SWINBURNE UNIVERSITY OF TECHNOLOGY

Model evaluation and data partitioning

SWINBURNE UNIVERSITY OF TECHNOLOGY

Model evaluation and data partitioning

Evaluation methods

- Confusion matrix
- Clustering assessment metrics
- Accuracy, Precision, Recall, F1 measure
- Cross validation

Model evaluation

Will my model betray me?

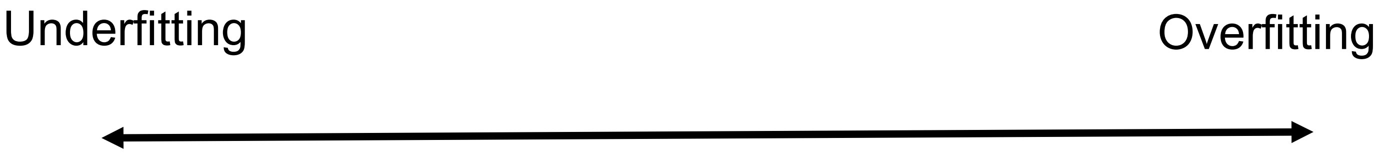
Is my model really good?

- My model shows an accuracy of 90% in the **training environment**
- Would the model be 90% accurate in **production environment**?

Generalization

- A predictive model should be able to handle any dataset coming from the same distribution as the training set

- Generalization refers to a model’s ability to handle any random variations of training data



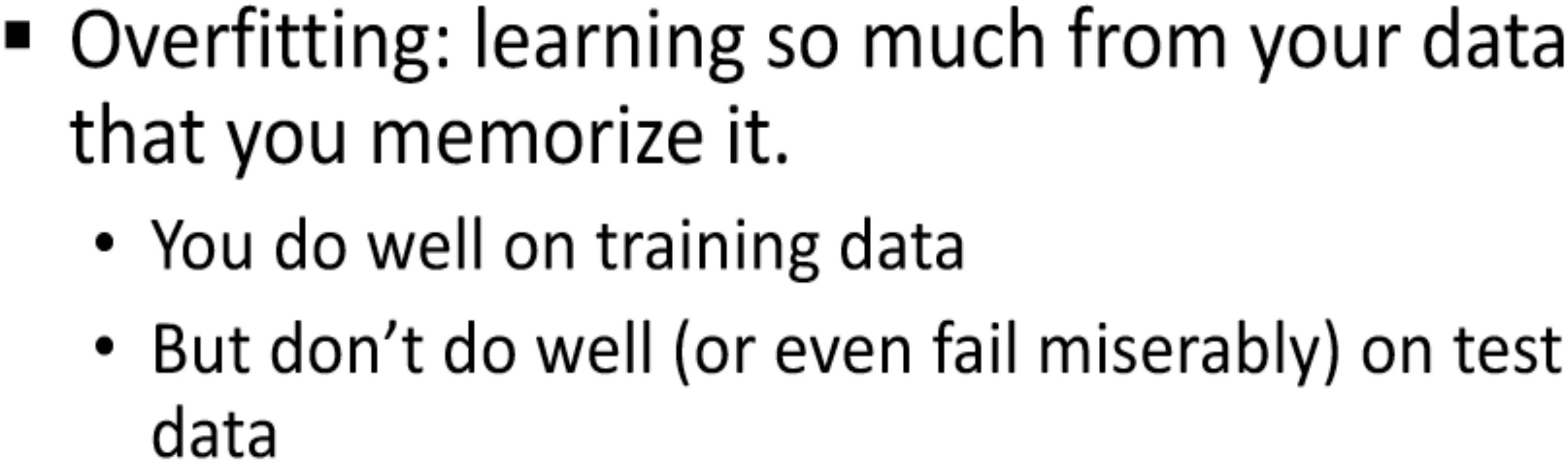
Underfitting

Overfitting

Underfitting (one extreme)

