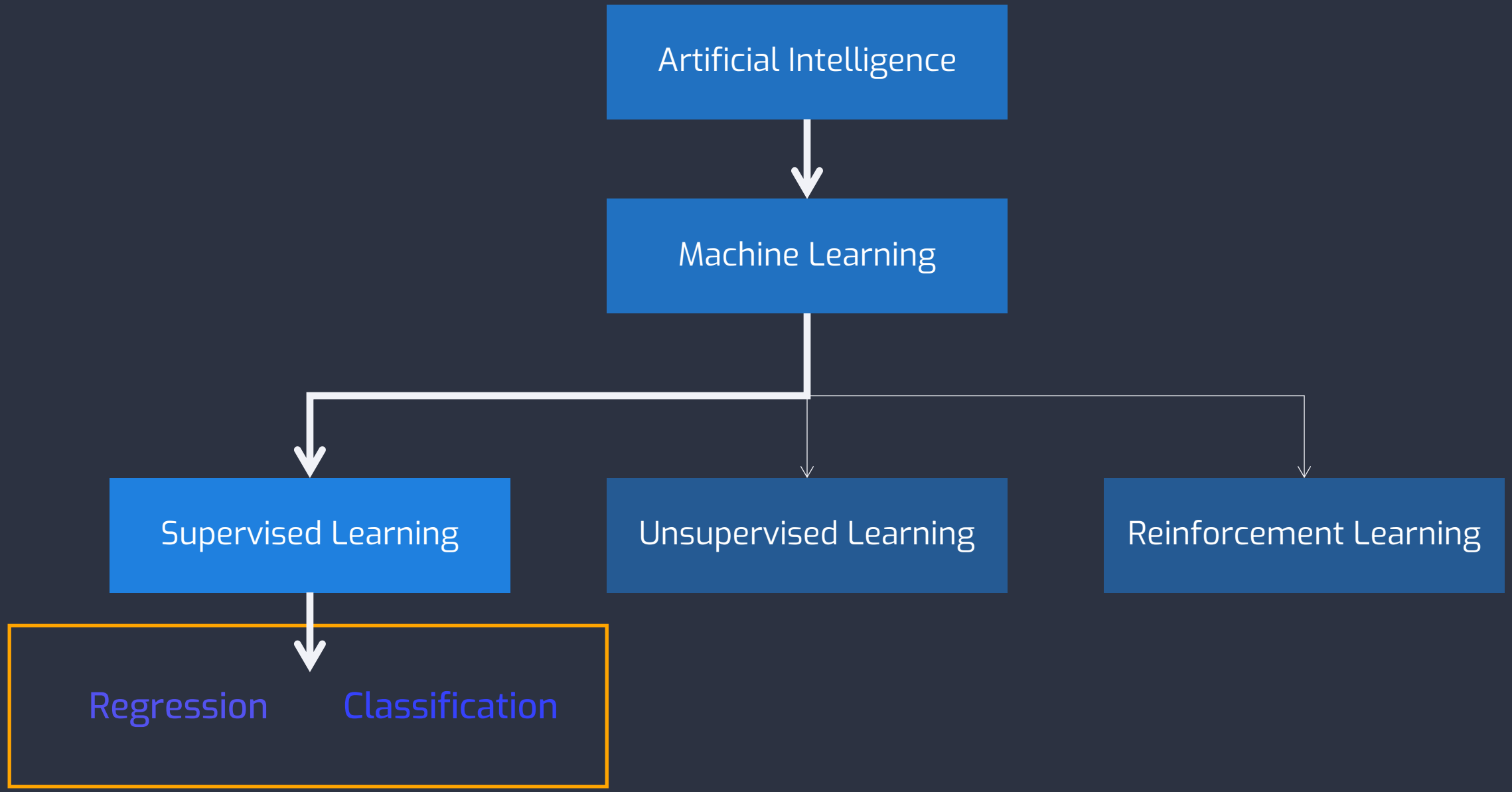
COMP2261 ARTIFICIAL INTELLIGENCE / MACHINE LEARNING

