MAC 0460 / 5832
Introduction to Machine Learning

14 – Performance evaluation

IME/USP (31/05/2021)

Once a hypothesis g is selected,
how do we estimate its performance?

✓ E_{out} ?

For that, we need an independent dataset D_{test}

(D_{test} is a set with samples not in D_{train})

E_{out} can be estimated just as E_{in}, but on D_{test}

Regression → MSE

Classification → cross-entropy

log

E_{in} and E_{out} are related to the loss function

Often, it is convenient to evaluate other performance metrics

Regression

We optimize MSE during training to obtain g

g can be evaluated on D_{test} in terms of

- MSE ✓
- MAE ✓
- coefficient of determination (R squared) ✓

Classification

We optimize the cross-entropy loss during training to obtain g

g outputs a score that can be interpreted as $P(y = 1|\mathbf{x})$

$[0,1]$

An obvious decision would be

$$\hat{y}_{\mathbf{x}} = \begin{cases} 1, & \text{if } g(\mathbf{x}) > 0.5, \\ 0, & \text{if } g(\mathbf{x}) \leq 0.5 \end{cases}$$

Types of errors in binary classification problems

Classes

- Positive
- Negative

Four possible cases:

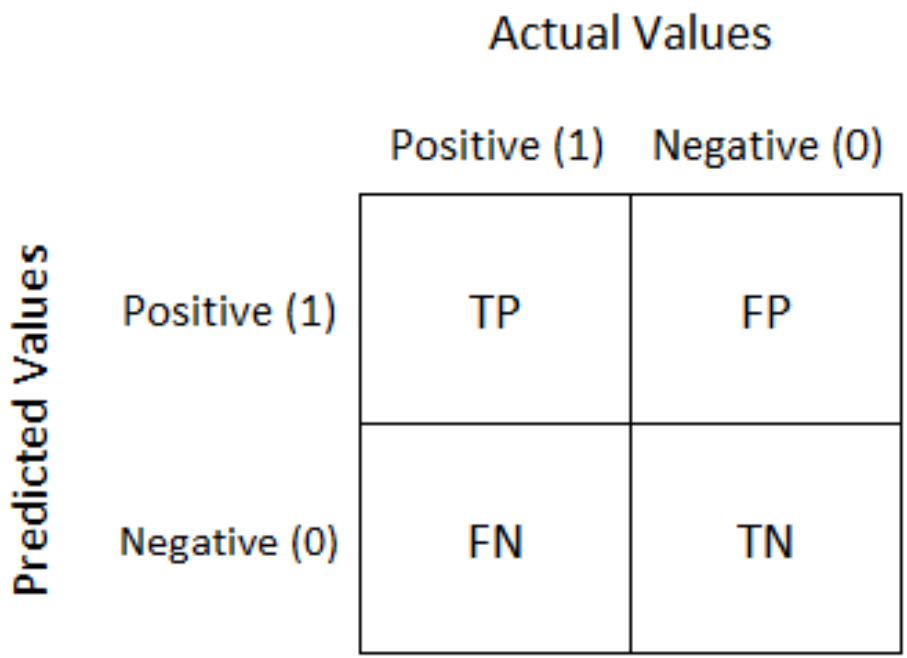
- True-positive (TP)

  $y = 1$ and $\hat{y} = 1$
- False-positive (FP)

  $y = 0$ and $\hat{y} = 1$
- False-negative (FN)

  $y = 1$ and $\hat{y} = 0$
- True-negative (TN)

  $y = 0$ and $\hat{y} = 0$



Metrics derived from TP, FP, TN, FN

Confusion matrix

	True condition			
Total population	Condition positive	Condition negative	Prevalence = $\frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$
Predicted condition positive	**True positive**, Power TP	**False positive**, Type I error FP	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$
Predicted condition negative	**False negative**, Type II error FN	**True negative** TN	False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$
	True positive rate (TPR), Recall, Sensitivity, probability of detection = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) = $\frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) = $\frac{\text{LR+}}{\text{LR−}}$
	False negative rate (FNR), Miss rate = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) = $\frac{\text{FNR}}{\text{TNR}}$	F_1 score = $2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$