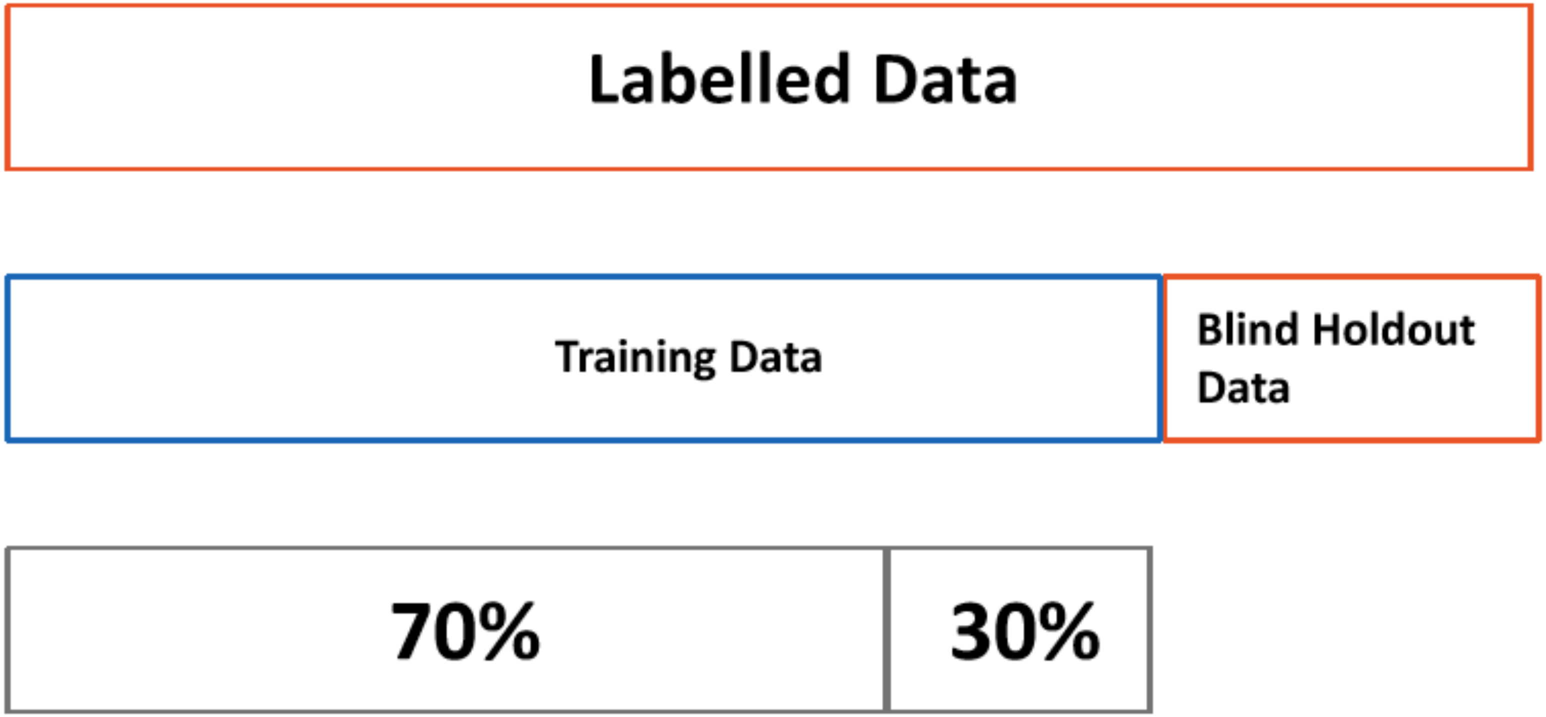
- Underfitting: Not learning enough from the data. An underfitting model does not fit the data at well at all
  - Underfitting models do not capture the underlying trends or patterns in the data
  - Both the training and test set prediction measures are pretty bad