Regression and Classification

Dr SHI Lei

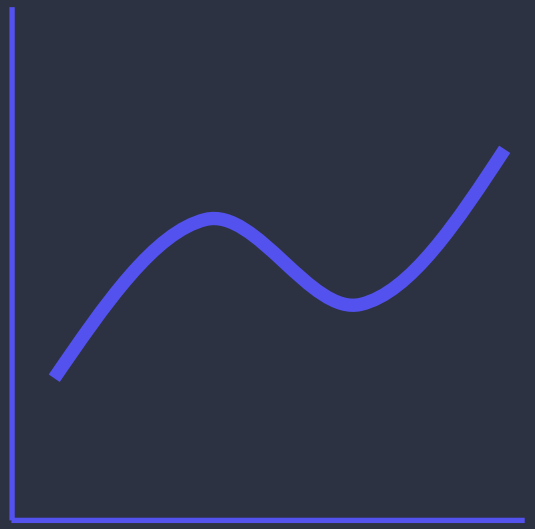
Durham University

learning lab

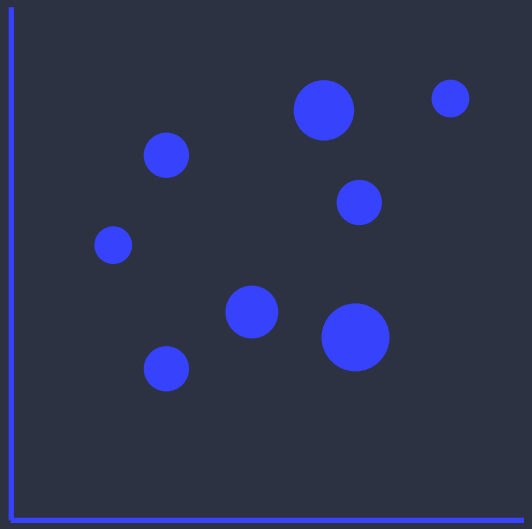

Artificial Intelligence
Machine Learning
Supervised Learning
Unsupervised Learning
Reinforcement Learning
Regression
Classification

Two tasks of supervised learning

- Regression: predicting a continuous (numerical) quantity output for an input.
- Classification: predicting a discrete (categorical) class label output for an input.



Regression



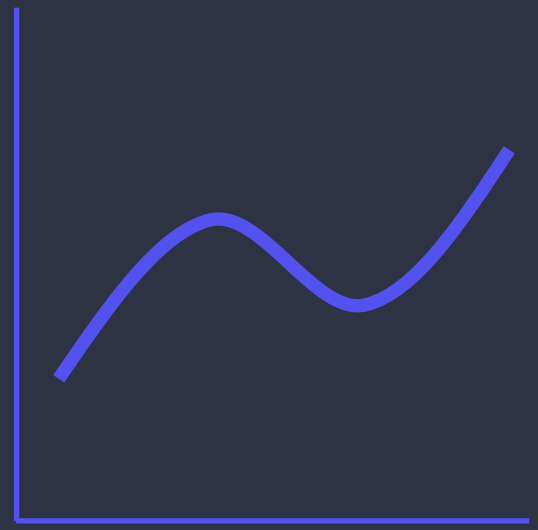
Classification

Learning Objectives

- Understand the two types of supervised learning
- Understand the differences between regression & classification
- Understand how to estimate their performance

