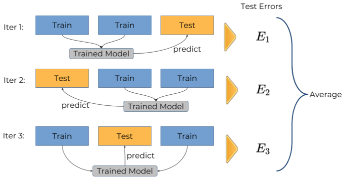
Introduction to Machine Learning

Evaluation: Resampling

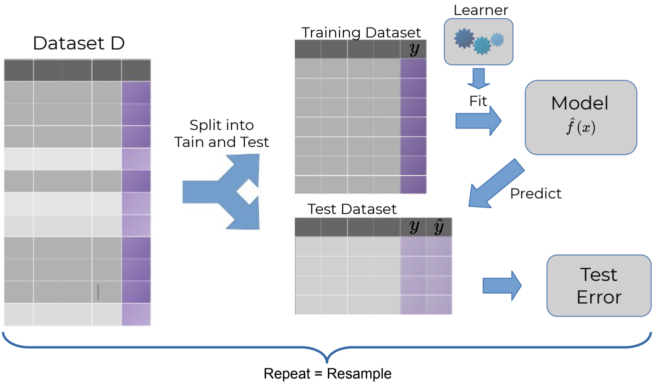


Learning goals

- Understand how resampling techniques extend the idea of simple train-test splits
- Understand the ideas of cross-validation, bootstrap and subsampling
- Understand what pessimistic bias means

RESAMPLING

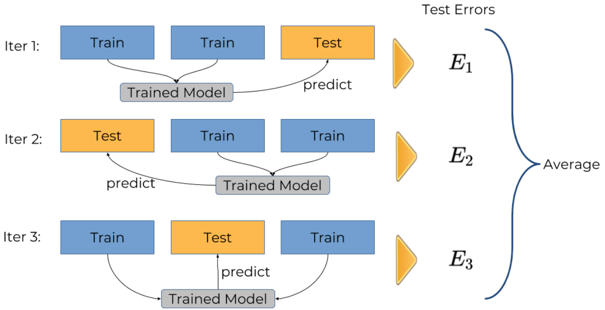
- Aim: Assess performance of learning algorithm.
- Make training sets large (to keep the pessimistic bias small), and reduce variance introduced by smaller test sets through many repetitions / averaging of results.



CROSS-VALIDATION

- Split the data into k roughly equally-sized partitions.
- Use each part once as test set and join the $k-1$ others for training
- Obtain k test errors and average.

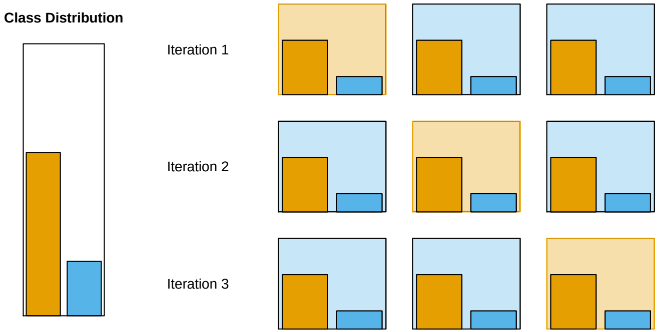
Example: 3-fold cross-validation:



CROSS-VALIDATION - STRATIFICATION

Stratification tries to preserve the distribution of the target class (or any specific categorical feature of interest) in each fold.

Example of stratified 3-fold cross-validation:

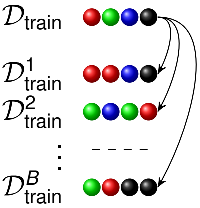


CROSS-VALIDATION

- 5 or 10 folds are common
- $k = n$ is known as leave-one-out (LOO) cross-validation
- Estimates of the generalization error tend to be pessimistically biased
  size of the training sets is $n - (n/k) < n$
  bias increases as k gets smaller.
- The k performance estimates are dependent because of the structured overlap of the training sets.
  $\Rightarrow$ Variance of the estimator increases for very large k (close to LOO), when training sets nearly completely overlap.
- Repeated k-fold CV (multiple random partitions) can improve error estimation for small sample sizes.