Source: Wikipedia

Metrics derived from TP, FP, TN, FN

	True condition				
Total population	Condition positive	Condition negative	Prevalence $= \frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) $= \frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	
Predicted condition positive	**True positive**, Power	**False positive**, Type I error	Positive predictive value (PPV), Precision $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$	
Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$	
	True positive rate (TPR), Recall, Sensitivity, probability of detection $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) $= \frac{\text{LR+}}{\text{LR−}}$	F_1 score $= 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$
	False negative rate (FNR), Miss rate $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) $= \frac{\text{FNR}}{\text{TNR}}$		

Source: Wikipedia

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

Metrics derived from TP, FP, TN, FN

	True condition				
Total population	Condition positive	Condition negative	Prevalence = $\frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	
Predicted condition positive	**True positive**, Power	**False positive**, Type I error	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$	
Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$	
	True positive rate (TPR), Recall, Sensitivity, probability of detection = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) = $\frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) = $\frac{\text{LR+}}{\text{LR−}}$	F_1 score = $2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$
	False negative rate (FNR), Miss rate = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) = $\frac{\text{FNR}}{\text{TNR}}$		

Source: Wikipedia

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

Metrics derived from TP, FP, TN, FN

	True condition				
Total population	Condition positive	Condition negative	Prevalence $= \frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) $= \frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	
Predicted condition positive	**True positive**, Power	**False positive**, Type I error	Positive predictive value (PPV), Precision $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$	
Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$	
	True positive rate (TPR), Recall, Sensitivity, probability of detection $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) $= \frac{\text{LR+}}{\text{LR−}}$	F_1 score $= 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$
	False negative rate (FNR), Miss rate $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) $= \frac{\text{FNR}}{\text{TNR}}$		

Source: Wikipedia

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

Metrics derived from TP, FP, TN, FN

	True condition			
Total population	Condition positive	Condition negative	Prevalence $= \frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) $= \frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$
Predicted condition positive	**True positive**, Power	**False positive**, Type I error	Positive predictive value (PPV), Precision $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$
Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$
	True positive rate (TPR), Recall, Sensitivity, probability of detection $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) $= \frac{\text{LR+}}{\text{LR−}}$
	False negative rate (FNR), Miss rate $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) $= \frac{\text{FNR}}{\text{TNR}}$	F_1 score $= 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$

Source: Wikipedia

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN}$$

Metrics derived from TP, FP, TN, FN

	True condition			
Total population	Condition positive	Condition negative	Prevalence $= \frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) $= \frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$
Predicted condition positive	**True positive**, Power	**False positive**, Type I error	Positive predictive value (PPV), Precision $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$
Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$
	True positive rate (TPR), Recall, Sensitivity, probability of detection $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) $= \frac{\text{LR+}}{\text{LR−}}$
	False negative rate (FNR), Miss rate $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) $= \frac{\text{FNR}}{\text{TNR}}$	F_1 score $= 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$

Source: Wikipedia

$$\text{F1-score} = 2\frac{Precision * Recall}{Precision + Recall}$$

TP, FP, TN, FN in multiclass classifictaion problems?

Let TP_j, FP_j, TN_j, FN_j, for each j (class j against the rest)

Micro-averaging

- Compute $TP = \sum TP_j$, $FP = \sum FP_j$, $TN = \sum TN_j$, $FN = \sum FN_j$
- Compute the performance metrics from TP, FP, TN, FN

Macro-averaging

- Compute the performance metrics for each class, from TP_j, FP_j, TN_j, FN_j
- Compute the mean of each metric

TP, FP, TN, FN in multiclass classifictaion problems?

Let TP_j, FP_j, TN_j, FN_j, for each j (class j against the rest)

Micro-averaging

- Compute $TP = \sum TP_j$, $FP = \sum FP_j$, $TN = \sum TN_j$, $FN = \sum FN_j$
- Compute the performance metrics from TP, FP, TN, FN
- assigns same importance to all examples $\rightsquigarrow$ larger classes dominate

Macro-averaging

- Compute the performance metrics for each class, from TP_j, FP_j, TN_j, FN_j
- Compute the mean of each metric
- assigns same importance to all classes

(There is no consensus about which is the right one)

TP, FP, TN, FN depend on the threshold T

$$\hat{y} = \begin{cases} 1, & \text{if } g(\mathbf{x}) > 0.5, \\ 0, & \text{if } g(\mathbf{x}) \leq 0.5 \end{cases}$$