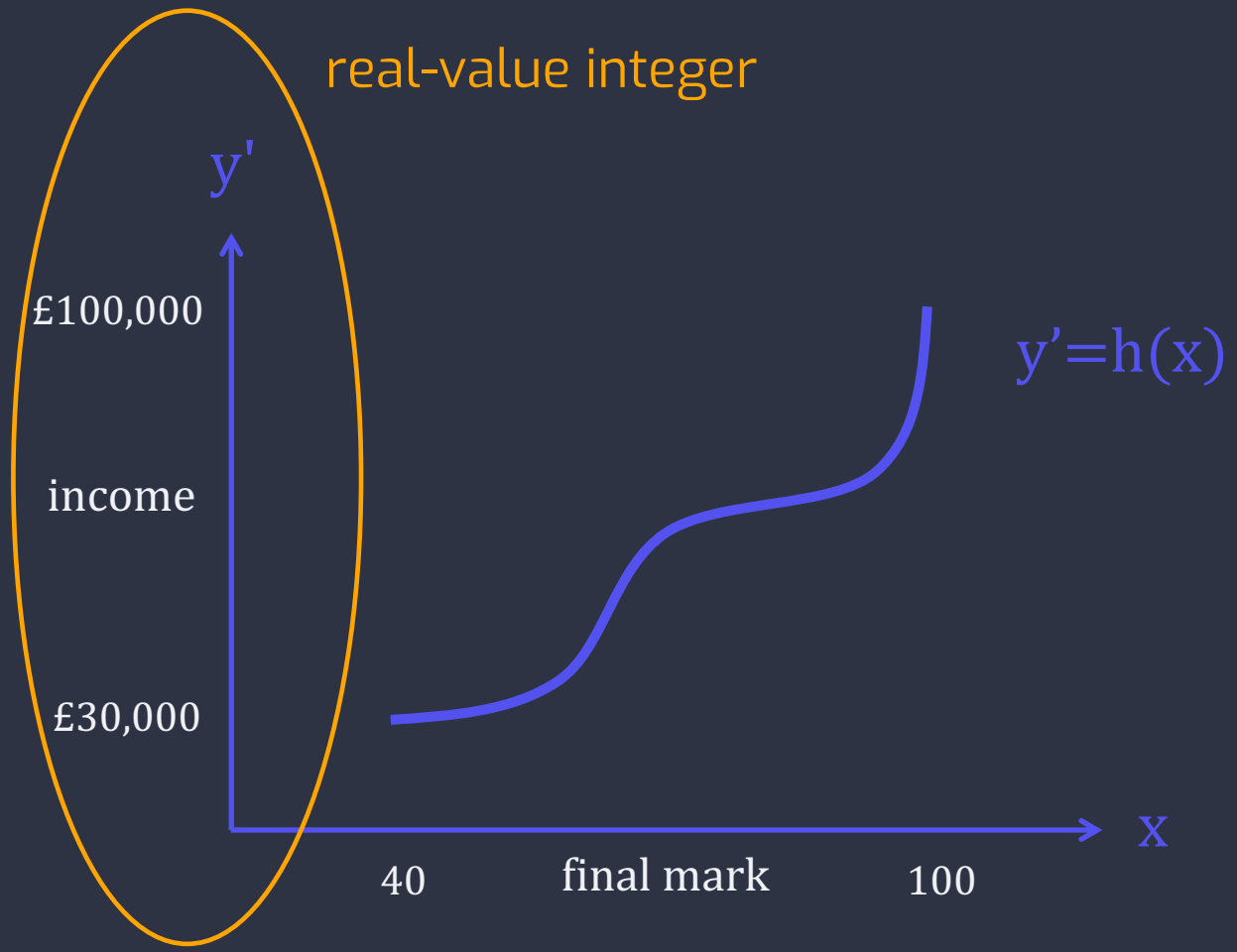
Regression

Regression



- Trying to learn from labelled input-output pairs of instances.
- To find an approximation function that maps from input variables to numerical and continuous output variables.
- The output variables are real-valued data such as integers or floating-point values; often quantities e.g. sizes, weights and amounts.

EXAMPLE. Final mark of ML module to predict income



Regression

- Can have real-value or discrete input variables;
- Aims to make prediction of a quantity;
- Is a univariate regression task if the input variable is a single value;
- Is a multivariate regression task is the input variable is a vector containing a group of single values.

Regression

- To estimate / evaluate how well a regression model performs:

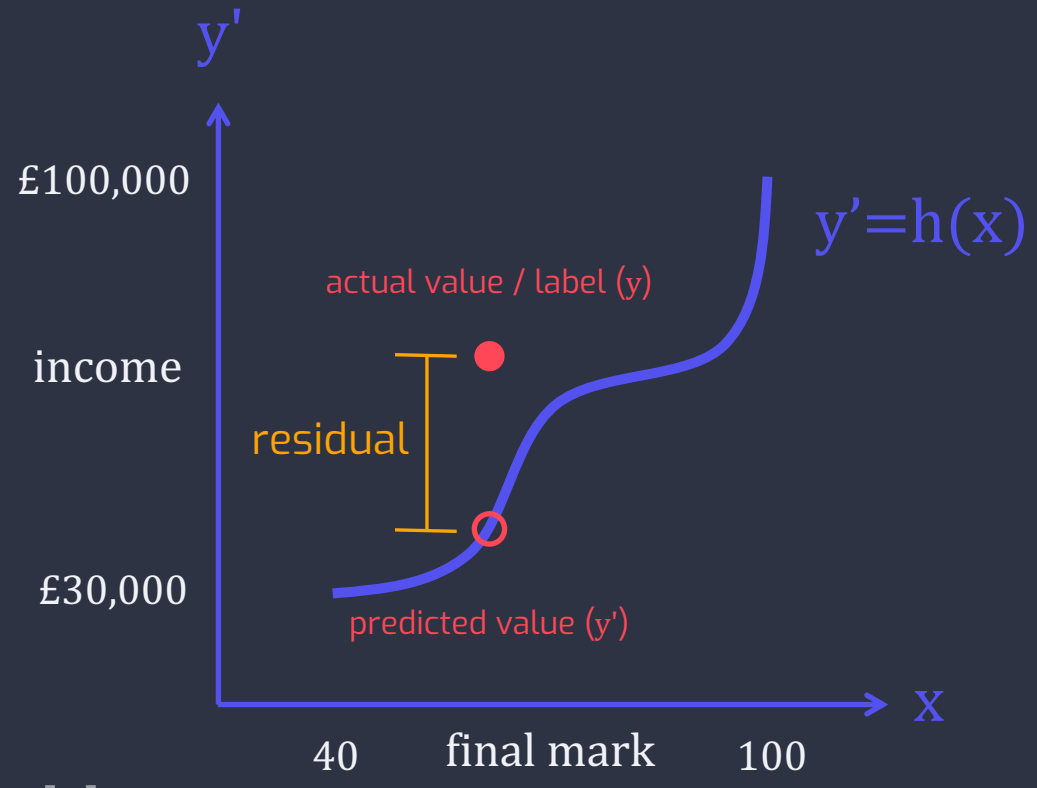
  Root Mean Squared Error (RMSE)