BOOTSTRAP

The basic idea is to randomly draw B training sets of size n with replacement from the original training set $\mathcal{D}_{\text{train}}$:



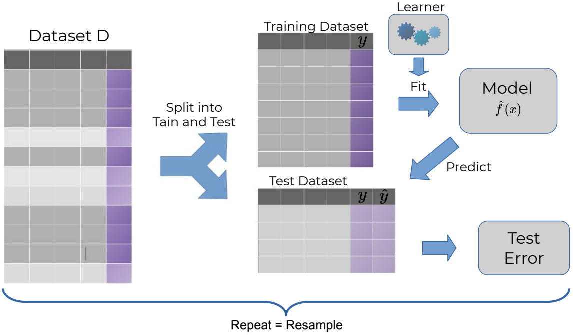
We define the test set in terms of out-of-bag observations $\mathcal{D}_{\text{test}}^{b} = \mathcal{D}_{\text{train}} \setminus \mathcal{D}_{\text{train}}^{b}$.

BOOTSTRAP

- Typically, B is between 30 and 200.
- The variance of the bootstrap estimator tends to be smaller than the variance of k-fold CV.
- The more iterations, the smaller the variance of the estimator.
- Tends to be pessimistically biased (because training sets contain only about 63.2% of the unique observations).
- Bootstrapping framework allows for inference (e.g. detect significant performance differences between learners).
- Extensions exist for very small data sets that also use the training error for estimation: B632 and B632+.

SUBSAMPLING

- Repeated hold-out with averaging, a.k.a. Monte Carlo CV
- Similar to bootstrap, but draws without replacement
- Typical choices for splitting: $4/5$ or $9/10$ for training



- The smaller the subsampling rate, the larger the pessimistic bias.
- The more subsampling repetitions, the smaller the variance.

RESAMPLING DISCUSSION

In ML we fit, at the end, a model on all our given data.

Problem: We need to know how well this model performs in the future, but no data is left to reliably do this.
$\Rightarrow$ Approximate using hold-out / CV / bootstrap / resampling estimate

But: pessimistic bias because we don't use all data points

Final model is (usually) computed on all data points.

RESAMPLING DISCUSSION

- 5CV or 10CV have become standard
- Do not use hold-out, CV with few iterations, or subsampling with a low subsampling rate for small samples, since this can cause the estimator to be extremely biased, with large variance.
- If $n < 500$, use repeated CV
- A $\mathcal{D}$ with $|\mathcal{D}| = 100.000$ can have small-sample properties if one class has few observations
- Research indicates that subsampling has better properties than bootstrapping. The repeated observations can cause problems in training.

Einführung in das Statistische Lernen

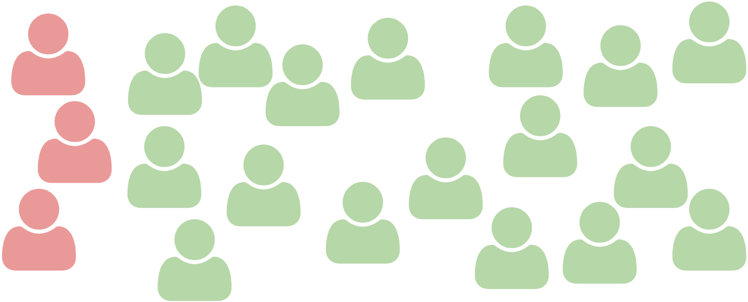
Evaluation: Measures for Binary Classification: ROC Measures

		Actual Class y		
		Positive	Negative	
$\hat{y}$ **Pred.**	Positive	**True Positive** (TP) = 20	**False Positive** (FP) = 180	Positive predictive value = TP / (TP + FP) = 20 / (20 + 180) = **10%**
	Negative	**False Negative** (FN) = 10	**True Negative** (TN) = 1820	Negative predictive value = TN / (FN + TN) = 1820 / (10 + 1820) ≈ **99.5%**
		True Positive Rate = TP / (TP + FN) = 20 / (20 + 10) ≈ **67%**	True Negative Rate = TN / (FP + TN) = 1820 / (180 + 1820) = **91%**	

Learning goals

- Understand why accuracy is not an optimal performance measure for imbalanced labels
- Understand the different measures computable from a confusion matrix
- Be aware that each of these measures has a variety of names

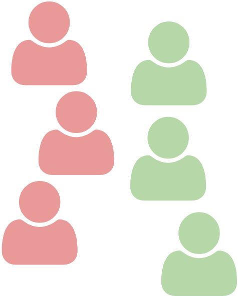
IMBALANCED BINARY LABELS



Classify all as “no disease” (green) $\rightarrow$ high accuracy.

