Accuracy

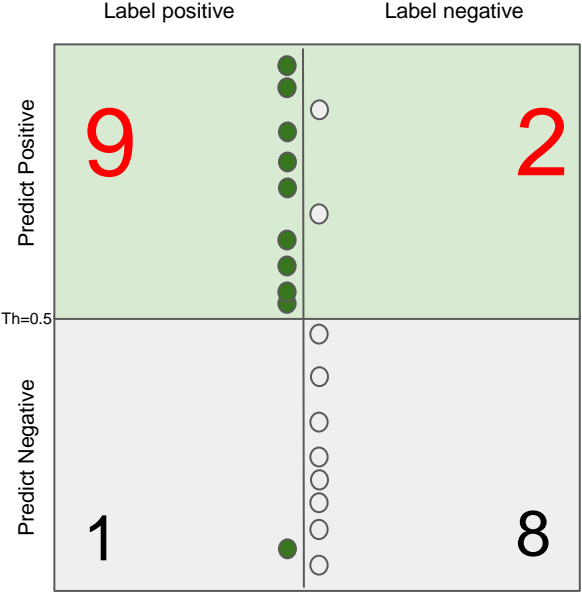


Th	TP	TN	FP	FN	Acc
0.5	9	8	2	1	0.85

Equivalent to 0-1 Loss!

		Predicted condition			
		Positive (PP)	Negative (PN)	Informedness, bookmaker informedness (BM) = TPR + TNR - 1	Prevalence threshold (PT) = $\frac{\sqrt{\text{TPR} \times \text{FPR}} - \text{FPR}}{\text{TPR} - \text{FPR}}$
Actual condition	Total population = P + N				
	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 = $\frac{\text{TP}}{\text{P}} = 1 - \text{FNR}$	False negative rate (FNR), miss rate = $\frac{\text{FN}}{\text{P}} = 1 - \text{TPR}$
	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 = $\frac{\text{FP}}{\text{N}} = 1 - \text{TNR}$	True negative rate (TNR), specificity (SPC), selectivity = $\frac{\text{TN}}{\text{N}} = 1 - \text{FPR}$
	Prevalence = $\frac{\text{P}}{\text{P} + \text{N}}$	Positive predictive value (PPV), precision = $\frac{\text{TP}}{\text{PP}} = 1 - \text{FDR}$	False omission rate (FOR) = $\frac{\text{FN}}{\text{PN}} = 1 - \text{NPV}$	Positive likelihood ratio (LR+) = $\frac{\text{TPR}}{\text{FPR}}$	Negative likelihood ratio (LR-) = $\frac{\text{FNR}}{\text{TNR}}$
	Accuracy (ACC) = $\frac{\text{TP} + \text{TN}}{\text{P} + \text{N}}$	False discovery rate (FDR) = $\frac{\text{FP}}{\text{PP}} = 1 - \text{PPV}$	Negative predictive value (NPV) = $\frac{\text{TN}}{\text{PN}}$ = 1 - FOR	Markedness (MK), deltaP (Δp) = PPV + NPV - 1	Diagnostic odds ratio (DOR) = $\frac{\text{LR}^+}{\text{LR}^-}$
	Balanced accuracy (BA) = $\frac{\text{TPR} + \text{TNR}}{2}$	F_1 score = $\frac{2\text{PPV} \times \text{TPR}}{\text{PPV} + \text{TPR}} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}$	Fowlkes–Mallows index (FM) = $\sqrt{\text{PPV} \times \text{TPR}}$	Matthews correlation coefficient (MCC) = $\frac{\sqrt{\text{TPR} \times \text{TNR} \times \text{PPV} \times \text{NPV}} - \sqrt{\text{FNR} \times \text{FPR} \times \text{FOR} \times \text{FDR}}}{\sqrt{\text{FNR} \times \text{FPR} \times \text{FOR} \times \text{FDR}}}$	Threat score (TS), critical success index (CSI), Jaccard index = $\frac{\text{TP}}{\text{TP} + \text{FN} + \text{FP}}$

