CMT307: Applied ML Session 3

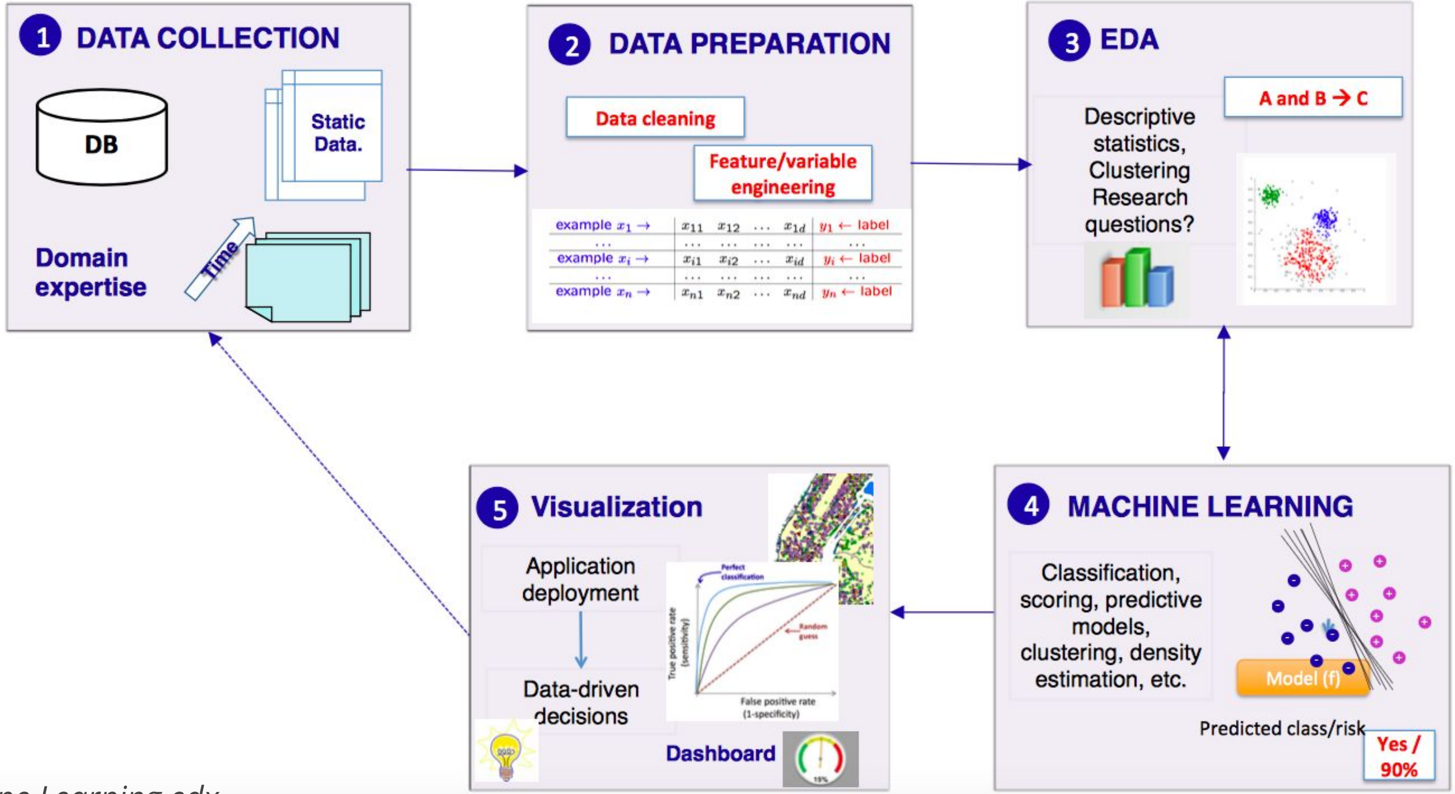
Experimental design

Jose Camacho Collados

Reminders (last session)

- Machine learning pipeline + Feature engineering/selection/extraction.
- Python notebook on basic **feature engineering/selection** with *sklearn* (Solutions to the exercises now available in learning central).