Accuracy Paradox

IMBALANCED COSTS



Classify incorrectly as “no disease” $\rightarrow$ very high cost

CONFUSION MATRIX

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

- $+$: "positive" class
- $-$: "negative" class
- n_+: number of observations in $+$
- n_-: number of observations in $-$

LABELS: ROC METRICS

From the confusion matrix (binary case), we can calculate "ROC" metrics.

		True Class y		
		$+$	$-$	
Pred.	$+$	TP	FP	PPV = $\frac{\text{TP}}{\text{TP}+\text{FP}}$
$\hat{y}$	$-$	FN	TN	NPV = $\frac{\text{TN}}{\text{FN}+\text{TN}}$
		TPR = $\frac{\text{TP}}{\text{TP}+\text{FN}}$	TNR = $\frac{\text{TN}}{\text{FP}+\text{TN}}$	Accuracy = $\frac{\text{TP}+\text{TN}}{\text{TOTAL}}$

- True Positive Rate: How many of the true 1s did we predict as 1?
- True Negative Rate: How many of the true 0s did we predict as 0?
- Positive Predictive Value: If we predict 1 how likely is it a true 1?
- Negative Predictive Value: If we predict 0 how likely is it a true 0?

HISTORY ROC

ROC = receiver operating characteristics

Initially developed by electrical engineers and radar engineers during World War II for detecting enemy objects in battlefields.

http://media.iwm.org.uk/iwm/mediaLib//39/media-39665/large.jpg

Still has the funny name.

LABELS: ROC

Example

		Actual Class y		
		Positive	Negative	
$\hat{y}$ **Pred.**	Positive	**True Positive** (TP) = 20	**False Positive** (FP) = 180	Positive predictive value = TP / (TP + FP) = 20 / (20 + 180) = **10%**
	Negative	**False Negative** (FN) = 10	**True Negative** (TN) = 1820	Negative predictive value = TN / (FN + TN) = 1820 / (10 + 1820) ≈ **99.5%**
		True Positive Rate = TP / (TP + FN) = 20 / (20 + 10) ≈ **67%**	True Negative Rate = TN / (FP + TN) = 1820 / (180 + 1820) = **91%**	

MORE METRICS AND ALTERNATIVE TERMINOLOGY

Unfortunately, for many concepts in ROC, 2-3 different terms exist.

		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	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 $= \frac{1}{\frac{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}}{2}}$
		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−) $= \frac{\text{FNR}}{\text{TNR}}$		

▸ Clickable version/picture source

▸ Interactive diagram

LABELS: F_1-MEASURE

A measure that balances two conflicting goals

1. Maximising Positive Predictive Value
2. Maximising True Positive Rate

is the harmonic mean of PPV and TPR:

$$F_1 = 2\frac{PPV \cdot TPR}{PPV + TPR}$$

Note: still doesn't account for the number of true negatives.

LABELS: F_1-MEASURE

Tabulated F_1-Score for different TPR (rows) and PPV (cols) combinations.

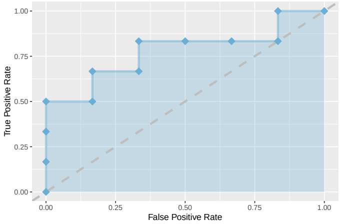
	0.0	0.2	0.4	0.6	0.8	1.0
0.0	0	0.00	0.00	0.00	0.00	0.00
0.2	0	0.20	0.27	0.30	0.32	0.33
0.4	0	0.27	0.40	0.48	0.53	0.57
0.6	0	0.30	0.48	0.60	0.69	0.75
0.8	0	0.32	0.53	0.69	0.80	0.89
1.0	0	0.33	0.57	0.75	0.89	1.00

$\rightarrow$ Tends more towards the lower of the 2 combined values.