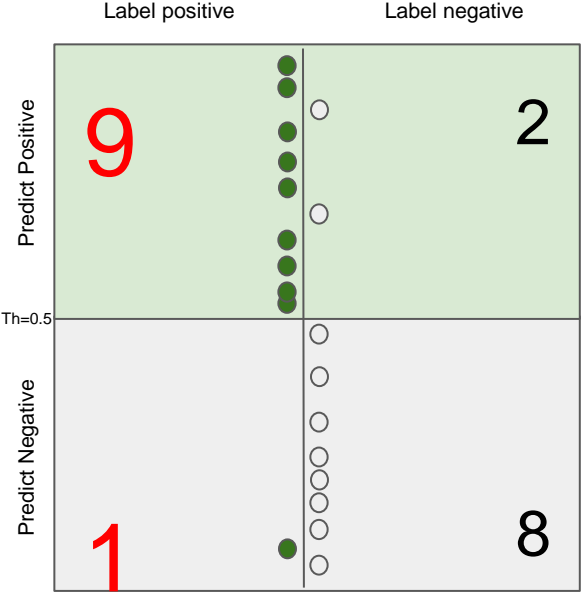
Point metrics: Precision



Th	TP	TN	FP	FN	Acc	Prec
0.5	9	8	2	1	0.85	0.81

		Predicted condition			
		Positive (PP)	Negative (PN)	Informedness, bookmaker informedness (BM) = TPR + TNR - 1	Prevalence threshold (PT) = $\frac{\sqrt{\text{TPR} \times \text{FPR}} - \text{FPR}}{\text{TPR} - \text{FPR}}$
Actual condition	Total population = P + N				
	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 = $\frac{\text{TP}}{\text{P}} = 1 - \text{FNR}$	False negative rate (FNR), miss rate = $\frac{\text{FN}}{\text{P}} = 1 - \text{TPR}$
	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 = $\frac{\text{FP}}{\text{N}} = 1 - \text{TNR}$	True negative rate (TNR), specificity (SPC), selectivity = $\frac{\text{TN}}{\text{N}} = 1 - \text{FPR}$
Prevalence = $\frac{\text{P}}{\text{P} + \text{N}}$	Positive predictive value (PPV), precision = $\frac{\text{TP}}{\text{PP}} = 1 - \text{FDR}$	False omission rate (FOR) = $\frac{\text{FN}}{\text{PN}} = 1 - \text{NPV}$	Positive likelihood ratio (LR+) = $\frac{\text{TPR}}{\text{FPR}}$	Negative likelihood ratio (LR-) = $\frac{\text{FNR}}{\text{TNR}}$	
Accuracy (ACC) = $\frac{\text{TP} + \text{TN}}{\text{P} + \text{N}}$	False discovery rate (FDR) = $\frac{\text{FP}}{\text{PP}} = 1 - \text{PPV}$	Negative predictive value (NPV) = $\frac{\text{TN}}{\text{PN}}$ = 1 - FOR	Markedness (MK), deltaP (Δp) = PPV + NPV - 1	Diagnostic odds ratio (DOR) = $\frac{\text{LR}+}{\text{LR}-}$	
Balanced accuracy (BA) = $\frac{\text{TPR} + \text{TNR}}{2}$	F ₁ score = $\frac{2\text{PPV} \times \text{TPR}}{\text{PPV} + \text{TPR}} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}$	Fowlkes–Mallows index (FM) = $\sqrt{\text{PPV} \times \text{TPR}}$	Matthews correlation coefficient (MCC) = $\frac{\sqrt{\text{TPR} \times \text{TNR} \times \text{PPV} \times \text{NPV}} - \sqrt{\text{FNR} \times \text{FPR} \times \text{FOR} \times \text{FDR}}}{\sqrt{\text{FNR} \times \text{FPR} \times \text{FOR} \times \text{FDR}}}$	Threat score (TS), critical success index (CSI), Jaccard index = $\frac{\text{TP}}{\text{TP} + \text{FN} + \text{FP}}$	

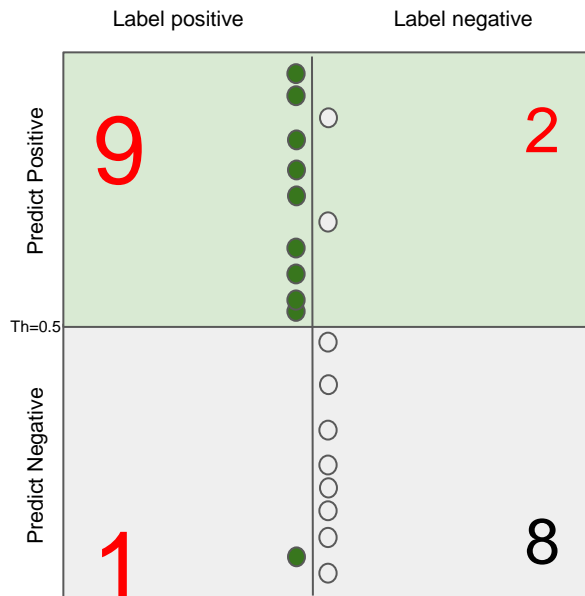
Point metrics: Positive Recall (Sensitivity)



