COMP47350: Data Analytics (Conv)

Dr. Georgiana Ifrim
georgiana.ifrim@ucd.ie

Insight Centre for Data Analytics
School of Computer Science
University College Dublin

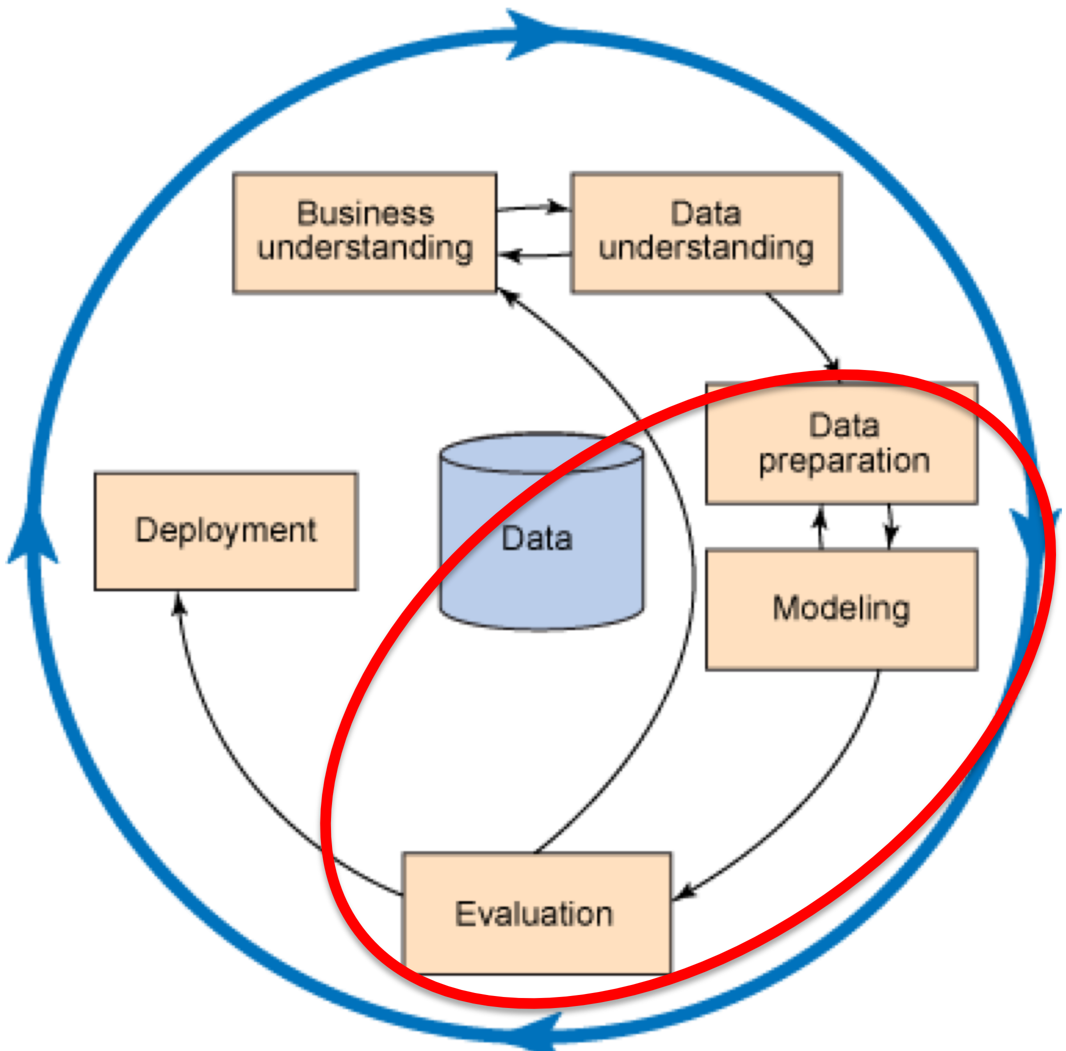
2018/19

Module Topics

- **Python Environment** (Anaconda, Jupyter Notebook)
- **Getting Data** (Web scrapping, APIs, DBs)
- **Understanding Data** (slicing, visualisation)
- **Preparing Data** (cleaning, transformation)
- **Modeling & Evaluation** (machine learning)

Data Analytics Project Lifecycle: CRISP-DM

CRISP-DM: **CR**oss-**I**ndustry **S**tandard **P**rocess for **D**ata **M**ining



Model Evaluation

Experiment Design

- Underfitting/Overfitting
- Out-of-sample Testing
- Cross-Validation

Evaluation Measures

- Regression
- Classification

Model Evaluation

Regression (numeric target):

- Root Mean Squared Error
- R^2

Classification (categorical target):

- Confusion Matrix (aka Error Matrix)
- ROC Curve (and AUC)

Classification: Evaluation Metrics

Confusion Matrix: table describing the performance of a classifier; needs YES/NO predictions

- Summarizes the agreement between the Actual and the Predicted values

Example: **Test for the presence of disease** (165 patients)

NO = negative test (can be coded as 0; this is the negative class)

YES = positive test (can be coded as 1; this is the positive class)

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

Errors!