- $TPR = 0$ or $PPV = 0 \Rightarrow F_1$ of 0
- Predicting always "neg": $F_1 = 0$
- Predicting always "pos": $F_1 = 2PPV/(PPV + 1) = 2n_+/(n_+ + n)$, which will be rather small, if the size of the positive class n_+ is small.

Einführung in das Statistische Lernen

Evaluation: Measures for Binary Classification: ROC Visualization

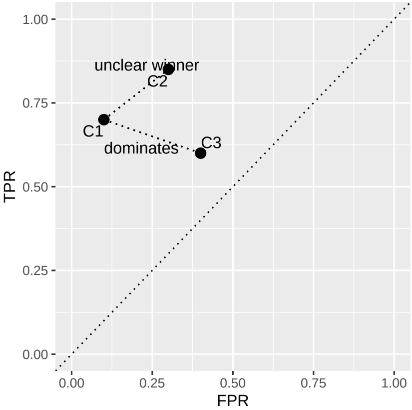


Learning goals

- Understand the ROC curve
- Be able to compute a ROC curve manually
- Understand the definition of AUC and what a certain value of AUC means (and what not!)

LABELS: ROC SPACE

Plot True Positive Rate and False Positive Rate:



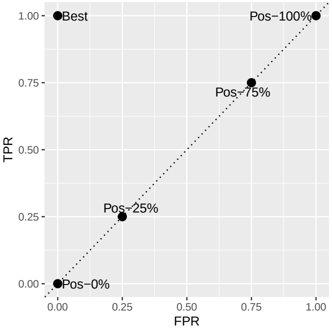
		True Class y	
		$+$	$-$
Pred.	$+$	TP	FP
$\hat{y}$	$-$	FN	TN

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

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

LABELS: ROC SPACE

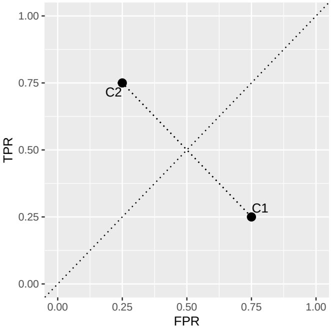
- The best classifier lies on the top-left corner
- The diagonal $\approx$ random labels (with different proportions).
  Assign positive *x* as "pos" with 25% probability $\rightarrow$ $TPR = 0.25$.
  Assign negative *x* as "pos" with 25% probability $\rightarrow$ $FPR = 0.25$.



LABELS: ROC SPACE

In practice, we should never obtain a classifier below the diagonal.

Inverting the predicted labels ($0 \to 1$ and $1 \to 0$) will result in a reflection at the diagonal.



LABEL DISTRIBUTION IN TPR AND FPR

TPR and FPR are insensitive to the class distribution:
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

MCE = 35/100
$TPR = 0.8$
$FPR = 0.5$

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

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

MCE = 45/150 = 30/100
$TPR = 0.8$
$FPR = 0.5$

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

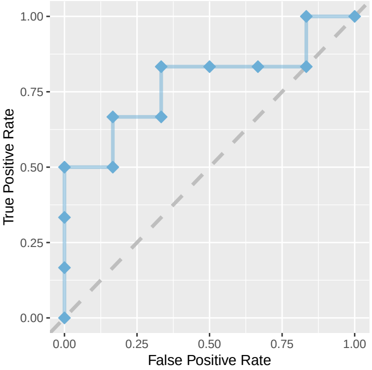
FROM PROBABILITIES TO LABELS: ROC CURVE

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

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

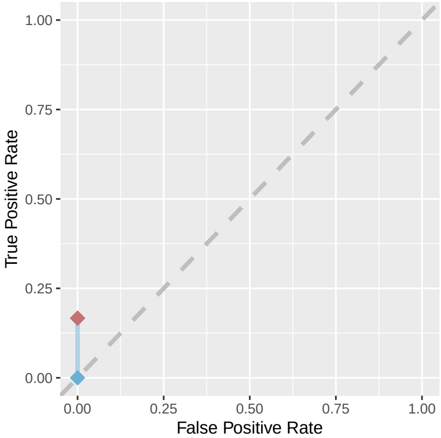
To draw a ROC curve:
Iterate through all possible thresholds c

