Model Selection and Evaluation in supervised learning

Daryn Ramsden
thisisdaryn@gmail.com

Last updated: June 4, 2019

Two types of learning problems:

- **Supervised learning**:
  - training data available
  - correct answers are available for training data
- **Unsupervised learning**:
  - correct answers not available
  - Used to draw inferences about the data or perhaps make suggestions about how to group the data

Common supervised learning tasks

- **Regression**: Estimating the relationship between a dependent variable and independent variables.
  - Possible output values along a continuum
  - Used when you want to predict a value e.g. the
- **Classification**: Predicting which group an observation belongs to based on values of independent variables
  - Output values are from a small known set

Regression Types

Regression can be done to fit data to numerous functional forms, including:

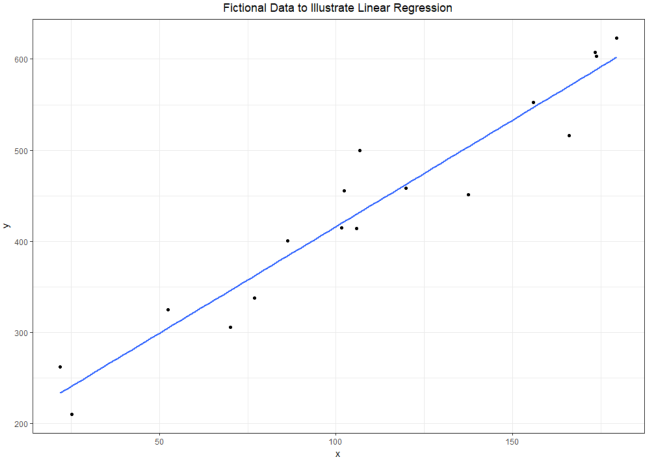
- linear
- logistic
- quadratic

Generally up to the analyst to choose which is appropriate.

Common form: Linear Regression

If n independent variables, find a model of the form:

$$\bar{y} = \alpha + \beta_1 x_1 \ldots \beta_n x_n \tag{1}$$



Evaluating Regression Fit

You need a means of telling how well your regression model performs.
This requires an error metric:

- quantifies the discrepancy between model predictions and actual values
- sum of squared error is a common choice

The Danger of Overfitting

Regression models face the danger of **overfitting**: the model fits the training data too closely and is not as successful when applied to other data sets.

Results from:

- too many variables relative to the number of observations
- unnecessarily complex choice of models

Using Training, Cross-validation and Test sets

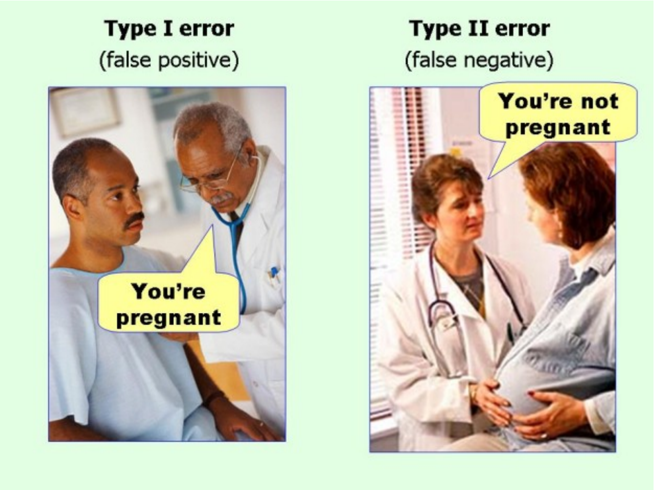
- **Best practice:** if possible, partition available data into **training**, **cross-validation**, and **test sets**.
  - Steps:
    1. Use training set to formulate candidate models
    2. Apply good candidates to the cross-over set to choose best model
    3. Evaluate generalisation error by using the chosen model on the test set
  - If available data is sufficient in quantity, **60-20-20** ratio is suggested
- Alternative: appraise model over a trial period

Using regularisation and Information Criteria to combat overfitting

- **Regularisation**:
- **Information Criteria**: heuristics that penalise model complexity, often quantified by the number of model parameters (variables). Examples are:
  - **Akaike Information Criterion (AIC)** (very common)
  - **Bayesian Information Criterion (BIC)**

Appraising Classification Models

Often a tradeoff between Type 1 and Type 2 errors



The Confusion Matrix (2 Classes)

		Prediction outcome		
		Yes	**No**	**Total**
Actual value	**Yes**	True Positive	False Negative	TP + FN
	No	False Positive	True Negative	FP + TN
	Total	TP + FP	FN + TN	

Sensitivity: What fraction of the time do you correctly identify positive results?

Prediction outcome

		Yes	No	Total
Actual value	**Yes**	True Positive	False Negative	TP + FN

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Specificity: What fraction of the time do you correctly identify negative results?

Prediction outcome

		Yes	**No**	**Total**
Actual value	**No**	False Positive	True Negative	FP + TN

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

Positive Predictive Value: can you trust a positive result?

		Prediction outcome
		Yes
Actual value	**Yes**	True Positive
	No	False Positive
	Total	TP + FP

$$\text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Negative Predictive Value: can you trust a negative result?

		Prediction outcome
		No
Actual value	**Yes**	False Negative
	No	True Negative
	Total	FN + TN

$$\text{NPV} = \frac{\text{TN}}{\text{FN} + \text{TN}}$$

Main takeaway

- Different metrics are generally at odds with each other e.g. have to sacrifice specificity to improve sensitivity
- Problem-specific knowledge is necessary to choose the best metric
  - Are false negatives or false positives more important to avoid?