Machine Learning pipeline (last sesssion)



1 DATA COLLECTION

DB

Static Data.

Domain expertise

Time

2 DATA PREPARATION

Data cleaning

Feature/variable engineering

example $x_1 \rightarrow$	x_{11}	x_{12}	$\cdots$	x_{1d}	$y_1 \leftarrow$ label
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
example $x_i \rightarrow$	x_{i1}	x_{i2}	$\cdots$	x_{id}	$y_i \leftarrow$ label
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
example $x_n \rightarrow$	x_{n1}	x_{n2}	$\cdots$	x_{nd}	$y_n \leftarrow$ label

3 EDA

Descriptive statistics, Clustering Research questions?

A and B → C

4 MACHINE LEARNING

Classification, scoring, predictive models, clustering, density estimation, etc.

Model (f)

Predicted class/risk

Yes / 90%

5 Visualization

Application deployment

Data-driven decisions

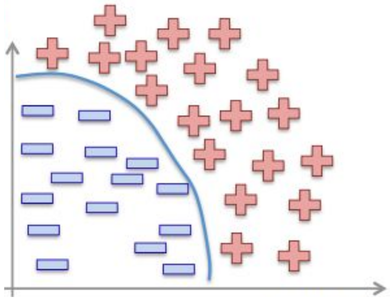
Dashboard

Perfect classification
Random guess
True positive rate (sensitivity)
False positive rate (1-specificity)

Slide credit: *Machine Learning edx*

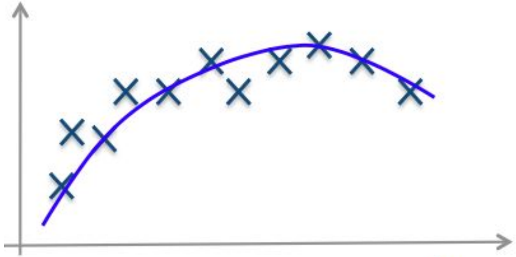
Classification vs. Regression (last session)

Classification (binary)



