'SHOW ME THE MONEY'

PREDICTING INCOME VIA SUPPORT VECTOR MACHINE

Mark Healy

Motivation

- **Binary Classification using Machine Learning**
  - The aim of this project is to tackle a binary classification problem, using a machine learning approach (to approach an old problem with a new technique if you will).
  - Through using a supervised learning model; a Support Vector Machine (SVM), which 'learns' from the data shall be fitted and it's performance assessed.

Overview

- **Data**
  - The data used is extracted from the US Census, and contains demographic values on circa 30,000 respondents.

- **Research Question**
  - *"Based on a respondent's demographics, do they earn (i) less than or equal to $50k per annum, or (ii) in excess of $50k per annum".*

Approach (I of II)

- **Support Vector Machine (SVM)**
  - SVM is a supervised learning classification algorithm

- **How does it work?**
  - Uses a separating hyperplane to separate the classes
  - Classes: {≤ \$50k, > \$50k}
  - Model learns from **training** data (70%) and is **tested** on unseen data (30%)

Approach (II of II)

- **Non-Linear Separable Case**
  - For non-linearly separable data, a kernel function is applied to the data
    - This maps the data to a higher dimensional space
    - The data is now separable
  - Simplistic example illustrated below: Fig 1.1 → Kernel → Fig 1.2

Fig 1.1: Linearly Inseparable

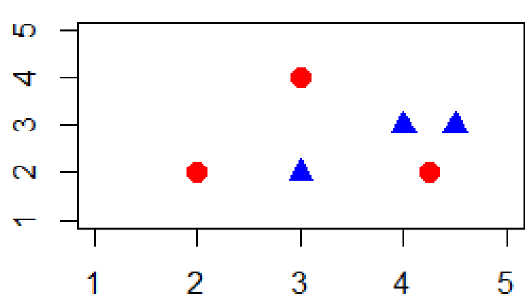
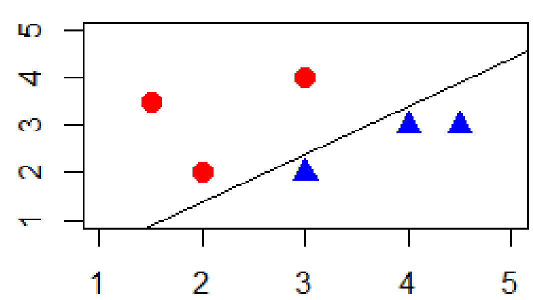


Fig 1.2: Linearly Separable



Model Validation

- **Model Validation is comprised of the following:**

- **Approach 1**
  - K-Folds Cross Validation

- **Approach 2**
  - Confusion Matrices

- **Approach 3**
  - Receiver-Operator Characteristic (ROC) Curve

Model Validation, Approach1(I of III)

- **K-Folds Cross Validation (K=10)**
- **How does it work?**
  - Data is divided into 10 Folds
  - One is Test Data, 9 are Train Data
  - Replicate 10 times
- **Why this approach?**
  - All observations are used for both training and validation

Model Validation, Approach 1 (II of III)

- K-Folds Cross Validation (K=10) Overview

Iteration	← Training Data: 10 Folds (K = 10) →									
K = 1	Test	Train	Train	Train	Train	Train	Train	Train	Train	Train
K = 2	Train	Test	Train	Train	Train	Train	Train	Train	Train	Train
K = 3	Train	Train	Test	Train	Train	Train	Train	Train	Train	Train
K = 4	Train	Train	Train	Test	Train	Train	Train	Train	Train	Train
K = 5	Train	Train	Train	Train	Test	Train	Train	Train	Train	Train
K = 6	Train	Train	Train	Train	Train	Test	Train	Train	Train	Train
K = 7	Train	Train	Train	Train	Train	Train	Test	Train	Train	Train
K = 8	Train	Train	Train	Train	Train	Train	Train	Test	Train	Train
K = 9	Train	Train	Train	Train	Train	Train	Train	Train	Test	Train
K = 10	Train	Train	Train	Train	Train	Train	Train	Train	Train	Test

Model Validation, Approach1 (III of III)