$$\text{RMSE} = \text{sqrt}(\text{average}(\text{error}\text{\^{}}2))$$

(standard deviation of the Residuals, i.e. prediction errors)

Regression

- Residuals : a measure of how far the regression line is from the data points.
- RMSE : a measure of how spread out these residuals are, i.e. how concentrated the data point is around the line of best fit.



EXAMPLE.

	predicted	actual
Student 1	£40k	£45k
Student 2	£80k	£70k

RMSE = sqrt(average(error^2))

= sqrt(((40 - 45)^2 + (80 - 70)^2) / 2)

= sqrt(62.5) = 7.91k

Regression

Regression Model --

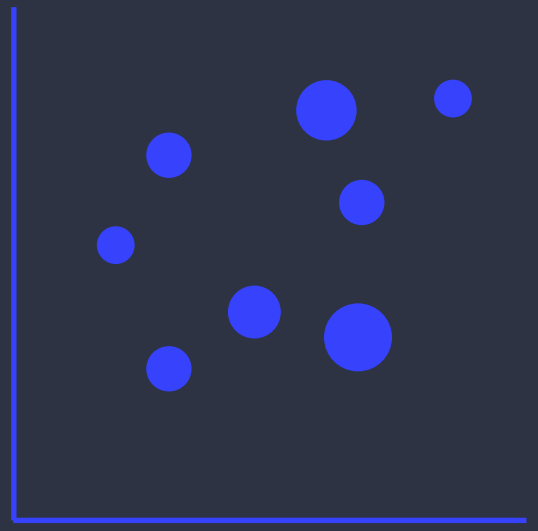
The model that makes such prediction.

Regression Algorithm --

The algorithm that learns from historical data to create a regression predictive model.