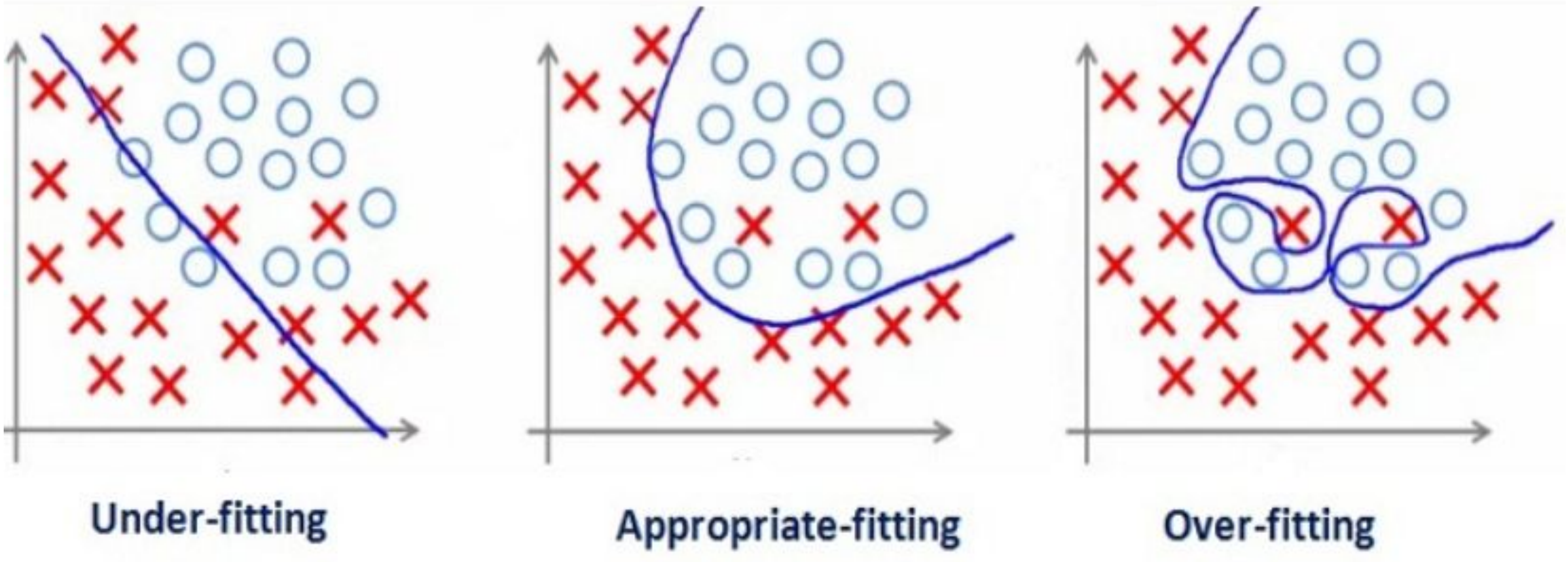
Categories

Regression



Continuous values

Overfitting vs. underfitting (last session)



Under-fitting

Appropriate-fitting

Over-fitting

Slide credit: *GreyAtom*

Outline

- ➢ Housekeeping/opportunities
- ➢ Experimental design
- ➢ Evaluation measures
- ➢ Coursework
- ➢ Hands on!

Panopto: Video-recorded lectures

Currently not working properly (IT issue)

It should be fixed for this week (hopefully!)

Two first sessions may be lost - apologies!

Data and Knowledge Engineer Seminars

Weekly seminars (on Mondays from 1pm to 2pm) about machine learning, knowledge representation and data mining.

Location: Room WX/3.07 (West extension, Queen’s building)

- **Last:** *Interface between AI and Social Sciences and why it matters*
- **Next (28 Oct):** *Sentence embeddings (NLP) + Classifying memory reactivation during sleep using machine learning*

Link:
https://www.cardiff.ac.uk/computer-science/events/data-and-knowledge-engineering-seminars

Opportunities in Cardiff: China Scholarship Council

Tuition fee PhD scholarships in the fields of science and technology

How to apply: CV+Personal Statement

First-Come-First-Served! Apply early = More chances

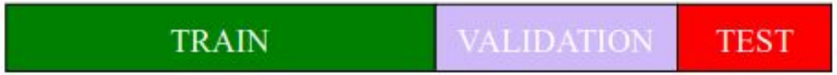
Application deadline: 15 January 2020. **Start date**: October 2020

Link: https://www.cardiff.ac.uk/study/international/funding-and-fees/international-scholarships/phd-research-scholarships

[Only for Chinese students]

Experimental design

Train, validation and test

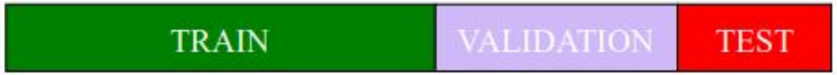


Datasets are usually split into three parts:

- **Training set:** Set of instances (examples) used to train a model.
- **Development (validation) set:** Instances that are used to train the parameters of your model -> Useful to avoid overfitting!
- **Test set:** Split where the model is evaluated (only used once!).

Slide credit: *Machine Learning edx*

Train, validation and test



TRAIN VALIDATION TEST

How to split the dataset?

No general rule. Generally training data larger, but test data should be large enough for the evaluation to be meaningful (and development for the model to generalize well).