- **SVM Model Parameters**
- **Cost (C)**
  - Penalises the misclassification cost
    - A small C corresponds to a lower misclassficiation Cost
- **Gamma (γ)**
  - Relates to the Kernel applied
    - A small γ results in low bias and high variance

Optimal parameters found using the `svm.tune()` function

C = 1.1   γ = 0.3

10-Fold Cross-Validation Error Rate = 16.5%

Model Validation, Approach 2 (I of II)

- **Confusion Matrix**
  - A table of Actual vs. Predicted Classification for the data.
  - Computed for both the Train and Test data

- **Why this approach?**
  - A variety of performance metrics can be derived from these tables such as Error Rates, Sensitivity and Specificity etc.

Model Validation, Approach 2 (II of II)

- Confusion Matrices

		Actual	
		Yes	No
Predicted	Yes	True Positive (TP)	False Positive (FP)
	No	False Negative (FN)	True Negative (TN)

Train Data

		Actual	
		<=50K	>50K
Predicted	<=50K	14,672	2,026
	>50K	1,138	3228

Test Data

		Actual	
		<=50K	>50K
Predicted	<=50K	6,205	924
	>50K	582	1316

	Train	Test
Error Rate	15.02%	16.68%

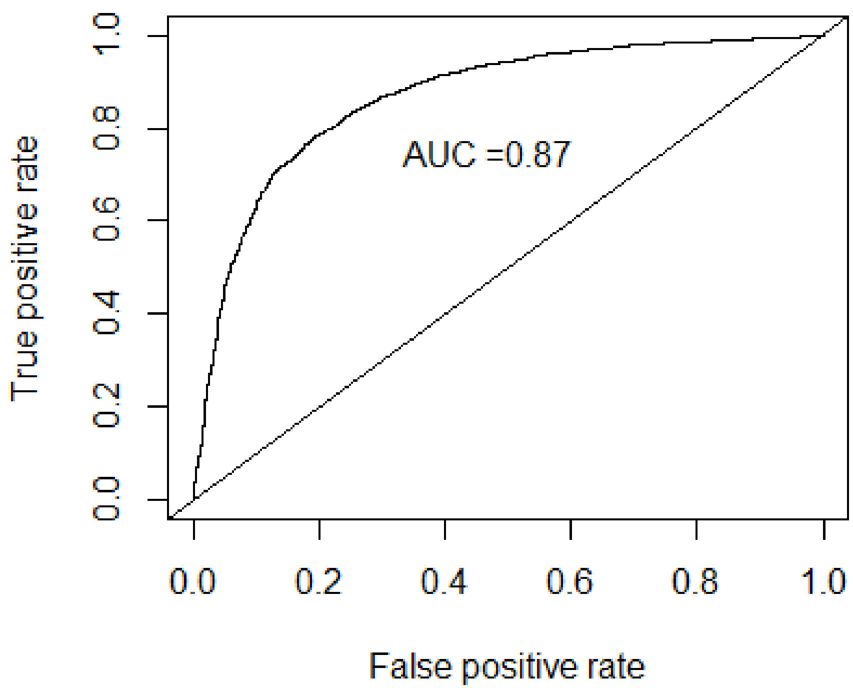
Model Validation, Approach 3 (I of II)

- **Receiver-Operator-Characteristic (ROC) Curve**
  - A visual tool used to ascertain the discriminatory ability of a classification model
  - Displays the trade-off between the sensitivity and specificity of a model

- **Why this approach?**
  - Visual approach
  - Easy to interpret
  - Accompanied by metric for interpretation/comparison: Area under the Curve (AUC)

Model Validation, Approach 3 (II of II)

- ROC Curve Interpretation
  - Curve indicates a model with good discriminatory ability
  - AUC = 0.87 (Perfect model AUC = 1.00)



Conclusions

- **SVM Model Findings**
  - The SVM model produced showed itself to be a good classifying tool, as evidenced by the model validation metrics:
    - Training Error Rate: 15.02%
    - 10-Fold Cross-Validation Error Rate: 16.5%
    - Test Error Rate: 16.68%
    - AUC: 0.87
  - A feature in this model; the `svm.tune()` command, helps ensure badly chosen parameters are not an issue when developing this model.