Th	TP	TN	FP	FN	Acc	Prec	Recall
0.5	9	8	2	1	0.85	0.81	0.90

		Predicted condition			
		Positive (PP)	Negative (PN)	Informedness, bookmaker informedness (BM) = TPR + TNR - 1	Prevalence threshold (PT) = $\frac{\sqrt{\text{TPR} \times \text{FPR}} - \text{FPR}}{\text{TPR} - \text{FPR}}$
Actual condition	Total population = P + N				
	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 = $\frac{\text{TP}}{\text{P}} = 1 - \text{FNR}$	False negative rate (FNR), miss rate = $\frac{\text{FN}}{\text{P}} = 1 - \text{TPR}$
	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 = $\frac{\text{FP}}{\text{N}} = 1 - \text{TNR}$	True negative rate (TNR), specificity (SPC), selectivity = $\frac{\text{TN}}{\text{N}} = 1 - \text{FPR}$
	Prevalence = $\frac{\text{P}}{\text{P} + \text{N}}$	Positive predictive value (PPV), precision = $\frac{\text{TP}}{\text{PP}} = 1 - \text{FDR}$	False omission rate (FOR) = $\frac{\text{FN}}{\text{PN}} = 1 - \text{NPV}$	Positive likelihood ratio (LR+) = $\frac{\text{TPR}}{\text{FPR}}$	Negative likelihood ratio (LR-) = $\frac{\text{FNR}}{\text{TNR}}$
	Accuracy (ACC) = $\frac{\text{TP} + \text{TN}}{\text{P} + \text{N}}$	False discovery rate (FDR) = $\frac{\text{FP}}{\text{PP}} = 1 - \text{PPV}$	Negative predictive value (NPV) = $\frac{\text{TN}}{\text{PN}}$ = 1 - FOR	Markedness (MK), deltaP (Δp) = PPV + NPV - 1	Diagnostic odds ratio (DOR) = $\frac{\text{LR}+}{\text{LR}-}$
	Balanced accuracy (BA) = $\frac{\text{TPR} + \text{TNR}}{2}$	F ₁ score = $\frac{2\text{PPV} \times \text{TPR}}{\text{PPV} + \text{TPR}} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}$	Fowlkes–Mallows index (FM) = $\sqrt{\text{PPV} \times \text{TPR}}$	Matthews correlation coefficient (MCC) = $\frac{\sqrt{\text{TPR} \times \text{TNR} \times \text{PPV} \times \text{NPV}} - \sqrt{\text{FNR} \times \text{FPR} \times \text{FOR} \times \text{FDR}}}{\sqrt{\text{FNR} \times \text{FPR} \times \text{FOR} \times \text{FDR}}}$	Threat score (TS), critical success index (CSI), Jaccard index = $\frac{\text{TP}}{\text{TP} + \text{FN} + \text{FP}}$

Point metrics: Precision vs Recall



Th	TP	TN	FP	FN	Acc	Prec	Recall
0.5	9	8	2	1	0.85	0.81	0.90

Trivial 100% recall?

Pull everybody above the threshold.

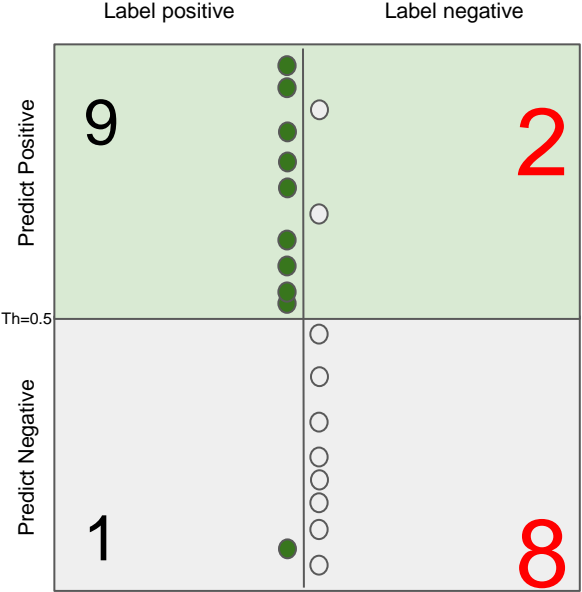
Trivial 100% precision?

Push everybody below the threshold except 1 green on top. (Hopefully no gray above it!)

Striving for good precision with 100% recall = pulling up the lowest green as high as possible in the ranking.

Striving for good recall with 100% precision = pushing down the top gray as low as possible in the ranking.