Some usual train/dev/test splits: 80%/10%/10% or 60%/20%/20%

Slide credit: *Machine Learning edx*

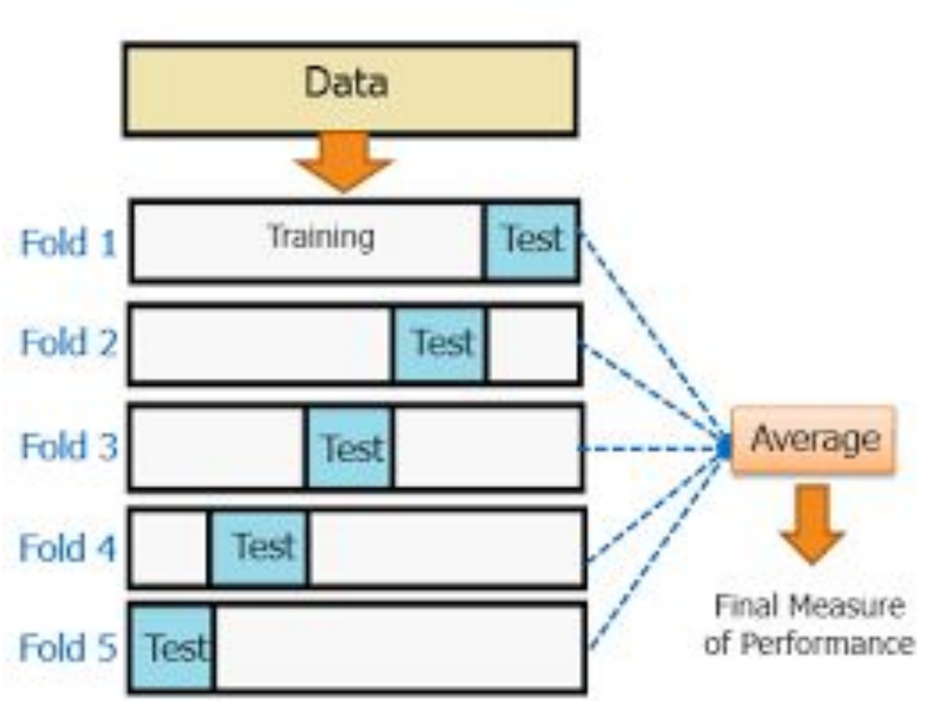
Cross-validation

Automatic creation of train/test splits (dev if needed) using *k* folds.

Useful when the dataset is not large, or with a very skewed distribution.

Important: No preprocessing or feature selection can be done in the initial full dataset! All steps must be performed within the folds.

Example: 5-fold cross-validation



Slide credit: *Sunny Srinidhi*

Evaluation measures

Evaluation measures (binary classification)

Binary classifiers are evaluated by comparing their predictions to a reference label set, also referred to as ground truth or **gold standard**.

The results of a classifier and an understanding of its behaviour can be easily condensed in a **confusion matrix** (or contingency table).

Slide credit: *Luis Espinosa*

Evaluation measures (binary classification)

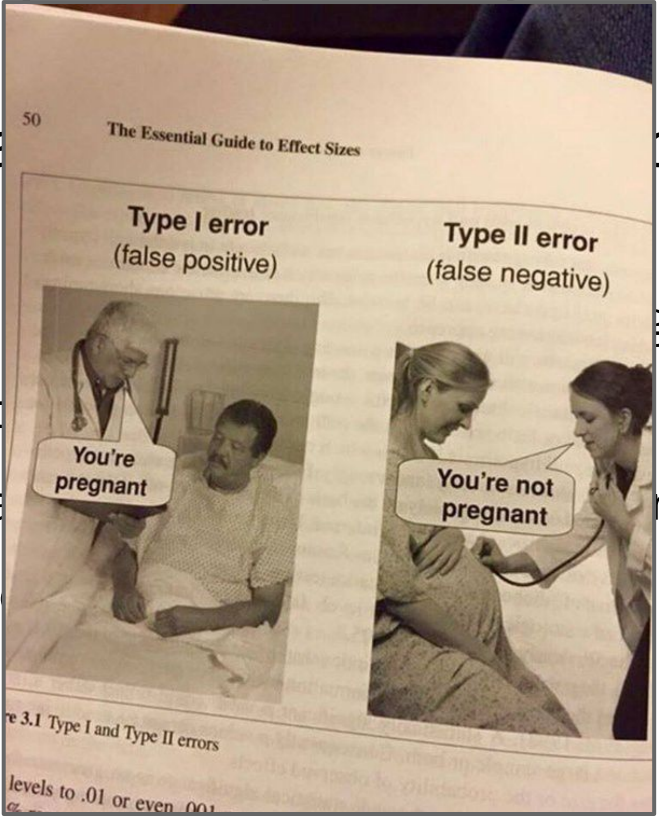
Example: Predicting whether a person has diabetes or not

- ➢ **True Positive (TP):** Our prediction came out positive, and it was right
- ➢ **True Negative (TN):** Our prediction came out negative, and it was right.
- ➢ **False Positive (FP):** Our prediction came out positive, but it was wrong.
- ➢ **False Negative (FN):** Our prediction came out negative, but it was wrong.

