Predictive Analytics – Session 5

Predicting Classifications

Associate Professor Ole Maneesoonthorn

Associate Professor in Econometrics and Statistics

Melbourne Business School

O.Maneesoonthorn@mbs.edu

What is “classification”?

What are the challenges in modelling/predicting them?

Classifications & Challenges

What is classification?

- Qualitative data: e.g.
    - Yes/No responses
    - Agree/Neutral/Disagree responses
    - Product choices (brands)

- … basically output variables that are non-numeric

Classifications & Challenges

How do we model & predict classifications?

- What you might see:
  - Convert the labels to numbers
  - Apply the linear regression techniques
- NOT APPROPRIATE! Why?
  - It assumes an inherent ranking in the labels
  - It assumes that shifting from label A to label B is the same as shifting from label B to label C

Classifications & Challenges

How do we model & predict classifications?

- What you should do:
  - Model and predict the likelihood of belonging to a certain label
  - That is, look at probability!
  - Input variables then explain how these probabilities may vary
- Prediction?
  - Label with highest probability wins!
  - Or introduce a threshold of probability according to your context

Classifications & Challenges

Assessing predictive accuracy

- The predictive assessment process remains the same as numerical data
  - Split data into training and test set
  - Train the model using training data set
  - Evaluate predictive accuracy using the test set
- Metrics have to reflect the data type
  - Accurate prediction: actual = predicted
  - Inaccurate prediction: actual $\neq$ predicted

Classifications & Challenges

Assessing predictive accuracy

- Focus on two labels classification
- Two-label classification typically done using 0/1 coding
- But concepts generalizable to multiple labels

Classifications & Challenges

Assessing predictive accuracy

- Hit/miss table (confusion table)

		Actual Observation	
		Yes	No
Predicted Outcome	Yes	True Positive	False Positive
	No	False Negative	True Negative

$$Overall\ Accuracy = \frac{True\ Positive + True\ Negative}{Total}$$

Classifications & Challenges

Assessing predictive accuracy

- Hit/miss table (confusion table)

		Actual Observation	
		Yes	No
Predicted Outcome	Yes	True Positive	False Positive
	No	False Negative	True Negative

$$Precision = \frac{True\ Positive}{Predicted\ Positive}$$

Classifications & Challenges

Assessing predictive accuracy

- Hit/miss table (confusion table)

		Actual Observation	
		Yes	No
Predicted Outcome	Yes	True Positive	False Positive
	No	False Negative	True Negative

$$Recall = \frac{True\ Positive}{Actual\ Positive}$$

Also known as **"Sensitivity"**

Classifications & Challenges

Assessing predictive accuracy

- Hit/miss table (confusion table)

		Actual Observation	
		Yes	No
Predicted Outcome	Yes	True Positive	False Positive
	No	False Negative	True Negative