Point metrics: Negative Recall (Specificity)



Th	TP	TN	FP	FN	Acc	Prec	Recall
0.5	9	8	2	1	0.85	0.81	0.90

Sens.	Spec.
0.90	0.80

		Predicted condition			
		Positive (PP)	Negative (PN)	Informedness, bookmaker informedness (BM) = TPR + TNR - 1	Prevalence threshold (PT) = $\frac{\sqrt{\text{TPR} \times \text{FPR}} - \text{FPR}}{\text{TPR} - \text{FPR}}$
Actual condition	Total population = P + N				
	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 = $\frac{\text{TP}}{\text{P}} = 1 - \text{FNR}$	False negative rate (FNR), miss rate = $\frac{\text{FN}}{\text{P}} = 1 - \text{TPR}$
	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 = $\frac{\text{FP}}{\text{N}} = 1 - \text{TNR}$	True negative rate (TNR), specificity (SPC), selectivity = $\frac{\text{TN}}{\text{N}} = 1 - \text{FPR}$
	Prevalence = $\frac{\text{P}}{\text{P} + \text{N}}$	Positive predictive value (PPV), precision = $\frac{\text{TP}}{\text{PP}} = 1 - \text{FDR}$	False omission rate (FOR) = $\frac{\text{FN}}{\text{PN}} = 1 - \text{NPV}$	Positive likelihood ratio (LR+) = $\frac{\text{TPR}}{\text{FPR}}$	Negative likelihood ratio (LR-) = $\frac{\text{FNR}}{\text{TNR}}$
	Accuracy (ACC) = $\frac{\text{TP} + \text{TN}}{\text{P} + \text{N}}$	False discovery rate (FDR) = $\frac{\text{FP}}{\text{PP}} = 1 - \text{PPV}$	Negative predictive value (NPV) = $\frac{\text{TN}}{\text{PN}}$ = 1 - FOR	Markedness (MK), deltaP (Δp) = PPV + NPV - 1	Diagnostic odds ratio (DOR) = $\frac{\text{LR}+}{\text{LR}-}$
	Balanced accuracy (BA) = $\frac{\text{TPR} + \text{TNR}}{2}$	F ₁ score = $\frac{2\text{PPV} \times \text{TPR}}{\text{PPV} + \text{TPR}} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}$	Fowlkes–Mallows index (FM) = $\sqrt{\text{PPV} \times \text{TPR}}$	Matthews correlation coefficient (MCC) = $\frac{\sqrt{\text{TPR} \times \text{TNR} \times \text{PPV} \times \text{NPV}} - \sqrt{\text{FNR} \times \text{FPR} \times \text{FOR} \times \text{FDR}}}{1}$	Threat score (TS), critical success index (CSI), Jaccard index = $\frac{\text{TP}}{\text{TP} + \text{FN} + \text{FP}}$

Point metrics: F1-score



Th	TP	TN	FP	FN	Acc	Prec	Recall		F1
0.5	9	8	2	1	0.85	0.81	0.90		0.857
							Sens.	Spec.	
							0.90	0.80	

$$F_1 = \left(\frac{2}{\text{recall}^{-1} + \text{precision}^{-1}} \right) = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

		Predicted condition			
		Positive (PP)	Negative (PN)	Informedness, bookmaker informedness (BM) = TPR + TNR - 1	Prevalence threshold (PT) = $\frac{\sqrt{TPR \times FPR} - FPR}{TPR - FPR}$
Actual condition	Total population = P + N				
	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) = $\frac{TPR + TNR}{2}$	F₁ score = $\frac{2PPV \times TPR}{PPV + TPR} = \frac{2TP}{2TP + FP + FN}$	Fowlkes–Mallows index (FM) = $\sqrt{PPV \times TPR}$	Matthews correlation coefficient (MCC) = $\frac{\sqrt{TPR \times TNR \times PPV \times NPV} - \sqrt{FNR \times FPR \times FOR \times FDR}}$	Threat score (TS), critical success index (CSI), Jaccard index = $\frac{TP}{TP + FN + FP}$

Point metrics: Changing threshold



Th	TP	TN	FP	FN	Acc	Prec	Recall		F1
0.6	7	8	2	3	0.75	0.77	0.70		0.733
							Sens.	Spec.	
							0.70	0.80	