Evaluation measures (binary classification)

Example: Predicting whether a pe *ot*

- ➢ **True Positive (TP):** Our prediction ca ght
- ➢ **True Negative (TN):** Our prediction c right.
- ➢ **False Positive (FP):** Our prediction ca rong.
- ➢ **False Negative (FN):** Our prediction wrong.



50 The Essential Guide to Effect Sizes

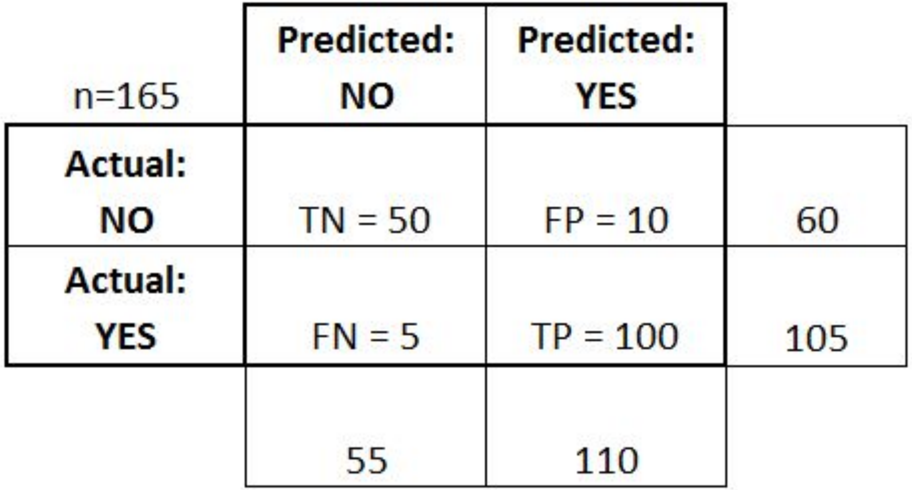
Type I error (false positive)

Type II error (false negative)

re 3.1 Type I and Type II errors

levels to .01 or even .001

Confusion matrix



n=165	Predicted: NO	Predicted: YES	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	