$\rightarrow$ Visual inspection of all possible thresholds / results



ROC CURVE

#	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



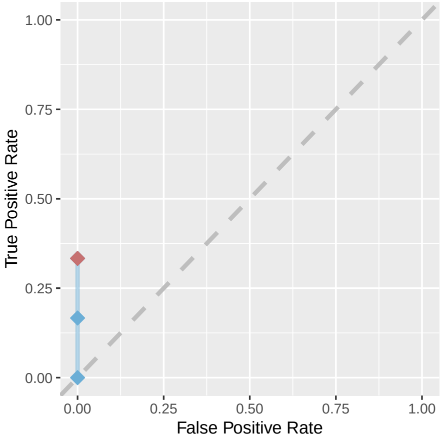
$c = 0.9$

$\rightarrow$ TPR = 0.167

$\rightarrow$ FPR = 0

ROC CURVE

#	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



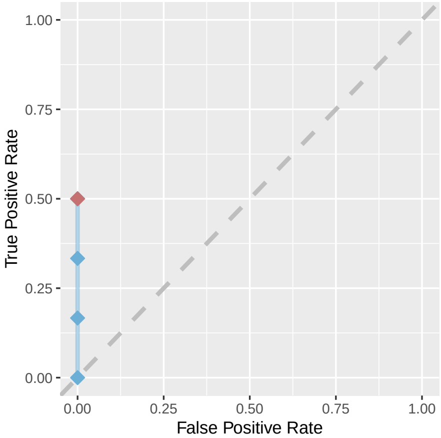
$c = 0.85$

$\rightarrow$ TPR = 0.333

$\rightarrow$ FPR = 0

ROC CURVE

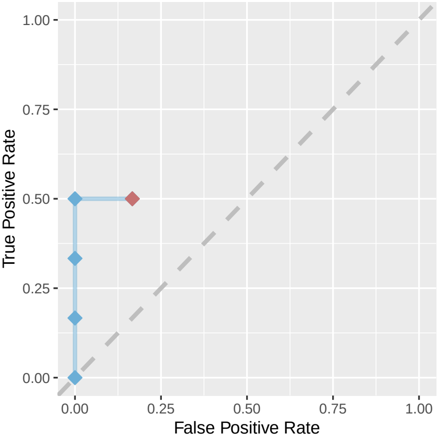
#	Truth	Score
1	Pos	0.95
2	Pos	0.86
3	Pos	0.69
4	Neg	0.65
5	Pos	0.59
6	Neg	0.52
7	Pos	0.51
8	Neg	0.39
9	Neg	0.28
10	Neg	0.18
11	Pos	0.15
12	Neg	0.06