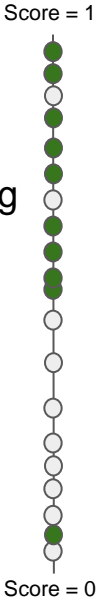
effective thresholds = # examples + 1

Threshold Scanning



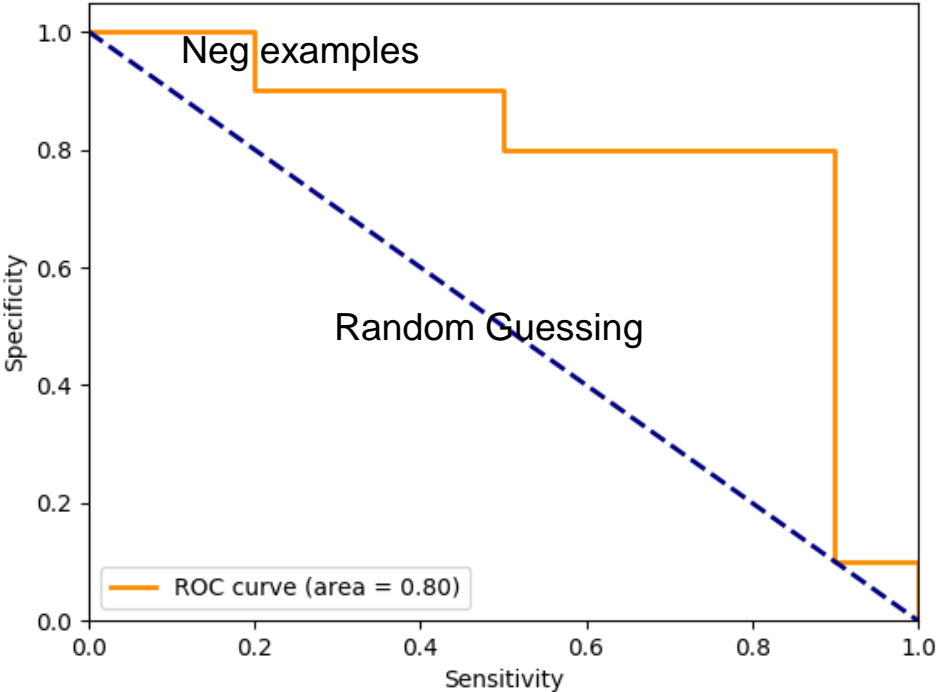
Threshold	TP	TN	FP	FN	Accuracy	Precision	Recall	Specificity	F1
1.00	0	10	0	10	0.50	1	0	1	0
0.95	1	10	0	9	0.55	1	0.1	1	0.182
0.90	2	10	0	8	0.60	1	0.2	1	0.333
0.85	2	9	1	8	0.55	0.667	0.2	0.9	0.308
0.80	3	9	1	7	0.60	0.750	0.3	0.9	0.429
0.75	4	9	1	6	0.65	0.800	0.4	0.9	0.533
0.70	5	9	1	5	0.70	0.833	0.5	0.9	0.625
0.65	5	8	2	5	0.65	0.714	0.5	0.8	0.588
0.60	6	8	2	4	0.70	0.750	0.6	0.8	0.667
0.55	7	8	2	3	0.75	0.778	0.7	0.8	0.737
0.50	8	8	2	2	0.80	0.800	0.8	0.8	0.800
0.45	9	8	2	1	0.85	0.818	0.9	0.8	0.857
0.40	9	7	3	1	0.80	0.750	0.9	0.7	0.818
0.35	9	6	4	1	0.75	0.692	0.9	0.6	0.783
0.30	9	5	5	1	0.70	0.643	0.9	0.5	0.750
0.25	9	4	6	1	0.65	0.600	0.9	0.4	0.720
0.20	9	3	7	1	0.60	0.562	0.9	0.3	0.692
0.15	9	2	8	1	0.55	0.529	0.9	0.2	0.667
0.10	9	1	9	1	0.50	0.500	0.9	0.1	0.643
0.05	10	1	9	0	0.55	0.526	1	0.1	0.690
0.00	10	0	10	0	0.50	0.500	1	0	0.667

Summary metrics: Rotated ROC (Sen vs. Spec)