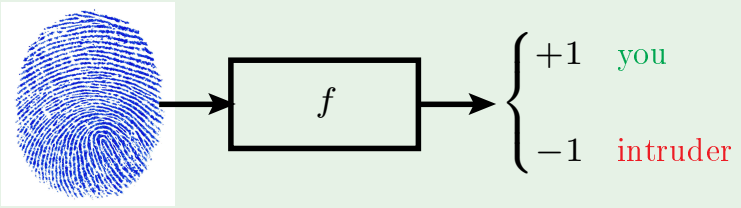
One can choose other values than 0.5 for the threshold T

The error measure - for supermarkets

Supermarket verifies fingerprint for discounts

False reject is costly; customer gets annoyed!

False accept is minor; gave away a discount and intruder left their fingerprint



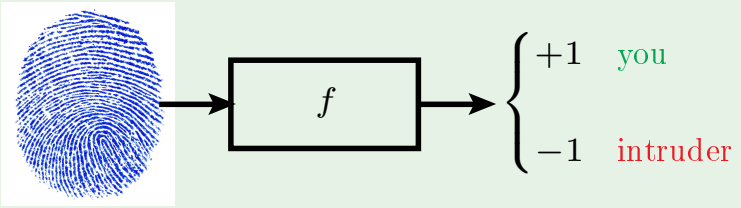
		f	
		$+1$	-1
h	$+1$	0	1
	-1	10	0

The error measure - for the CIA

CIA verifies fingerprint for security

False accept is a disaster!

False reject can be tolerated
Try again; you are an employee ☺



		f	
		$+1$	-1
h	$+1$	0	1000
	-1	1	0

supermarket

~~Finger print~~ case

Better false accept (FP) than false reject (FN)

Threshold T : 0.4

positive $\Leftrightarrow \underbrace{p(y=1/x)}_{g(x)} > T$

CIA case

Better false reject (FN) than false accept (FP)

Threshold T : 0.9

Finger print case

Better false accept (FP) than false reject (FN)

Threshold T : small

CIA case

Better false reject (FN) than false accept (FP)

Threshold T : large

Often we would like to maximize the true positives (TP)

$\longrightarrow$ high recall (TPR)

At the same time we would like to minimize the false positives (FP)

$\longrightarrow$ small FPR (high precision)

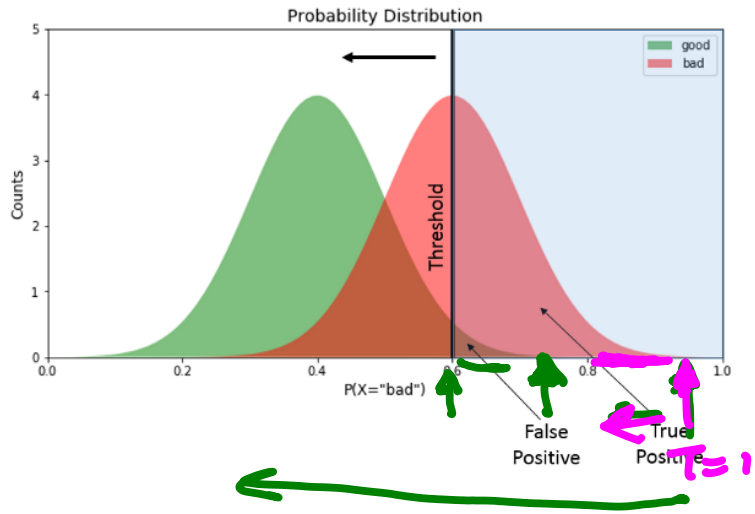
ROC curve and **PR curve** (shown next) are often used as tools to assess recall and precision simultanesously