Overfitting (Other extreme)



- Overfitting: learning so much from your data that you memorize it.
  - You do well on training data
  - But don’t do well (or even fail miserably) on test data

Train/Test partition is not enough



Labelled Data

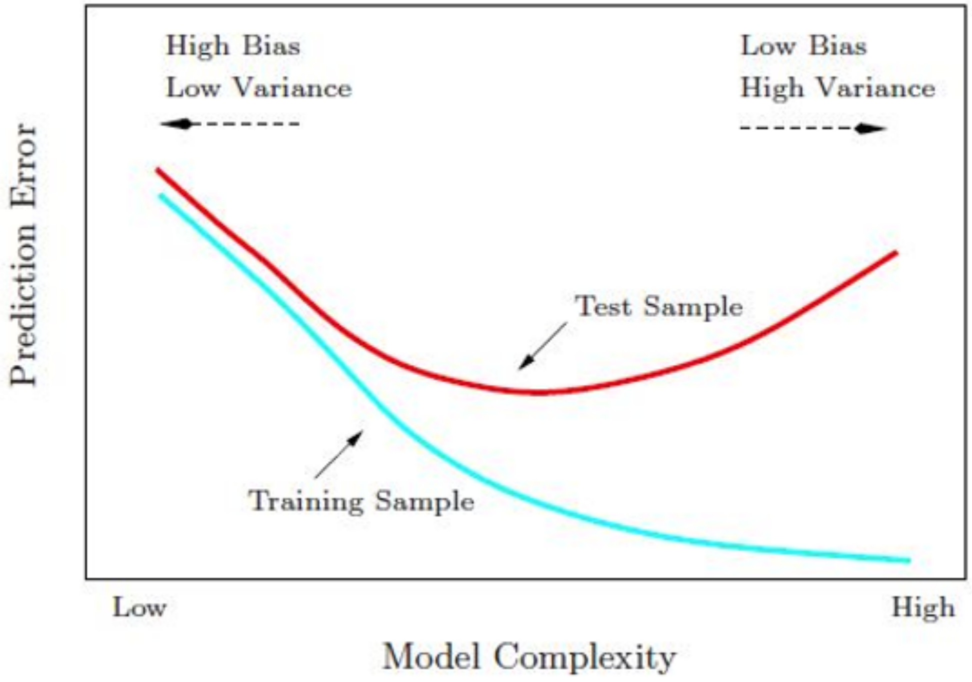
Training Data

Blind Holdout Data

70%

30%

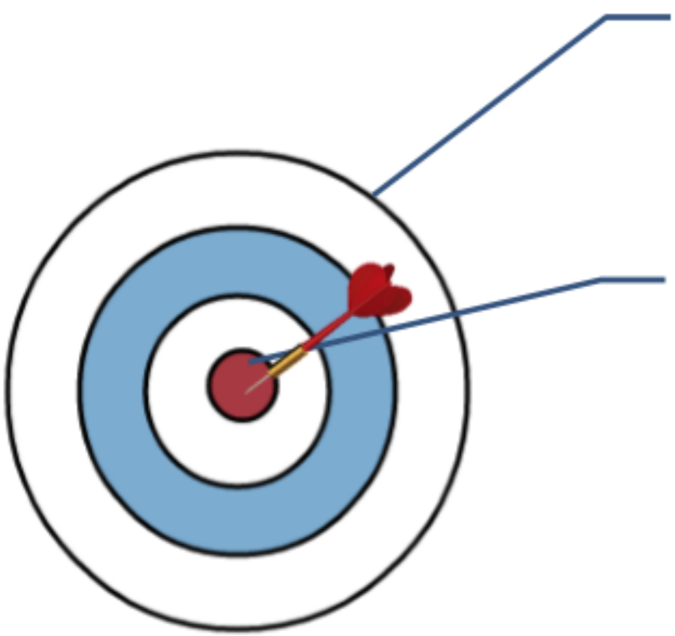
Bias/Variance Trade-off



- Bias is the is an error from erroneous assumptions in the learning algorithm. High bias can cause an algorithm to miss the relevant relations between features and target outputs (underfitting).