Specificity
= True Neg / Neg

Pos examples Receiver operating characteristic



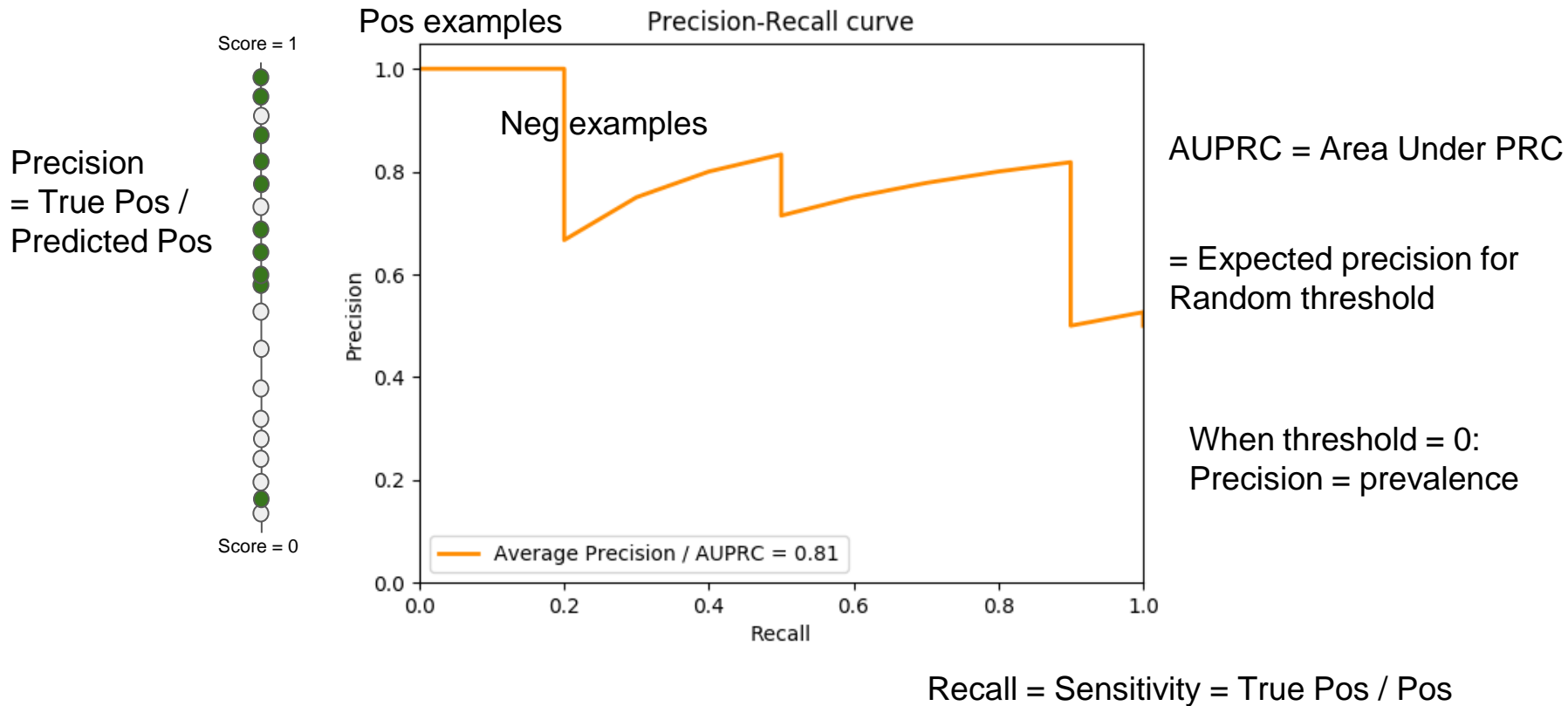
AUROC = Area Under ROC

= Prob[Random Pos ranked higher than random Neg]

Agnostic to prevalence!

Sensitivity = True Pos / Pos

Summary metrics: PRC (Recall vs. Precision)



Summary metrics:



Two models scoring the same data set. Is one of them better than the other?

Summary metrics: Log-Loss vs Brier Score

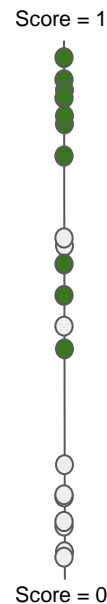
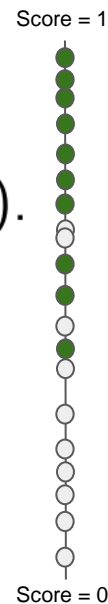
- Same ranking, and therefore the same AUROC, AUPRC, accuracy!

$$\text{Log Loss} = \frac{1}{N} \sum_{i=1}^N -y_i \log \hat{y}_i - (1 - y_i) \log (1 - \hat{y}_i).$$

- Rewards confident correct answers, heavily penalizes confident wrong answers.
 - One perfectly confident wrong prediction is fatal.
- > Well-**calibrated** model