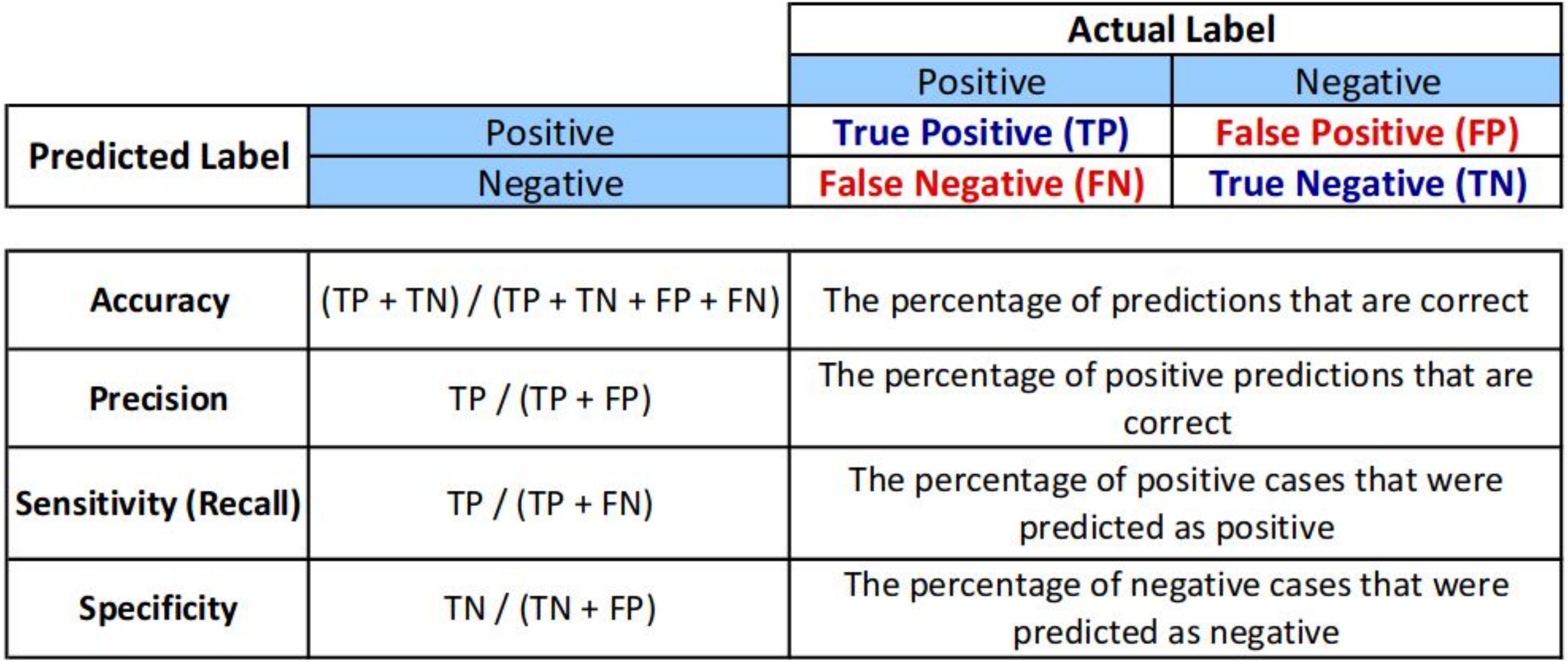
TN: True negative

FN: False negative

FP: False positive

TP: True positive

Evaluation measures (binary classification)



		Actual Label	
		Positive	Negative
Predicted Label	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

Accuracy	(TP + TN) / (TP + TN + FP + FN)	The percentage of predictions that are correct
Precision	TP / (TP + FP)	The percentage of positive predictions that are correct
Sensitivity (Recall)	TP / (TP + FN)	The percentage of positive cases that were predicted as positive
Specificity	TN / (TN + FP)	The percentage of negative cases that were predicted as negative

Slide credit: *Machine Learning edx*

Evaluation measures: F1-Score

While accuracy gives us the percentage of correct answers, sometimes this is not enough (e.g. a dataset heavily biased towards one class).

Another important measure if **F1-Score** (or F-Measure), which is the harmonic mean of precision and recall.

$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$

Evaluation measures (multi-class classification)

- ➢ **Micro-average:** compute the metric independently for each class and then take the average (i.e. treat all the classes equally)
- ➢ **Macro-average:** aggregate the contributions of all classes to compute the average metric (i.e. more frequent classes have more weight).

Micro-average useful for general **overview**, and macro-average useful if we want our classifier to **work well in all classes**.

We can then compute **micro/macro precision, recall or F1-score**.

More information: http://rushdishams.blogspot.com/2011/08/micro-and-macro-average-of-precision.html