- Variance is the error from sensitivity to small fluctuations in the training set. High variance can cause an algorithm to model the random noise in the training data, rather than the intended outputs (overfitting).

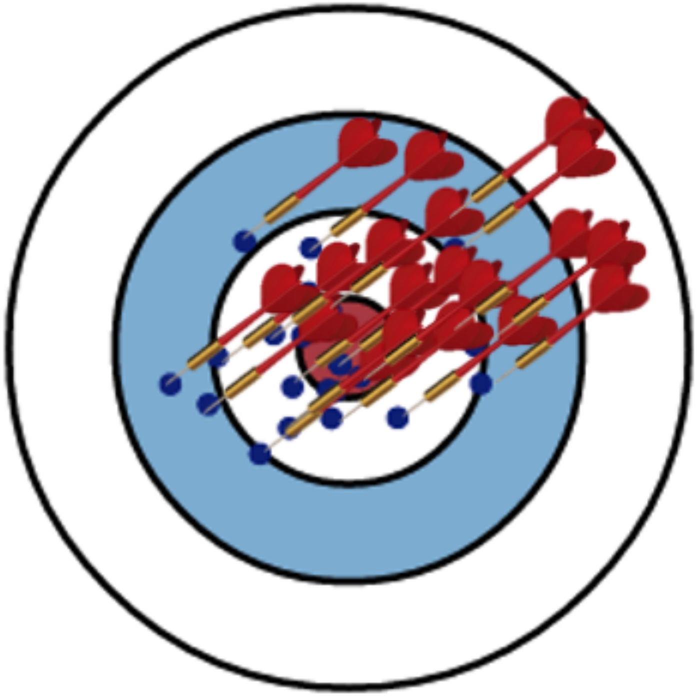
Bias/Variance Trade-off



Each dartboard represents a model

Bullseye is the theoretical best performance (accuracy, precision, recall or something else)