- **Proper** scoring rule: Minimized at $\hat{y} = y$

$$\text{Brier Score} = \frac{1}{N} \sum_{i=1}^N (\hat{y}_i - y_i)^2$$

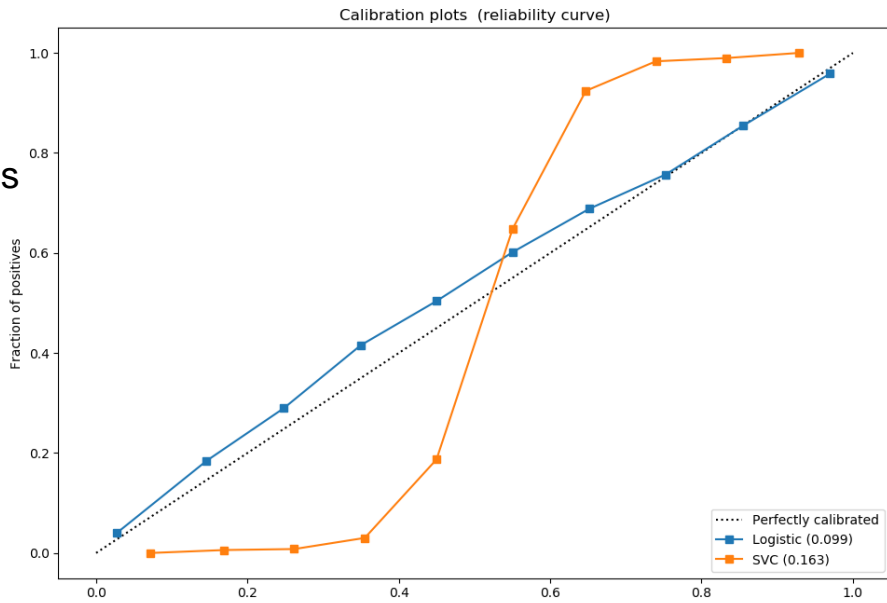


Calibration vs Discriminative Power

Logistic (th=0.5):
Precision: 0.872
Recall: 0.851
F1: 0.862
Brier: 0.099

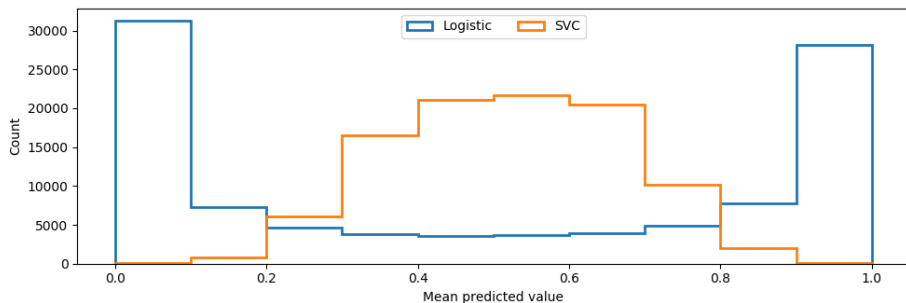
SVC (th=0.5):
Precision: 0.872
Recall: 0.852
F1: 0.862
Brier: 0.163

Fraction of Positives



Output

Histogram



Binary Classification Under Class Imbalance

Class Imbalance

Symptom: Prevalence $< 5\%$ (no strict definition)

Metrics: May not be meaningful.

Learning: May not focus on minority class examples at all

(majority class can overwhelm logistic regression, to a lesser extent SVM)

What happen to the metrics under class imbalance?

Accuracy: Blindly predicts majority class -> prevalence is the baseline.