ROC Curve

ROC : Receiver operating characteristic

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

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



$$TPR = \frac{True\ Positive}{Total\ Positive}$$

$$FPR = \frac{False\ Positive}{Total\ Negative}$$

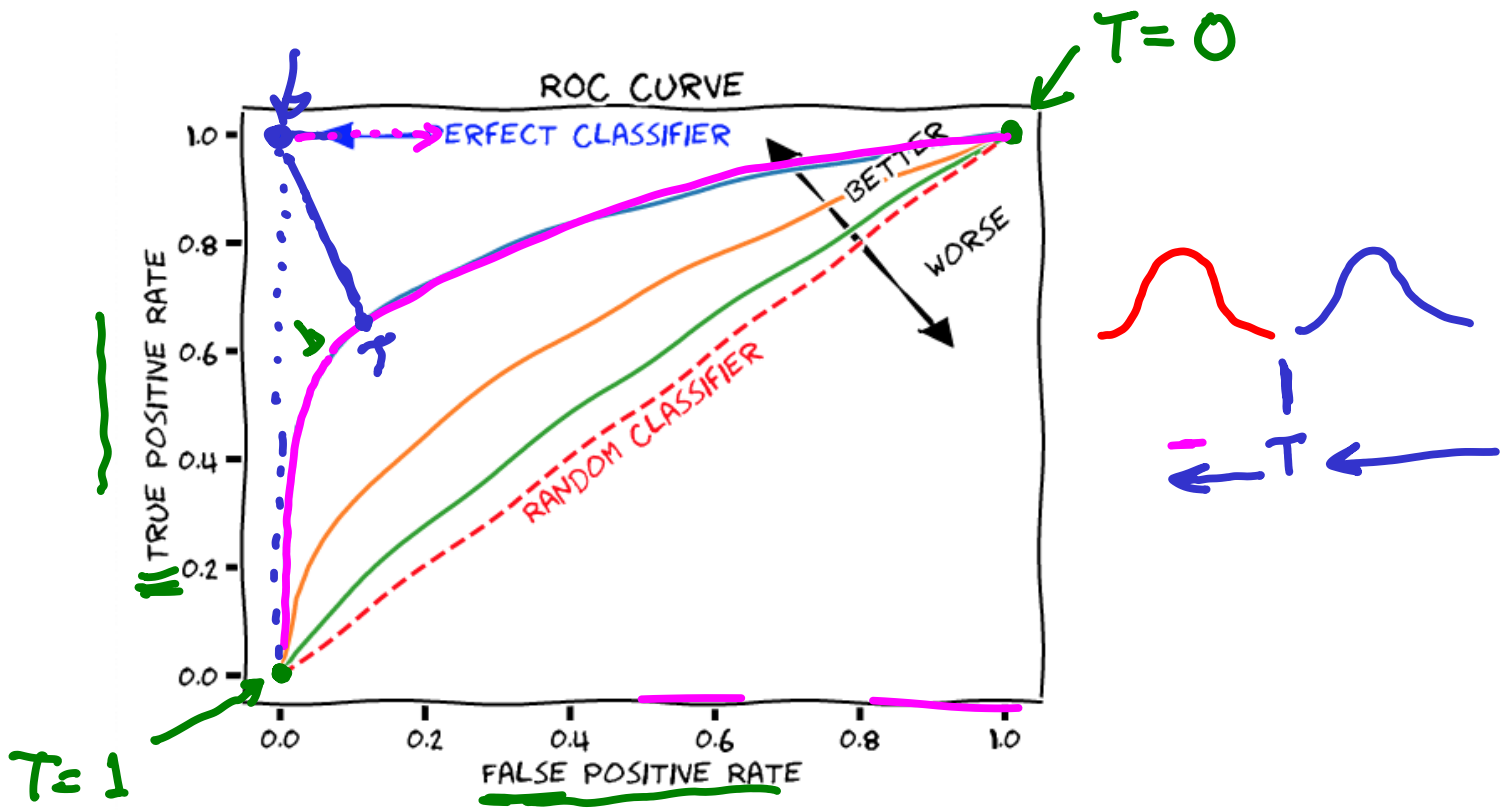
www.kdnuggets.com/2018/07/receiver-operating-characteristic-curves-demystified-python.html

$T = 1.0 \implies$ all inputs are classified as negative (TP=0% and FP=0%)

$T = 0.0 \implies$ all inputs are classified as positive (TP=100% and FP=100%)