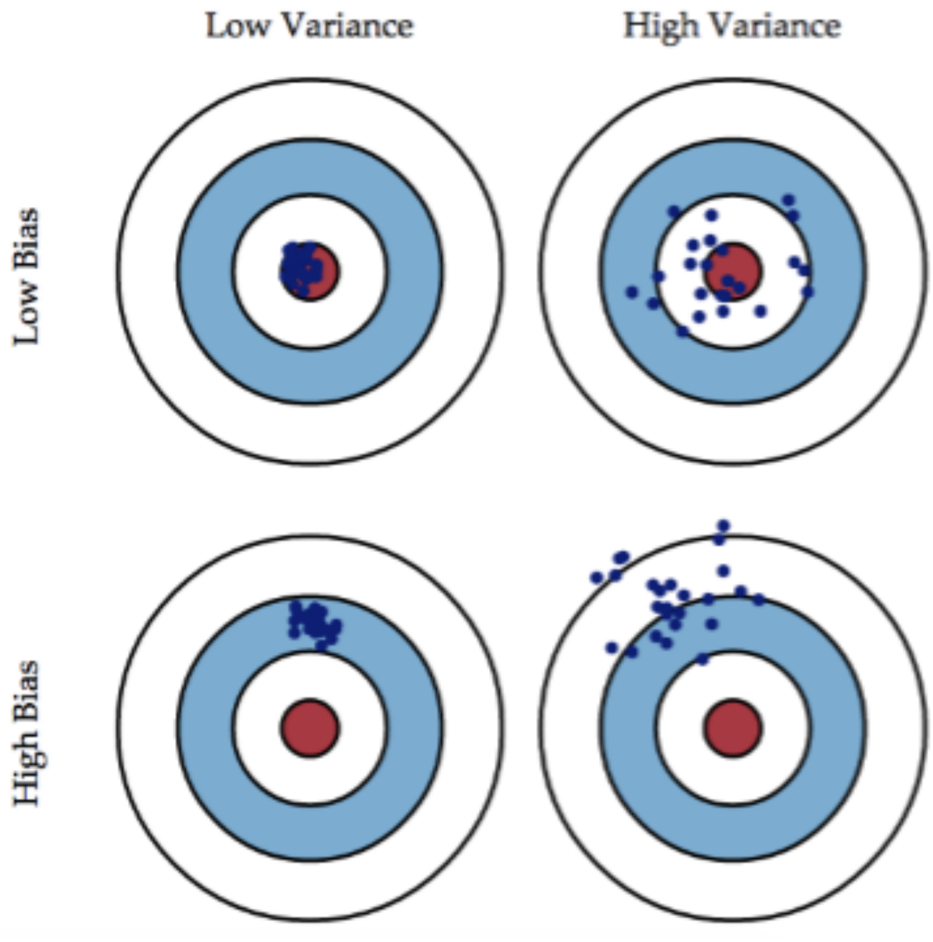
Bias/Variance Trade-off



- Test your model on several variations of the dataset

- Each dot represents a random variation of the test data set

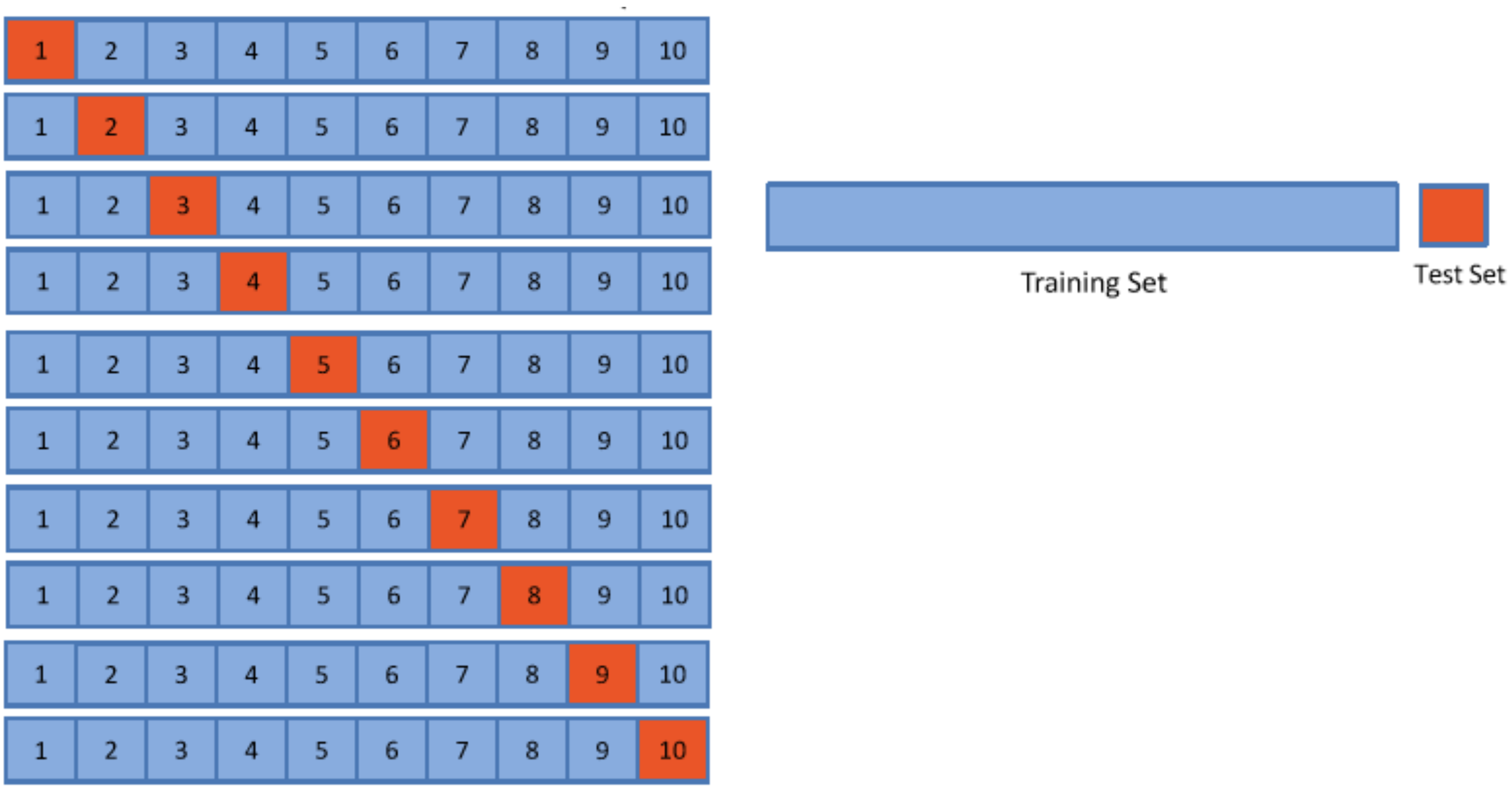
Bias/Variance Trade-off



Cross validation

- Split data into k disjoint partitions