Errors!

Confusion Matrix

- Example: **Test for the presence of disease** (165 patients)
  NO = negative test; YES = positive test

- The confusion matrix tells you:
  - How many classes are there?
  - How many patients?
  - How many times is disease predicted?
  - How many patients actually have the disease?

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

Confusion Matrix

<u>Basic Terminology:</u>

- **True Positives (TP):** Actual YES and Predicted YES
- **True Negatives (TN):** Actual NO and Predicted NO
- **False Positives (FP):** Actual NO and Predicted YES
- **False Negatives (FN):** Actual YES and Predicted NO

Most classification metrics are derived using the confusion matrix

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

Confusion Matrix

Accuracy (in [0,1], high is good):

- Overall, how often is the classifier **correct**?
- (TP + TN) / total = 150/165 = 0.91

Misclassification Rate (Error Rate, in [0,1], low is good):

- Overall, how often is the classifier **wrong**?
- (FP + FN) / total = 15/165 = 0.09

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

Confusion Matrix

True Positive Rate (in [0,1], high is good):
When actual value is **positive**, how often is the prediction **correct**?

- TP / actual yes = 100/105 = 0.95

False Positive Rate (in [0,1], low is good):

- When actual value is **negative**, how often is the prediction **wrong**?
- FP / actual no = 10/60 = 0.17

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

Confusion Matrix

Precision: correctly predicted positive / predicted positive

Recall: correctly predicted positive / actual positive

F1-measure: aggregation of Precision and Recall

$$\text{precision} = \frac{TP}{(TP + FP)}$$

$$\text{recall} = \frac{TP}{(TP + FN)}$$

$$\text{F}_1\text{-measure} = 2 \times \frac{(\text{precision} \times \text{recall})}{(\text{precision} + \text{recall})}$$

Classification Evaluation Measures

- Hands-on exercise to compute the previous evaluation metrics

Classification: Prediction Scores

- The Confusion Matrix assumes YES/NO class predictions, but most classifiers output a score that needs to be first thresholded to get a class decision

- All our classification prediction models return a score which is then thresholded.

Example

$$threshold(score, 0.5) = \begin{cases} positive & \text{if } score \geq 0.5 \\ negative & otherwise \end{cases} \tag{10}$$

Classification: AUC Measure

- Some evaluation metrics directly use the predicted score, without the need to set a fixed threshold: **AUC** (Area Under Roc Curve)

- Intuitively the AUC measures how many times the classifier places true positives above true negatives (i.e., the predicted score for true positives should be higher than the predicted score for true negatives).

- It first sorts all examples by predicted score, then it uses the TPR and FPR computed at different thresholds, to avoid sensitivity of classification decision to a particular threshold.

- AUC varies between 0.5 (random score predictions) and 1 (perfect ranking, when all positives have the predicted scores higher than the predicted scores of the negatives). AUC provides a means of comparison between classification models that may output very different score scales ro score distributions.

Classification: ROC Curve (AUC)

Email Number	Score	True Label
5	0.99	Spam
8	0.82	Spam
2	0.60	Spam
1	0.60	Ham
7	0.48	Spam
3	0.22	Ham
4	0.10	Ham
6	0.02	Ham