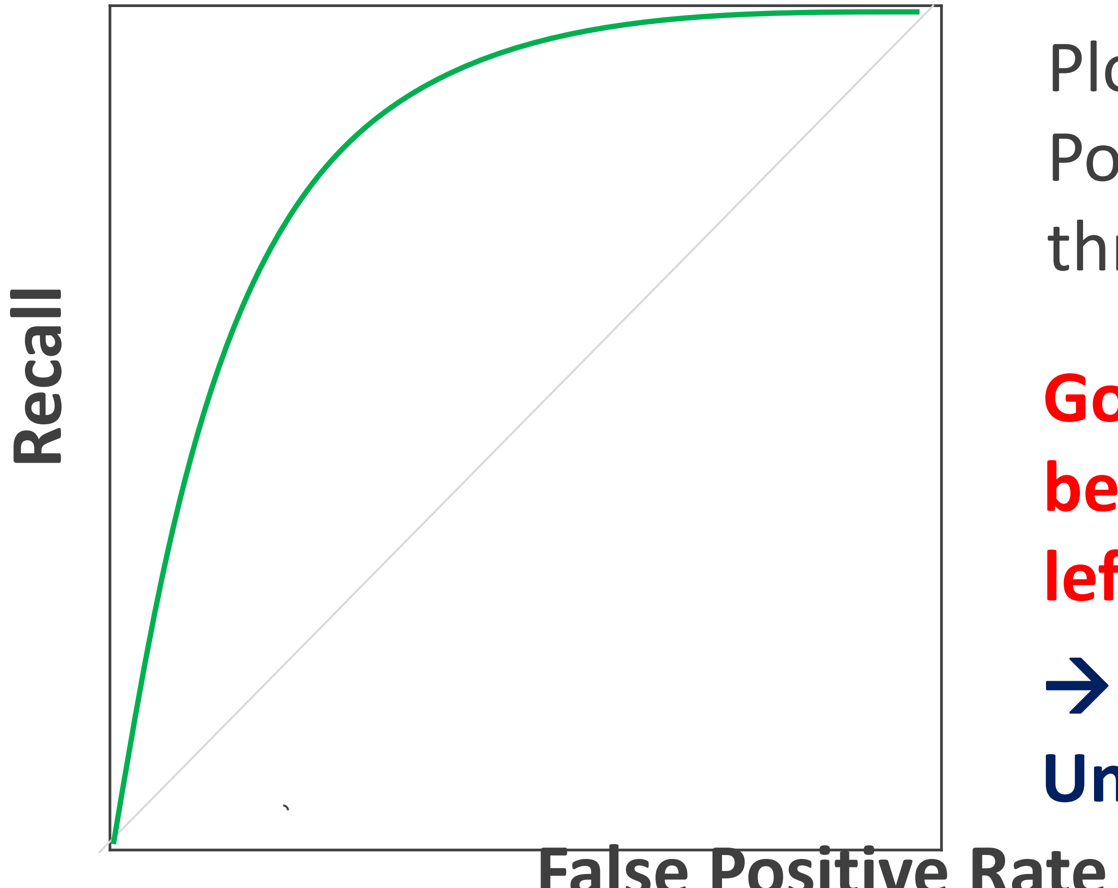
$$Specificity = \frac{True\ Negative}{Actual\ Negative}$$

False Positive Rate is $1 - Specificity$

Classifications & Challenges

Assessing predictive accuracy

- Receiver Operating Characteristics (ROC) curve



Plots Recall vs False Positive Rates at different thresholds

Goal: for the curve to be as close to the top left corner as possible

→ Maximize the Area Under Curve (AUC)

Classifications & Challenges

Assessing predictive accuracy

- Metrics associated with ROC curve:
  - Area Under Curve (AUC) – area under the ROC curve, high is good
  - Maximum achievable is 1
  - AUC = 0.5 → model is as good as a random guess
  - Look for AUC > 0.5

`

Classifications & Challenges

Assessing predictive accuracy

- Metrics associated with ROC curve:
  - Gini index – another interpretation of AUC
  - Recalculation of the AUC so that zero is a benchmark

$$Gini = (2 \times AUC) - 1$$

  - Gini = 0 → model predicts as good as a random guess
  - Negative Gini → model predicts worse than random guess
  - Look for positive Gini → model predicts better than random guess

Classifications & Challenges

Assessing predictive accuracy

- Incorporating context loss – focus on minimizing impacts of errors
- What are the costs if we get it wrong?
- To your business?

		Actual Observation	
		Yes	No
Predicted Outcome	Yes	True Positive	False Positive
	No	False Negative	True Negative

Classifications & Challenges

Assessing predictive accuracy

- Let us take the example of credit risk modelling
- Model determines which loan application is credit worthy
  - Good credit = loan approved
  - Bad credit = loan rejected

		Actual Observation	
		Good loan	Bad loan
Predicted Outcome	Approved	True Positive	False Positive
	Rejected	False Negative	True Negative

Classifications & Challenges

Assessing predictive accuracy

- Let us take the example of credit risk modelling
- Consequence of False Negative? → **Loss of income – opportunity cost**
- Consequence of False Positive? → **Loss of loan amount – real cost**
    - Loss of future income (for duration of agreed loan)
    - How much of the loan can be recovered?
    - Is there collateral on loan?
    - Very context specific!