- Train on k-1 partitions and test on 1
- Repeat k times

Cross validation



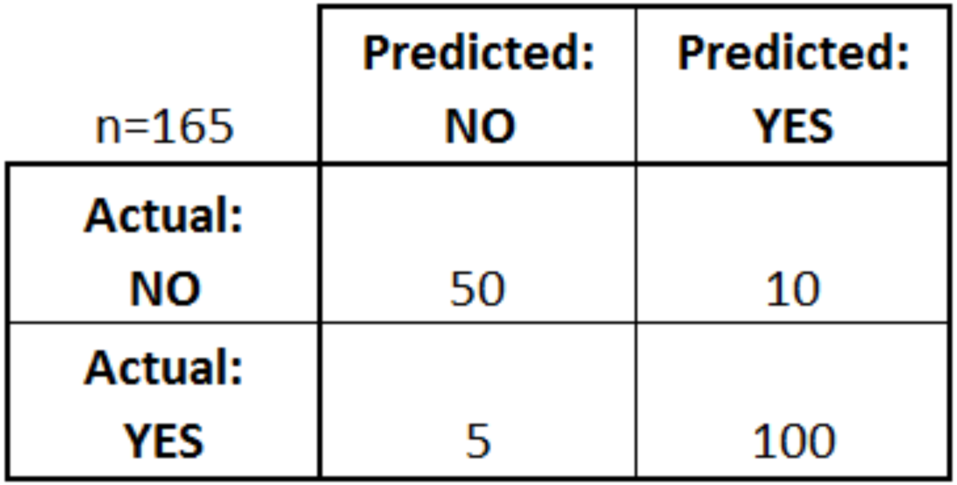
Stratified sampling

- Use when class distribution is skewed
- Ensures that all partitions have fixed ratio of classes
  - Same ratio as training set
  - If training set is 5% class 1, 95% class 2, so is each partition