As we vary T from 1.0 to 0.0

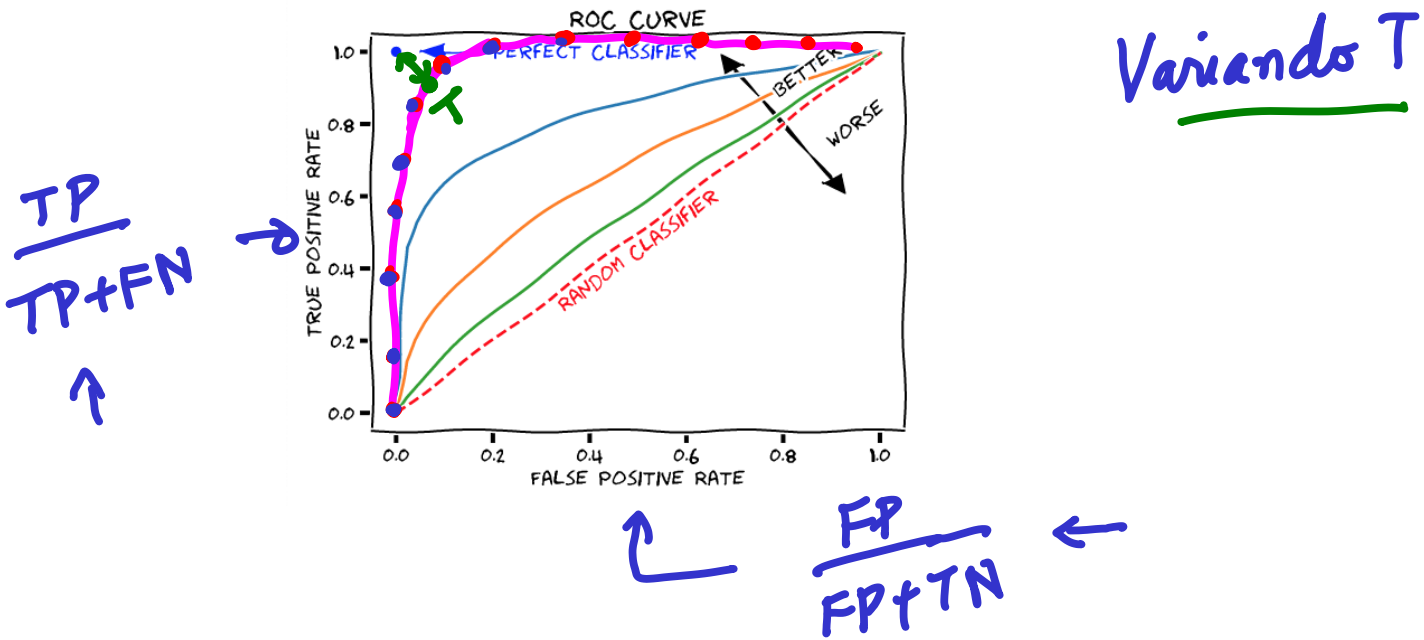
- **Perfectly separated classes**: TP will reach 100% while FP stays at 0%, and only after that FP will start to increase
- **General case**: TP will start to increase but so does FP too.



analyticsindiamag.com/beginners-guide-to-understanding-roc-curve-how-to-find-the-perfect-probability-threshold/