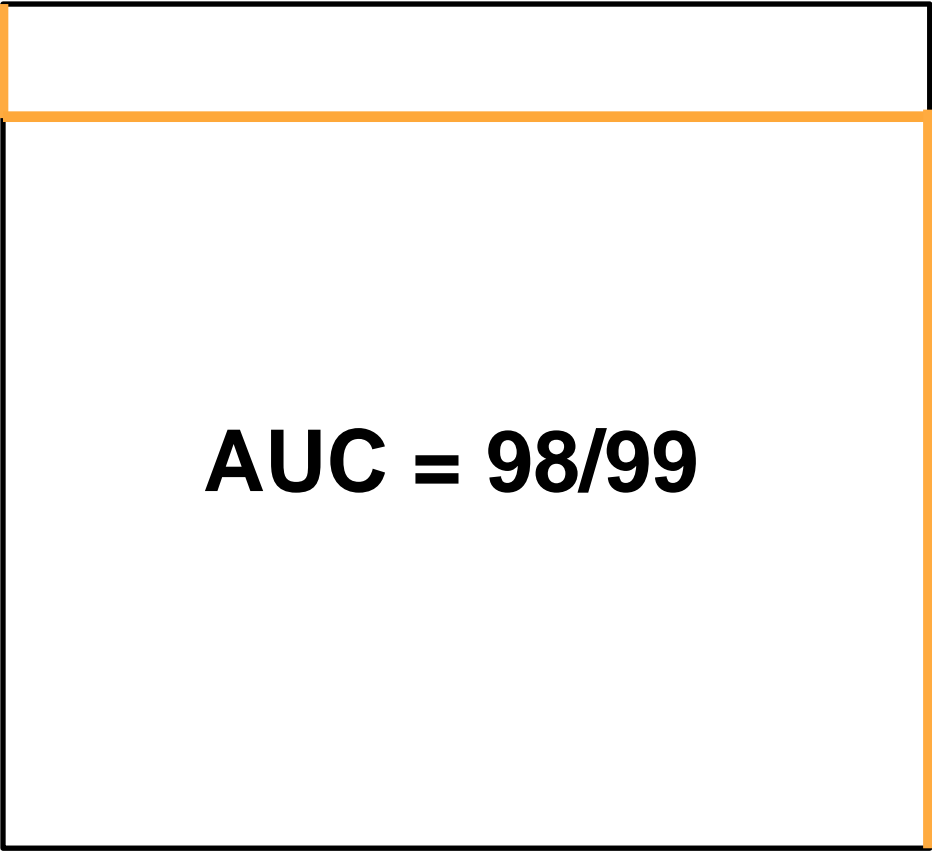
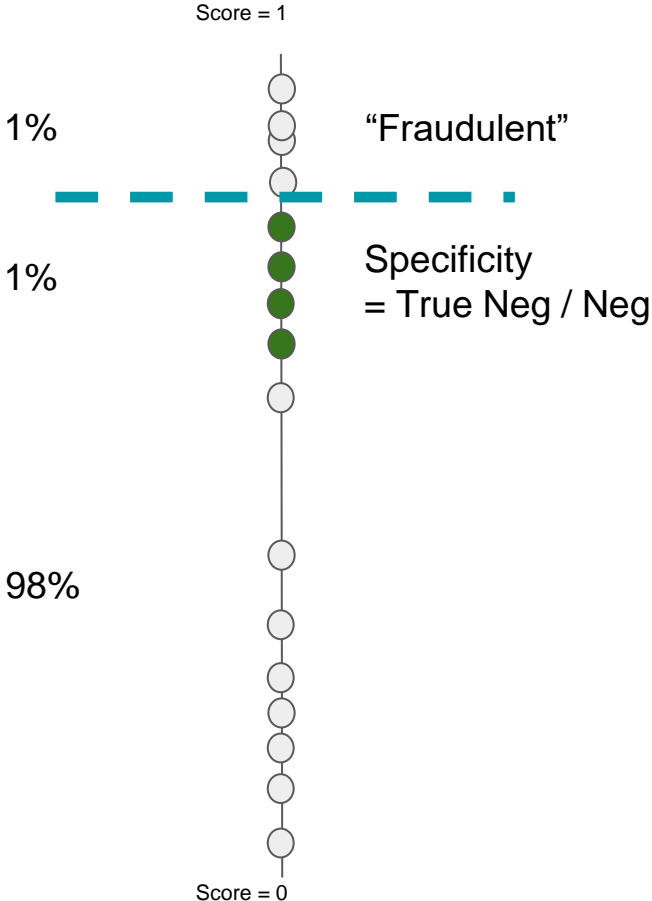
Log-Loss: Majority class can dominate the loss.

AUROC: Easy to keep AUC high by scoring most negatives very low.

AUPRC: More robust than AUROC. But other challenges.

In general: Accuracy < AUROC < AUPRC

Rotated ROC



Sensitivity = True Pos / Pos

Multi-Class Classification

Multi-class

(white board)

- Confusion matrix will be $N * N$ (still want heavy diagonals, light off-diagonals)
- Most metrics (except accuracy) generally analyzed as multiple 1-vs-many
- Multiclass variants of other metrics (micro vs macro averaging)
 - Macro: Class-wise, each class is given the same weightage
 - Micro: No class differences considered. $F1 = Pr = Re = Acc$
- Class imbalance is common (both in absolute and relative sense)
- Cost sensitive learning techniques (also helps in binary Imbalance)
 - Assign weights for each block in the confusion matrix.
 - Incorporate weights into the loss function.

Choosing Metrics

Choosing Metrics

Some common patterns:

- High precision is hard constraint, do best recall (search engine results, grammar correction): Intolerant to FP
 - Metric: Recall at Precision = XX %
- High recall is hard constraint, do best precision (medical diagnosis): Intolerant to FN
 - Metric: Precision at Recall = 100 %
- Capacity constrained (by K)
 - Metric: Precision in top-K.
-

Unsupervised Learning,
Regression,
Semantic Segmentation

Unsupervised Learning

- Log $P(x)$ is a measure of fit in Probabilistic models (GMM, Factor Analysis)
 - High log $P(x)$ on training set, but low log $P(x)$ on test set is a measure of overfitting
 - Raw value of log $P(x)$ hard to interpret in isolation
- K-means is trickier (because of fixed covariance assumption)

Regression

Common metrics:

- MSE: penalizes larger errors more
- MAE: penalizes all errors equally
- (others: Max Error, Mean squared logarithmic error, Mean absolute percentage error, Median absolute error, Pinball loss, ...)

Semantic Segmentation

- Binary Mask prediction: so all binary classification metrics hold

But more important:

- DICE Score: Dice Sørensen Coefficient
 - Equivalent to F1 Score
 - $\frac{2 |A \cap B|}{|A| + |B|}$
- IoU / Jaccard Metric

Thank You!