Evaluation - Confusion matrix

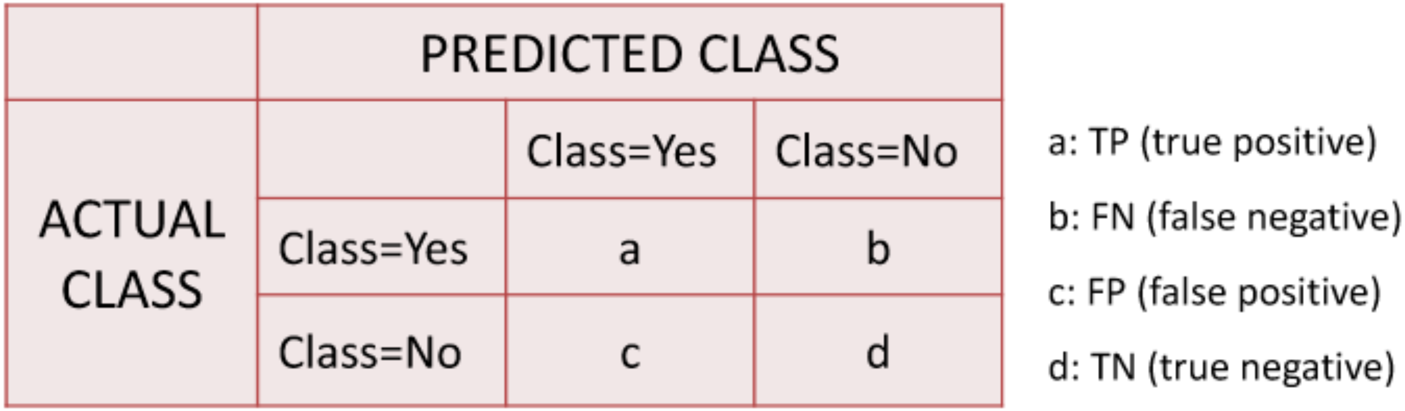
A confusion matrix is a table that is often used to **describe the performance of a classification model** (or "classifier") on a set of test data for which the true values are known. The confusion matrix itself is relatively simple to understand, but the related terminology can be confusing.

Lets assume the following is a confusion matrix of a classifier that predicts whether a patient has cancer or not



n=165	Predicted: NO	Predicted: YES
Actual: NO	50	10
Actual: YES	5	100

Evaluation - Confusion matrix



	PREDICTED CLASS		
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	a	b
	Class=No	c	d

a: TP (true positive)

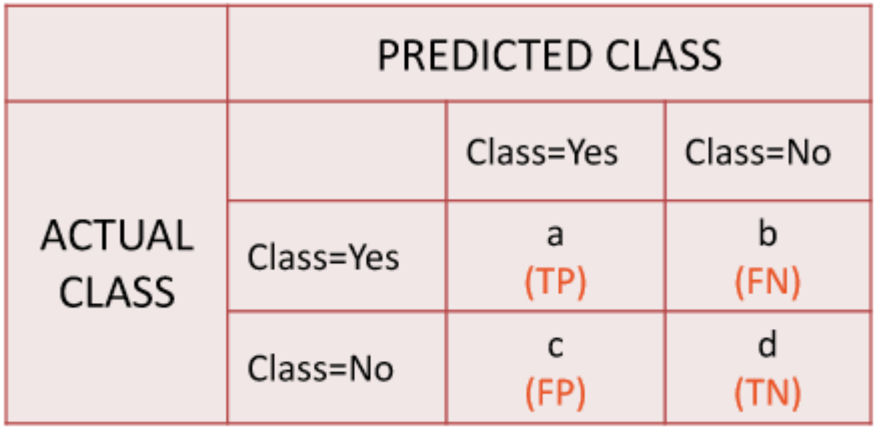
b: FN (false negative)

c: FP (false positive)

d: TN (true negative)

Accuracy

Also called recognition rate: percentage of test observations that are correctly classified



	PREDICTED CLASS		
ACTUAL CLASS		Class=Yes	Class=No
	Class=Yes	a (TP)	b (FN)
	Class=No	c (FP)	d (TN)