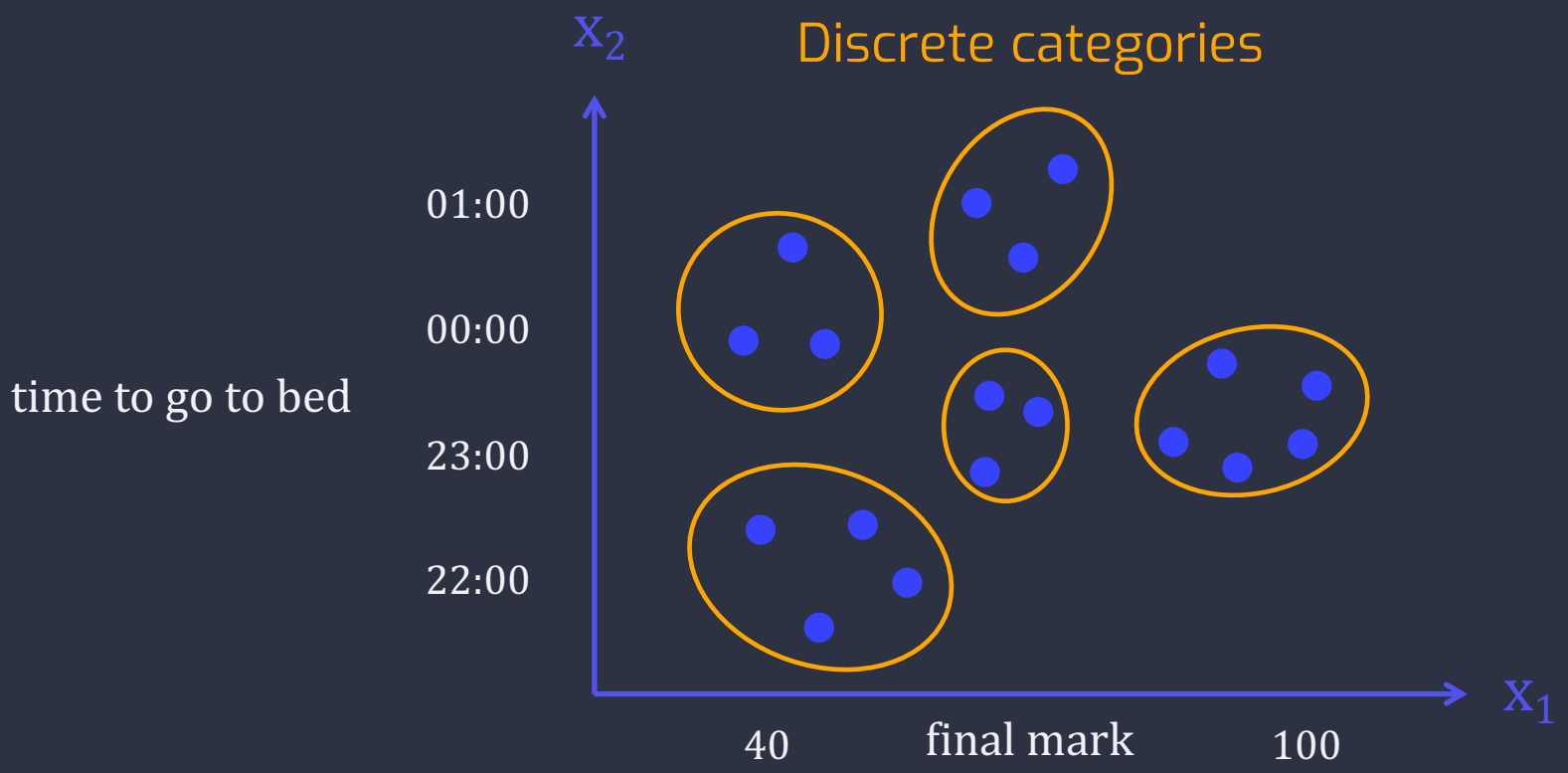
Classification

Classification



- Trying to learn from labelled input-output pairs of instances.
- To find an approximation function that maps from input variables to categorical and discrete output variables.
- The output variables are called categories / labels.

EXAMPLE. Final mark of ML module to predict degree award



Classification

- Can have real-value or discrete input variables;
- Aims to make prediction of a class or category;
- Is a binary classification task if the output variable could be 2 classes;
- Is multi-class classification task is the output variable could be >2 classes;

Classification

- To estimate / evaluate how well a classification model performs:

$$\text{accuracy} = \text{correct predictions} / \text{total predictions}$$

EXAMPLE.

If our model made predictions on 50 students' UG degree award, of which 45 were correct, then the classification accuracy would be: $\text{accuracy} = 45 / 50 = 90\%$

Other measurements: sensitivity specificity and precision

Classification

Classification Model --

The model that makes such prediction.

Classification Algorithm --

The algorithm that learns from historical data to create a classification predictive model

Regression versus Classification

Difference: prediction result / output variable

Regression

continuous quantity, e.g., 99, 19.85

Classification

discrete labels, e.g., {malignant, benign}, {orange, clementine, lemon}

Overlaps

Regression

may predict discrete values, but as integer quantity, e.g., a model could make predictions in a Likert scale, e.g. -2, -1, 0, 1, 2.

Classification

may predict continues values, but as probability for a class label, e.g. the output could be 98.6% and it means the model is 98.6% sure the cancer is malignant.

Method to estimate how they perform has no overlap

Regression

can be RMSE, but classification not

Classification

can be accuracy, but regression not

Takeaway Points

- Two types of supervised learning: regression & classification
- Main difference is output variables: continuous vs discrete
- Residual – difference between actual and predicted values
- RMSE to measure how spread out residuals are for regression
- Different method to estimate performance: RMSE vs accuracy