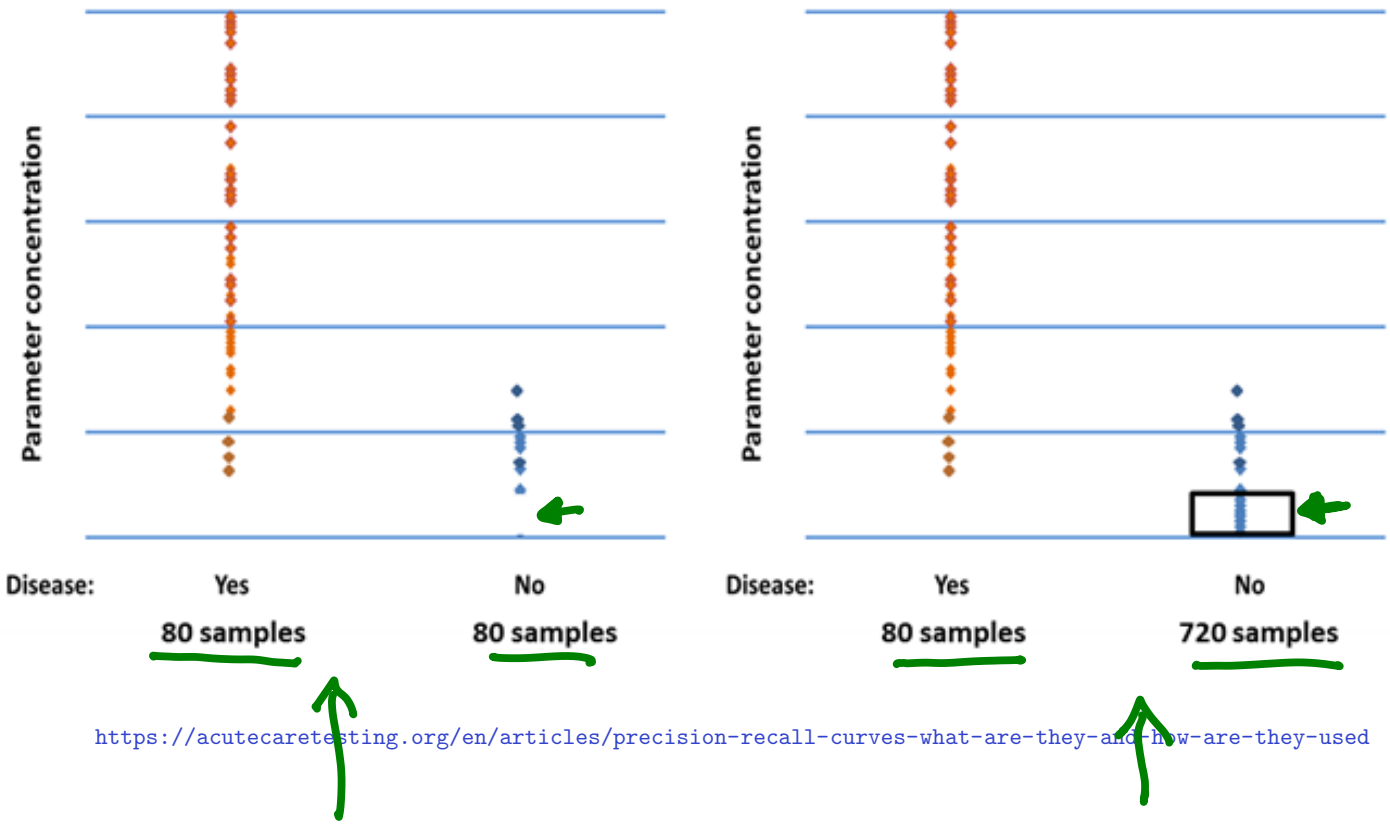
AUC ROC (area under the ROC curve)

It is often used as a performance metric
AUC vary from 0.5 to 1.0
The closer the AUC to 1.0, the better the classifier



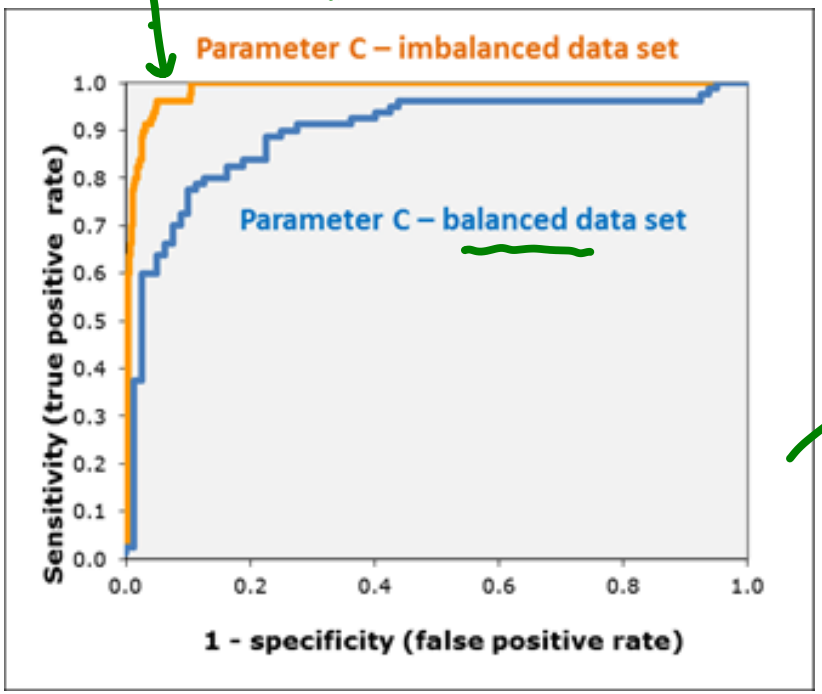
But AUC may be misleading when the dataset is unbalanced