Statistical significance tests

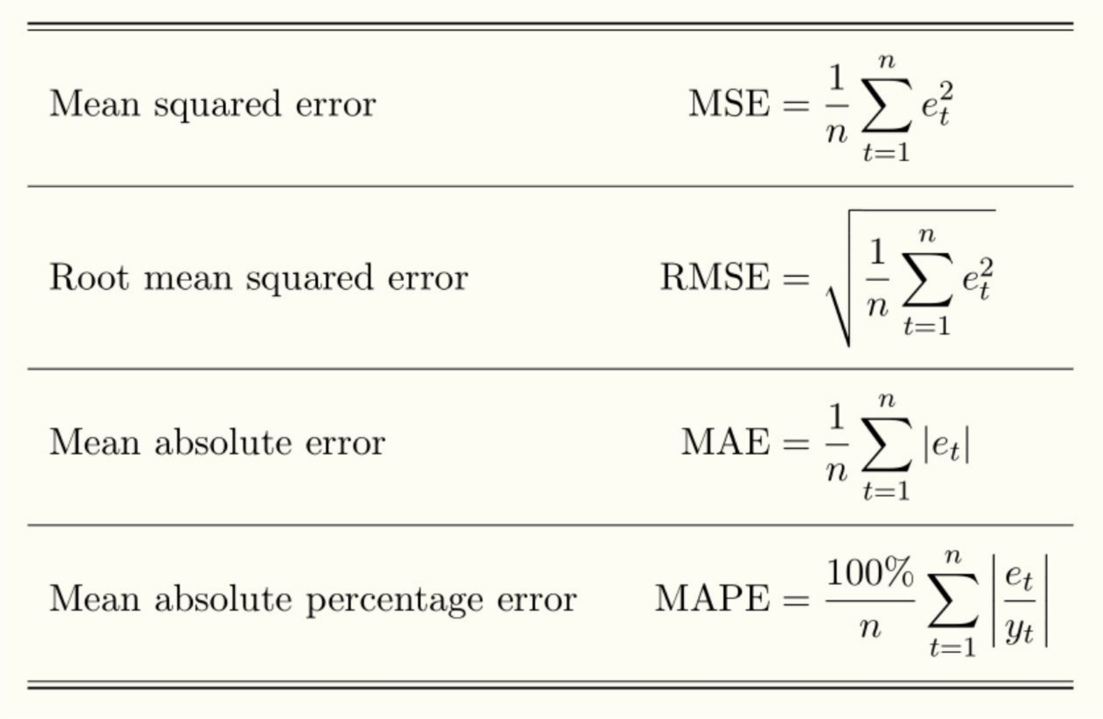
Many times we want to compare the performance between ML models. However, depending on the size of the dataset, the difference between two different models may be due to some statistical artifact.

To know how meaningful these differences are, we run **statistical hypothesis tests**.

Depending on the the case and evaluation metric, we may use different tests (**t-test**, **chi-squared, McNemar**, etc.).

More information:
https://machinelearningmastery.com/statistical-significance-tests-for-comparing-machine-learning-algorithms

Evaluation measures (regression)



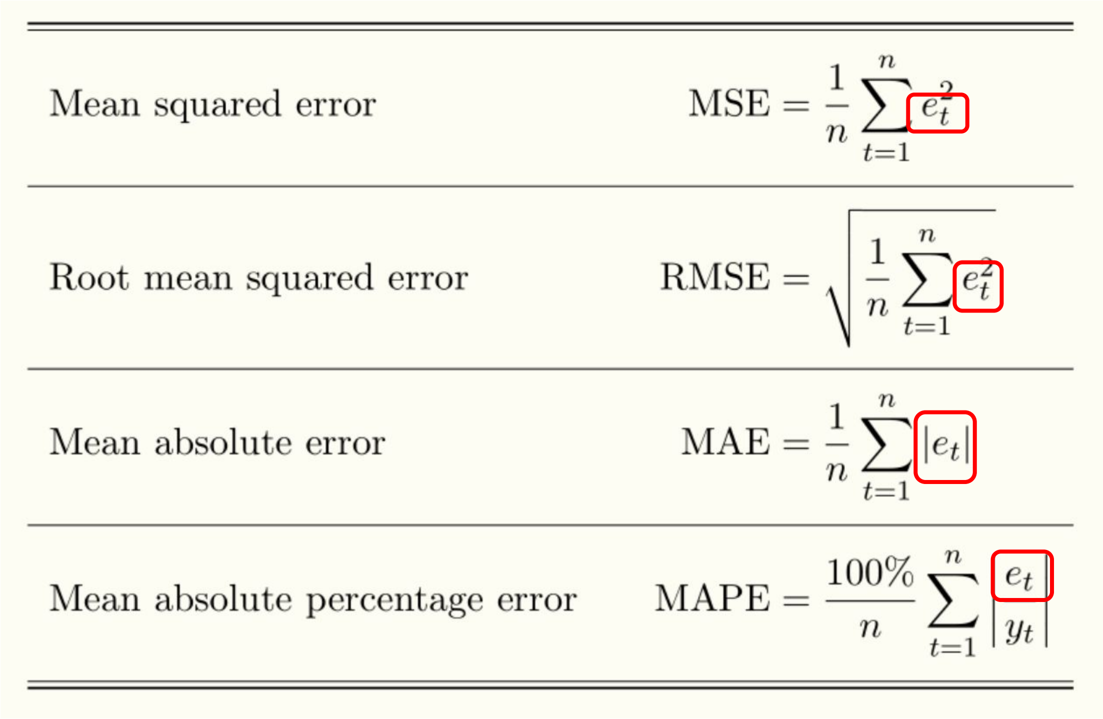
Mean squared error	$\mathrm{MSE} = \frac{1}{n}\sum_{t=1}^{n} e_t^2$
Root mean squared error	$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n} e_t^2}$
Mean absolute error	$\mathrm{MAE} = \frac{1}{n}\sum_{t=1}^{n} \lvert e_t \rvert$
Mean absolute percentage error	$\mathrm{MAPE} = \frac{100\%}{n}\sum_{t=1}^{n} \left\lvert \frac{e_t}{y_t} \right\rvert$