		Actual Observation	
		Good loan	Bad loan
Predicted Outcome	Approved	True Positive	False Positive
	Rejected	False Negative	True Negative

Loss is highly asymmetric!

Classifications & Challenges

Assessing predictive accuracy

- Let us take the example of credit risk modelling
- Example with numbers: loans with collateral

		Actual Observation	
		Good loan	Bad loan
Predicted Outcome	Approved	True Positive = 52%	False Positive = 8%
	Rejected	False Negative = 2%	True Negative = 38%

Classifications & Challenges

Assessing predictive accuracy

- Let us take the example of credit risk modelling
- Example with numbers: loans with collateral
- **False negative**
  - Avg. opportunity cost of 35% of portfolio value
- **False positive**
  - Loss of future income (after default), avg. cost of 30% of portfolio value
  - Costs associated with resell/release of collateral, avg. cost of 10% of portfolio value

Classifications & Challenges

Assessing predictive accuracy

- Let us take the example of credit risk modelling
- Example with numbers: loans with collateral
- Loss calculation for a \$10m. portfolio:

Expected Loss = $2\% \times (35\% \ \times 10m) + 8\% \times (40\% \times 10m)$

$$= 390k$$

- Different models will give different error rates
- Choose a model that minimize this loss!

		Actual Observation	
		Good loan	Bad loan
Predicted Outcome	Approved	True Positive 52%	False Positive 8%
	Rejected	False Negative 2%	True Negative 38%

Classifications & Challenges

Predictive Models

- Classification models produce probabilities

$$\Pr(Y = 1|X)$$

- Predictive labels:

$$\hat{Y} = 1 \text{ if } \Pr(Y = 1|X) > c$$

  - where c is the “threshold” probability, typically set at 0.5
  - This can be changed based on context
  - → Prediction rule set by context!

Classifications & Challenges

Predictive Models – regression based models

- Linear regression?
  - Not suitable
  - Probability bounded between 0 and 1, linear regression is unbounded
- Alternative to address boundary issue: LOGISTIC regression

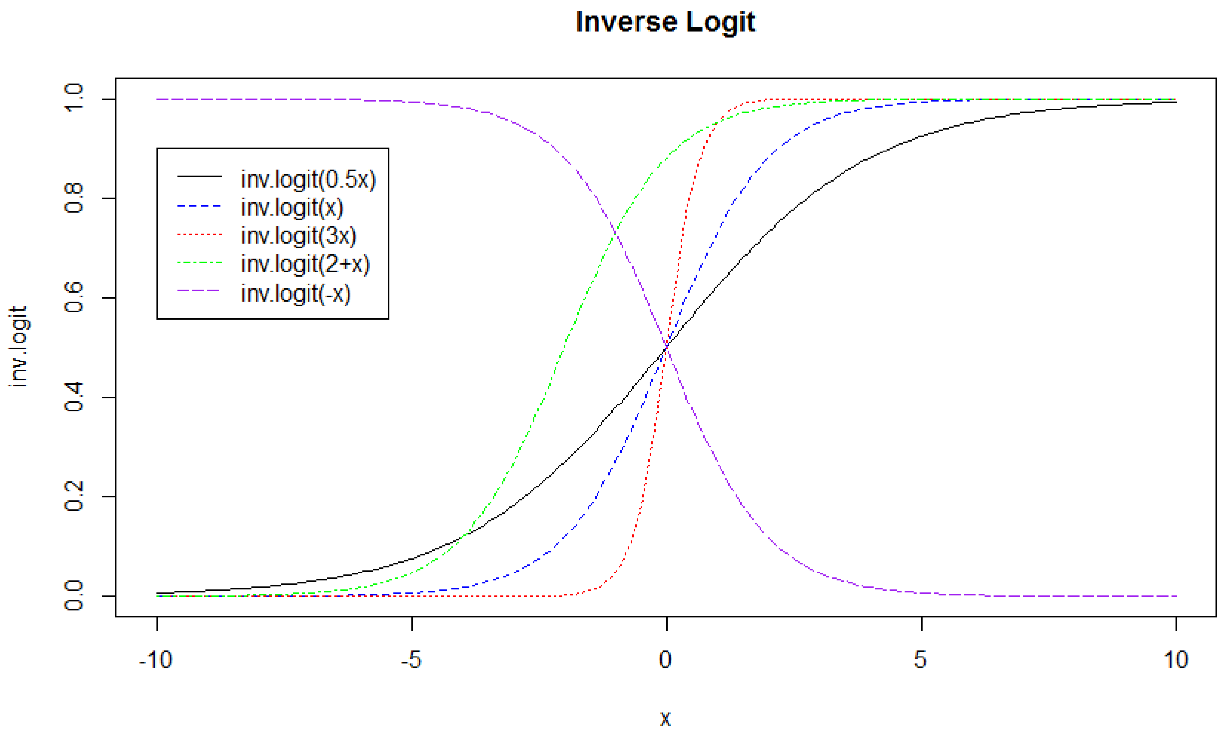
$$\Pr(Y|X) = f(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k)$$

