Model Evaluation & Validation

* Measures of Central Tendancy

* Mode - Most frequently occuring
  Median - Middle value when the data is ordered
  - Half way between the two middle values when there is an even number of points
  Mean - Average = $\sum_{i=1}^{n} x_i/n$

* Mode
  - distribution can have more than one mode (can be a range or set of values)
  - not all samples of a distⁿ contribute to it
  - not affected by outliers

  Mean
  - one value
  - all samples of a distⁿ contribute to it
  - affected by outliers

  Median
  - one value
  - all samples of a distⁿ contribute to it
  - not affected by outliers

* Order of measures

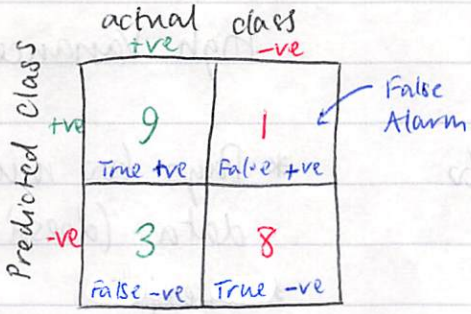
mode < median < mean

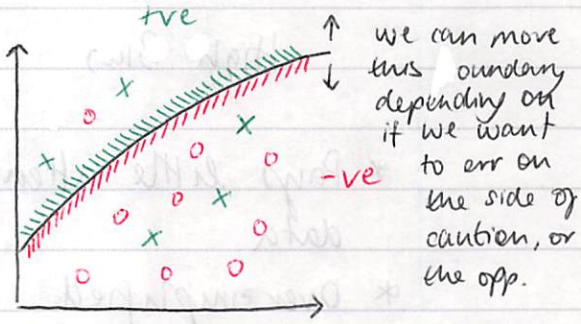
mode = median = mean

mode > median > mean

Evaluation Metrics

* Confusion Matrix Example:





* Precision & Recall

$$\text{Precision} = \mathbb{P}(\text{true} \mid \text{+ve}) = \frac{\#\text{ True +ve}}{\#\text{ True +ve} + \#\text{ False +ve}}$$

$$\text{Recall} = \mathbb{P}(\text{+ve} \mid \text{true}) = \frac{\#\text{ True +ve}}{\#\text{ True +ve} + \#\text{ False -ve}}$$

Terminology: { True, False } refers to the correctness of the prediction; { +ve, -ve } refers to the prediction result

* Example:

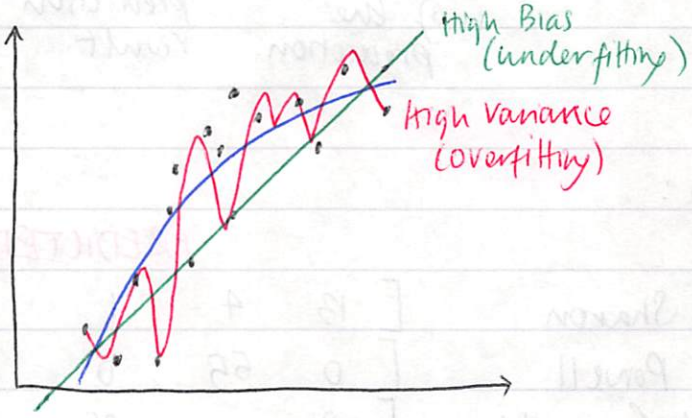
ACTUAL \ PREDICTED	Ariel Sharon	Colin Powell	Donald Rumsfeld	George W Bush	Gerhard Schroeder	Hugo Chavez	Tony Blair	precision	recall
Ariel Sharon	13	4	1	1	0	0	1		
Colin Powell	0	55	0	8	0	0	0		
Donald Rumsfeld	0	1	25	8	0	0	2		
George W Bush	0	3	0	123	0	0	1		
Gerhard Schroeder	0	1	0	7	14	0	4		
Hugo Chavez	0	3	0	2	1	10	0	= 1	= 5/8
Tony Blair	0	0	1	7	0	0	26		

Causes of Error

* Bias Variance Dilemma Solution

High Bias	High Variance
* Pays little attention to data	* Pays too much attention to data (doesn't generalise well)
* Oversimplified	* Overfits
* High error on training set (low r^2, high SSE)	* Much higher error on test set than training set
* Few features used	* Carefully minimised error over many features → optimised performance on training set

Want to use as few features as possible to
- maximise r^2
- minimise SSE



Data Types

* Numeric Data
  - Discrete
  - Continuous

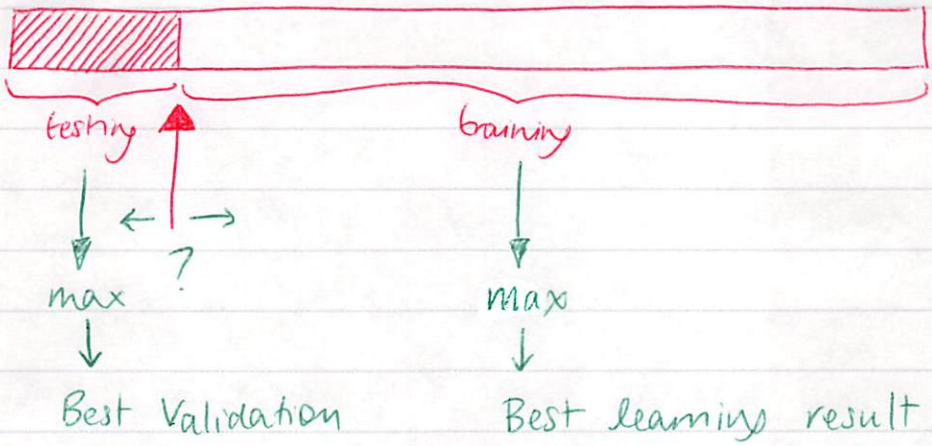
* Catagorical Data
  - Can take on numerical values but these don't have mathematical meaning i.e. mean or variance don't make sense
  - Ordinal data - Catagories with ordering / rank

* Time Series Data
  - Data point collected over time at

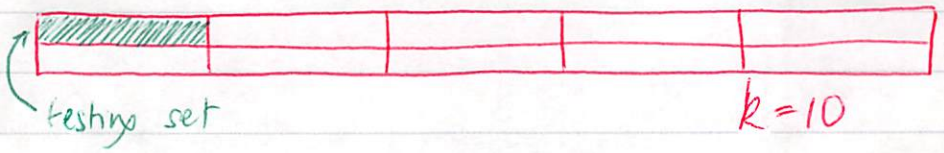
Cross Validation

k-Fold Cross Validation

Problems with splitting into training and testing data



K fold cross validation



Partion your data into k equal parts and run k separate learning experiments. for each experiment

- pick a different testing set
- train on the remainy data

Average test results from those k experiments

This way we use all our data for both training and testing. Clearly training will take longer in this case than the simple train then test method of validation but k-fold cross validation results in better accuracy in our model.

Representative Power of a Model

* Curse of Dimensionality

As the number of number of features or equivalently dimensions grows, the amount of data we need to generalize accurately, grows exponentially

→ You're better off getting more data than you are adding more features!