Number of instances=*n*

"*e*" refers to the error in one particular instance (generally, the difference between the prediction and the gold label)

Image credit: *Vimarsh Karbhari*

Evaluation measures (regression)



Mean squared error	$\mathrm{MSE} = \frac{1}{n}\sum_{t=1}^{n} \boxed{e_t^2}$
Root mean squared error	$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n} \boxed{e_t^2}}$
Mean absolute error	$\mathrm{MAE} = \frac{1}{n}\sum_{t=1}^{n} \boxed{\lvert e_t \rvert}$
Mean absolute percentage error	$\mathrm{MAPE} = \frac{100\%}{n}\sum_{t=1}^{n} \left\lvert \frac{\boxed{e_t}}{y_t} \right\rvert$

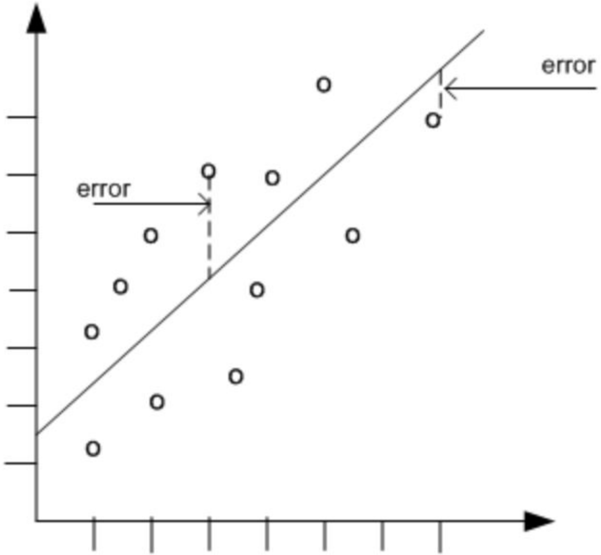


Image credit: *Vimarsh Karbhari*

Coursework

Coursework (1st semester)

- **Available in Learning Central:** ~Monday, October 28th
- **Submission date and time:** Tuesday, January 14th at 9:30am

Most of the exercises are highly practical (Python, sklearn, etc.)

Extra credit possible (a bit more challenging).

School’s private Stack Overflow (reminder)

https://stackoverflow.com/c/comsc

Post your questions related to the course here!

Add the tags ***cmt307*** and ***machine-learning*** to your question.

Next sessions

Next session: 7 November

Following four sessions (Session 4-7) with Dr. Yuhua Li.

In these sessions you will get to understand and delve into the actual machine learning algorithms (SVM, random forests, etc.). A bit more mathematical.

Session 8: Guest lecture + ethics and bias + recap

See you in Session 8!

Hands on!

Python notebook with exercises
about **experimental design** available at Learning Central