- What is this function $f(.)$?
  - Inverse logistic function $f(x) = \frac{e^x}{1+e^x}$
  - Translate a real number to a boundary between 0 and 1

Classifications & Challenges

Predictive Models – regression based models

- This gives the logistic regression
- Relate input variables to probabilities



Classifications & Challenges

Predictive Models – algorithm-based models

Tree segmentation

- Now known as “Decision Trees”
- Difference: observation at leaf nodes are used to calculate probabilities for each label
- Predictive process: trace the relevant tree branch using input variables
- Leaf node presents the predicted probability

Classifications & Challenges

Predictive Models – algorithm-based models

Neural network

- Trained in the usual way with hidden neurons
- Link function must translate output to probabilities
- Similar to logistic regression but with hidden layer → flexibility

Classifications & Challenges

Predictive Models – Unsupervised learning

K-means

- Input variables are used to obtain distinct segments
- Calculate probabilities of each label for each segment – these are used for predictions
- Note – input variables need to be numeric for this to work well

Classifications & Challenges

Predictive Models – Unsupervised learning

K-NN

- Input variables are used to train “neighbourhoods”
- Calculate probabilities of each label for each neighbourhood
- Bayes theorem used to obtain prediction of neighbourhood and classification

Classifications & Challenges

Predictive Models – algorithm-based models

Ensemble methods

- Bagging and random forest – algorithm remains as described in session 4
- With exception that the base learner is the "decision tree"
- Boosting – base learner can be a decision tree or logistic regression type model
    - glmboost() uses logistic regression as the base learner
    - Boosting algorithm remains as described in session 4

Classifications & Challenges

Predictive Models – algorithm-based models

SVM

- Conceptually – separation of the data into segments
- Goal: maximize the margin/gap between each segment
- But output variable plays a role in the definition of the objective function
    - As part of the constraint of the optimization problem

Classifications & Challenges

Remember:

- Scale your numeric input data for:
  - Neural network
  - K-means & K-NN
- Use the appropriate metrics for classification
  - Summaries of hit/miss table
  - Loss function constructed from user preference
- Each model still suffer from their respective pros & cons!

Classifications & Challenges

Multi-class output

- Discussion so far focuses on binary outcomes
- Most methods can be generalized to multi-class outputs
- Logistic regression – depends if the outcomes have a sense of ranking
  - If apparent ranking – ordered logistic regression (utility based concepts)
  - If no apparent ranking – multinomial logistic
- Tree, unsupervised and ensemble methods
  - Multiple categories and multiple probability calculations
  - Basic outline of algorithms remain as before

Classifications & Challenges

Multi-class output

- SVM – can only handle binary outputs!
- Possible solution:
  - Train SVM multiple times using “One vs ALL”, “One vs One” or “One vs Base”
  - Each algorithm designed to predict a particular class of the outputs
  - Predictions may not be consistent…..
- Ongoing research on this front