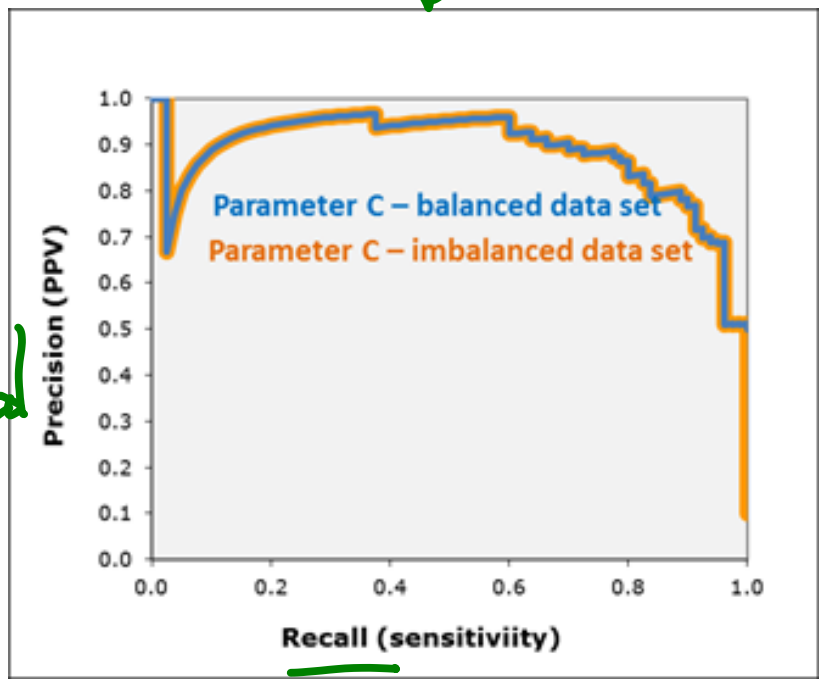
Example: When classes are highly unbalanced (right side graph has much more negatives)



https://acutecaretesting.org/en/articles/precision-recall-curves-what-are-they-and-how-are-they-used

ROC $\times$ PR (precison-recall) curves





https://acutecaretesting.org/en/articles/precision-recall-curves-what-are-they-and-how-are-they-used

$$TPR = \frac{TP}{TP+FN} \qquad FPR = \frac{FP}{FP+TN} \qquad \text{Precision} = \frac{TP}{TP+FP}$$

Summary

There is no single best metric for evaluating a hypothesis

In practice, how do we chosse a hypothesis ?

Which hypothesis space should we use?

Which algorithm should we use ?

How should hyperparameters be tuned ?

In practice we often see three types of datasets

They are used to estimate the hypothesis error

Training — E_{in} ✔

Validation — E_{val} ✔

Test — E_{test} ✔

They are computed at three different moments

- During training we compute E_{in}
- For hypothesis selection we compute E_{val}
- For estimating E_{out} of the selected model we compute E_{test}

Both E_{val} and E_{test} are estimations of E_{out}

- E_{val} is optimistically biased
- E_{test} is unbiased

Estimating E_{out}

Partition the existing dataset into two subsets:

$$D = D_{train} \cup D_{test}$$

D_{train} D_{test}

D_{train} is used for training and for computing E_{in}

D_{test} is used to compute E_{test}, an unbiased estimate of E_{out}

Potential problems of E_{test} as estimator of E_{out}

Let the size of the sets be:

- $|D_{train}|$
- $|D_{test}|$
- $|D| = |D_{train}| + |D_{test}|$

- large $|D_{test}| \rightsquigarrow$ small $|D_{train}|$ (hypothesis is no good)
- small $|D_{test}| \rightsquigarrow E_{test}$ hardly will be a good estimate of E_{out}
- when D_{train} and/or D_{test} have some bias

Problems when we use E_{test} for selecting a final hypothesis

E_{test} of the chosen hypothesis is no longer unbiased; it is an optimistically biased estimator of E_{out}

We will see later that:

Validation sets are commonly used for choosing a final hypothesis from the set of pre-selected hypotheses, based on E_{val}

To reduce bias of E_{val}, cross-validation methods are commonly used

Test set is used to get an unbiased estimate of E_{out} (if D_{test} is available ....)

No class

June 2nd

an

June 7th.