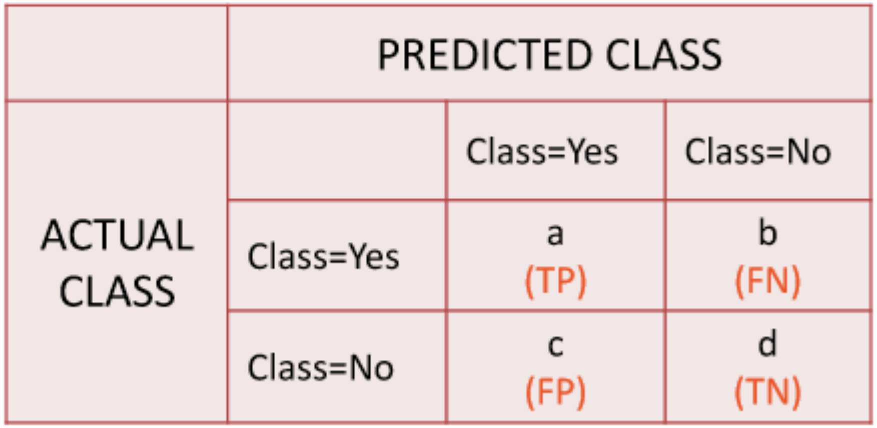
$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} = \frac{a + d}{a + b + c + d}$$

Error rate = 1- Accuracy

Precision

Also called exactness - what % of observations that the classifier labeled as positive are actually positive

$$p = \frac{TP}{TP + FP} = \frac{a}{a + c}$$



		PREDICTED CLASS	
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	a (TP)	b (FN)
	Class=No	c (FP)	d (TN)

Accuracy vs. Precision

Accuracy refers to the closeness of a measured value to a standard or known value

For example, if in lab you obtain a weight measurement of 3.2 kg for a given substance, but the actual or known weight is 10 kg, then your measurement is not accurate. In this case, your measurement is not close to the known value.

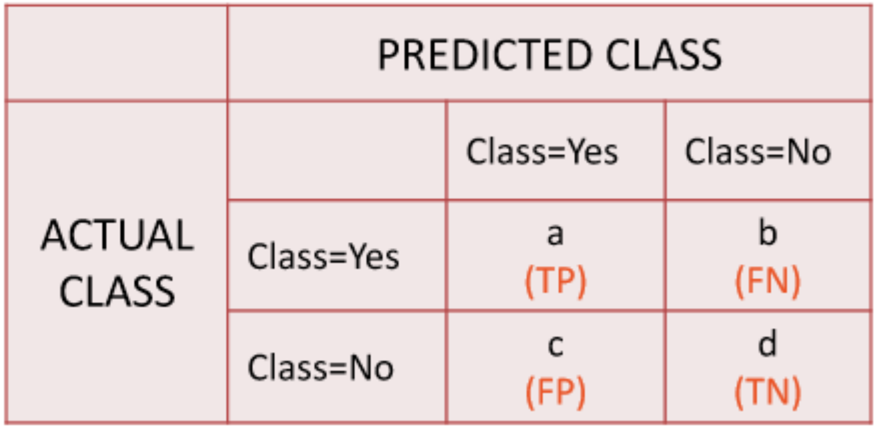
Precision refers to the closeness of two or more measurements to each other

Using the example above, if you weigh a given substance five times, and get 3.2 kg each time, then your measurement is very precise. Precision is independent from accuracy. You can be very precise but inaccurate, as described above. You can also be accurate but imprecise.

Recall/Sensitivity

Also called completeness – what % of positive observations did the classifier label as positive?

$$r = \frac{TP}{TP + FN} = \frac{a}{a + b}$$



	PREDICTED CLASS		
ACTUAL CLASS		Class=Yes	Class=No
	Class=Yes	a (TP)	b (FN)
	Class=No	c (FP)	d (TN)

F1-score

$$F1 = \frac{2rp}{r+p} = \frac{2a}{2a+b+c}$$

Harmonic mean of precision and recall

	PREDICTED CLASS		
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	a (TP)	b (FN)
	Class=No	c (FP)	d (TN)

Thank You for your attention.

Q&A