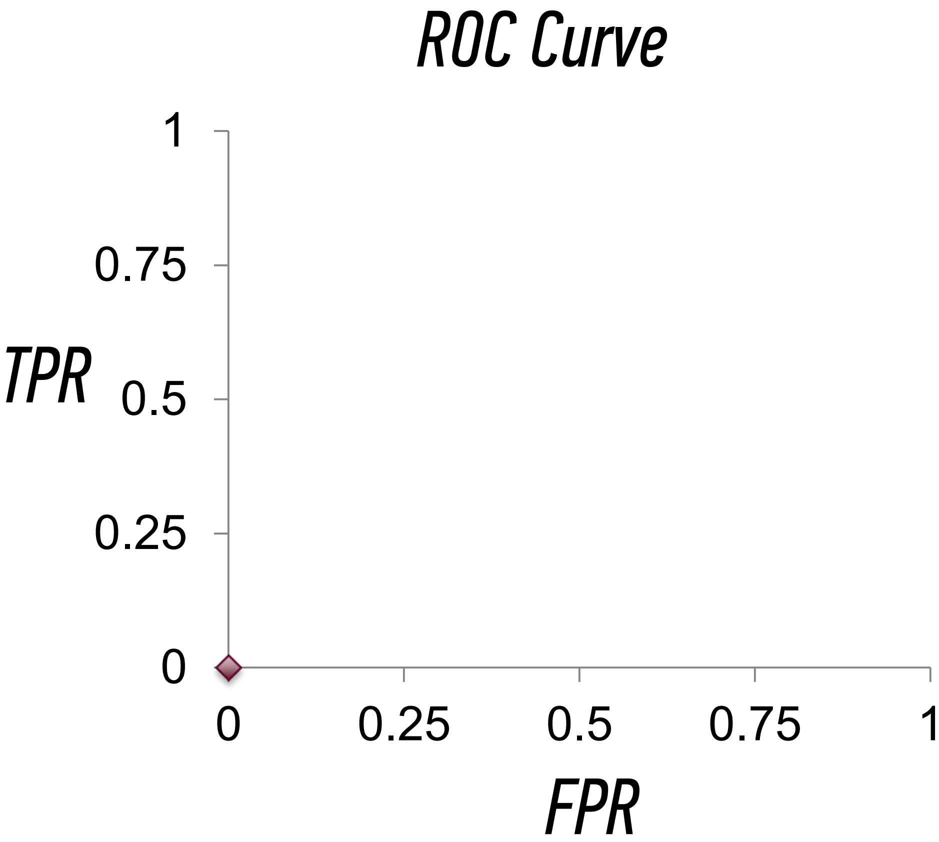
Example: Every email is assigned a "spamminess" score by the classifier. To actually make our class predictions, we choose a numeric cutoff, typically 0.5, for classifying as >=0.5: **spam** or < 0.5: **ham (not-spam).**

A **ROC Curve** helps us visualize how well our classifier is doing without having to choose a cutoff!

AUC (Area under ROC Curve) is a single number to quantify the quality of the ROC curve (between 0 and 1).

Classification: ROC Curve

Email Number	Score	True Label
5	0.99	Spam
8	0.82	Spam
2	0.60	Spam
1	0.60	Ham
7	0.48	Spam
3	0.22	Ham
4	0.10	Ham
6	0.02	Ham



Classification: ROC Curve

Email Number	Score	True Label
5	0.99	Spam
8	0.82	Spam
2	0.60	Spam
1	0.60	Ham
7	0.48	Spam
3	0.22	Ham
4	0.10	Ham
6	0.02	Ham

*TPR: When actual value is **spam**, how often is prediction **correct**?*

*FPR: When actual value is **ham**, how often is prediction **wrong**?*

Cutoff	TPR (y)	FPR (x)	Cutoff	TPR (y)	FPR (x)
0			0.50		
0.05			0.65		
0.15			0.85		
0.25			1		

Classification: ROC Curve

Email Number	Score	True Label
5	0.99	Spam
8	0.82	Spam
2	0.60	Spam
1	0.60	Ham
7	0.48	Spam
3	0.22	Ham
4	0.10	Ham
6	0.02	Ham

*TPR: When actual value is **spam**, how often is prediction **correct**?*

*FPR: When actual value is **ham**, how often is prediction **wrong**?*

Cutoff	TPR (y)	FPR (x)	Cutoff	TPR (y)	FPR (x)
0	1	1	**0.50**	0.75	0.25
0.05	1	0.75	**0.65**	0.5	0
0.15	1	0.5	**0.85**	0.25	0
0.25	1	0.25	**1**	0	0

Classification: ROC Curve

Email Number	Score	True Label
5	0.99	Spam
8	0.82	Spam
2	0.60	Spam
1	0.60	Ham
7	0.48	Spam
3	0.22	Ham
4	0.10	Ham
6	0.02	Ham

Q: Would the ROC Curve (and AUC) change if the ***scores*** *changed but the* ***ordering*** *remained the same?*

A: Not at all! The ROC Curve is only sensitive to ***rank ordering*** *and does not require* ***calibrated scores****.*

Calibrated scores = predicted scores accurately reflect ground truth;
For example, given 100 predictions, each with confidence of 0.8, we expect 80 examples to be correctly classified.
For AUC we could have all predicted score in a low range [0,0.005] and still correctly rank the examples by predicted score.