$c = 0.66$
$\rightarrow$ TPR = 0.5
$\rightarrow$ FPR = 0

ROC CURVE

#	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



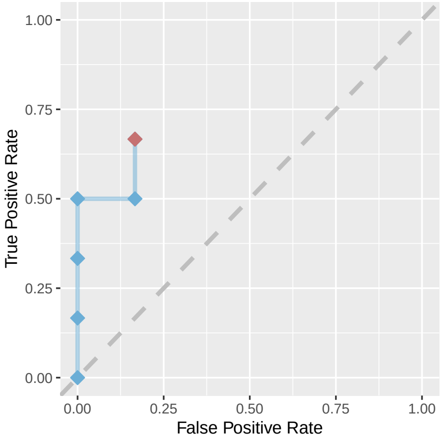
$c = 0.6$

$\rightarrow$ TPR = 0.5

$\rightarrow$ FPR = 0.167

ROC CURVE

#	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



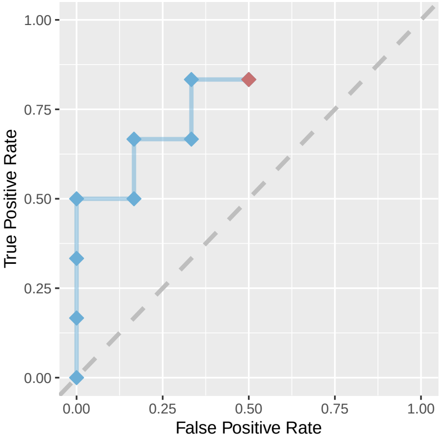
$c = 0.55$

$\rightarrow$ TPR = 0.667

$\rightarrow$ FPR = 0.167

ROC CURVE

#	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



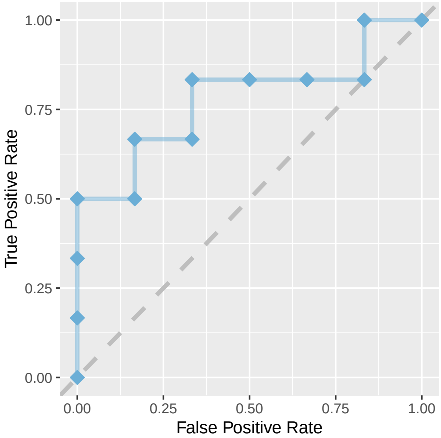
$c = 0.3$

$\rightarrow$ TPR = 0.833

$\rightarrow$ FPR = 0.5

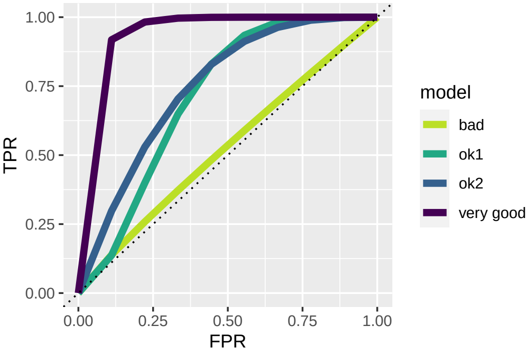
ROC CURVE

#	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



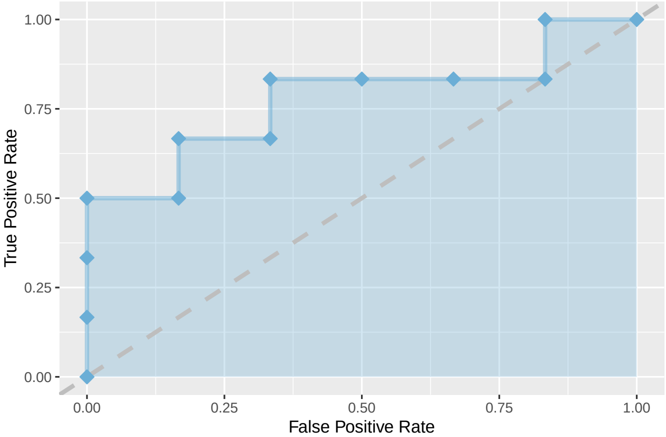
ROC CURVE

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



AUC: AREA UNDER ROC CURVE

- The AUC (in [0,1]) is a single metric to evaluate scoring classifiers
- AUC = 1: Perfect classifier
- AUC = 0.5: Randomly ordered

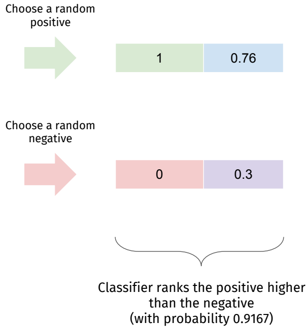


AUC: AREA UNDER ROC CURVE

Interpretation: Probability that classifier ranks a random positive higher than a random negative observation

Truth	Score
1	0.9
1	0.76
1	0.7
0	0.5
1	0.45
0	0.3
0	0.1

$\underbrace{\hspace{6em}}_{\text{AUC = 0.9167}}$



PARTIAL AUC

- Sometimes it can be useful to look at a specific region under the ROC curve $\Rightarrow$ partial AUC (pAUC).
- Examples: focus on a region with low FPR or a region with high TPR: