Data Mining 2

Topic 03 : Review of Model Building

Lecture 02 : Regression Models

Dr Kieran Murphy

Department of Computing and Mathematics, WIT.
(kmurphy@wit.ie)

Spring Semester, 2022

Outline

- Regression Models

Outline

Regression Models (Evaluating Numeric Prediction)

We have covered using the MSE

$$\text{MSE} = \frac{1}{m} \sum \left(f\left(\boldsymbol{X}^{(j)}; \boldsymbol{\theta}\right) - y^{(j)} \right)^2$$

as the cost function in our curve fitting example. Geometrically this is computed as follows*

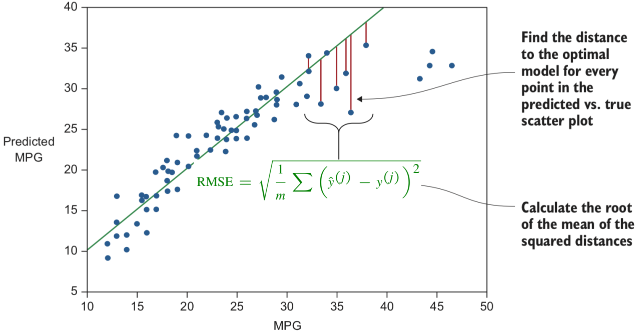

*Diagram (from *Real World Machine Learning*) shows the $\text{RMSE} = \sqrt{\text{MSE}}$

Common Cost Functions in Regression Models

Measure	Definition	Purpose/Advantage
Mean square error (MSE)	$\frac{(p_1-a_1)^2+\cdots+(p_m-a_m)^2}{m}$	Mathematically tractable but places greater emphasise on observations with large error
Root mean square error (RMSE)	$\sqrt{\frac{(p_1-a_1)^2+\cdots+(p_m-a_m)^2}{m}}$	Has same units as data
Mean absolute error (RMAE)	$\frac{\lvert p_1-a_1\rvert+\cdots+\lvert p_m-a_m\rvert}{m}$	Does not overemphasise observations with large error (as MSE does)
Relative square error (RSE)	$\frac{(p_1-a_1)^2+\cdots+(p_m-a_m)^2}{(p_1-\bar{a})^2+\cdots+(p_m-\bar{a})^2}$	Relative metric compares the error in the predictions with errors in the simplest model possible (a model just always predicting the average value of y)
Root Relative square error (RRSE)	$\sqrt{\frac{(p_1-a_1)^2+\cdots+(p_m-a_m)^2}{(p_1-\bar{a})^2+\cdots+(p_m-\bar{a})^2}}$	
Relative absolute error (RAE)	$\frac{\lvert p_1-a_1\rvert+\cdots+\lvert p_m-a_m\rvert}{\lvert p_1-\bar{a}\rvert+\cdots+\lvert p_m-\bar{a}\rvert}$	

where a_j is the actual value, p_j is the predicted value, m is the number of observations, and $\bar{a}$ represents the mean of the a_j.

Assumptions of (Linear) Regression Model

- **Multivariate normality** — each of the independent variables must be normally distributed.
  - Graphical: histograms, Q-Q plots,
  - Numerical: goodness of fit tests, e.g., the Kolmogorov-Smirnov test, ...
  - Fix: non-linear transformations such as log, power, Box-Cox, etc
- No or little **multicollinearity** — independent variables should not be too highly correlated with each other.
  - Numerical: correlation matrix using Pearson?s bivariate correlation coefficient.
  - Fix: Centre the data, filter out some of the independent variables,
- No **auto-correlation** — the residuals should be independent, and normally distributed.
  - Graphical: residual plot.
  - Numerical: Durbin-Watson test.
- **homoscedasticity** — constant variance in residuals.
  - Graphical: residual plot.
  - : Fix: transform data or use non-linear model.

And, in addition, for linear regression

- **Linearity** — relationship between the independent variables and the dependent variable is linear.