Classification: ROC Curve

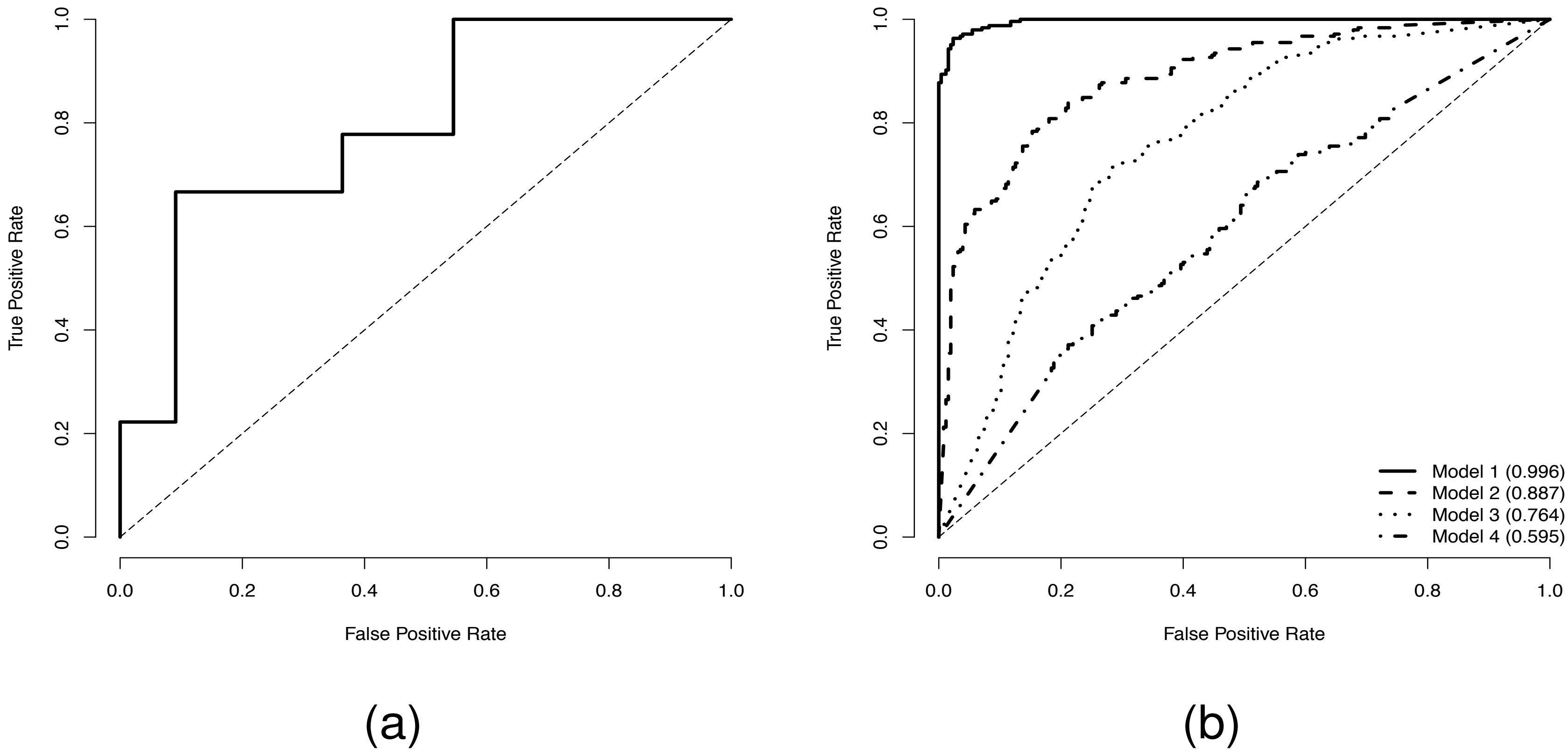


Figure: (a) A complete ROC curve for the email classification example; (b) a selection of ROC curves for different models trained on the same prediction task.

References

- Chapter8 from **FMLPDA Book: Fundamentals of Machine Learning for Predictive Data Analytics**, by J. Kelleher, B. Mac Namee and A. D'Arcy, MIT Press, 2015 (machinelearningbook.com)

- https://github.com/justmarkham/DAT4/blob/master/slides/10_model_evaluation_metrics.pdf
- http://www.theanalysisfactor.com/assessing-the-fit-of-regression-models/
- Pitfalls of the AUC:
  https://www.kdnuggets.com/2010/09/pub-is-auc-the-best-measure.html
- Arguments in favor of the AUC:
  http://www.icml-2011.org/papers/385_icmlpaper.pdf
- Calibrated prediction scores:
  https://arxiv.org/pdf/1706.04599.pdf