Week 9: Regularisation

① The Big Picture

◎ Things that influence your Bias and Variance trade off (Week 1)

① Understanding Your Data

Visualisation
- Week 5 • Parallel co-ord Plot
         • Mosaic Plot
Dimension Reduction
         Week 5
         • Tour

Unsupervised Learning
         Week 4
         • LDA (Dim reduction)
         • PCA

② Selecting Your Model

Parametric Models

Regression
- Week 2
  • Linear Models (GLM, Splines, etc)

Classification
- Week 3
  • Logit
  • LDA classification

Non-Parametric Models
- Week 6
  • Regression Trees
- Week 6
  • Classification Trees
- Week 8
  • Neural Networks
- Week 7
  • Support vector Machines

③ Understanding Your MODEL

Regularisation

Regularisation on The model
- Week 7
  • Ensemble Models
    ▷ Boosting Week 9
    ▷ Bagging — Lasso/ridge
              — subset selection

Regularisation on The estimators
- Week 3
  • Boot strapping
  • K-CV
  • Test & Training

Visualisation
- Week 1
  • Residual & other diagnostic plots

Week 9
Model Assessment
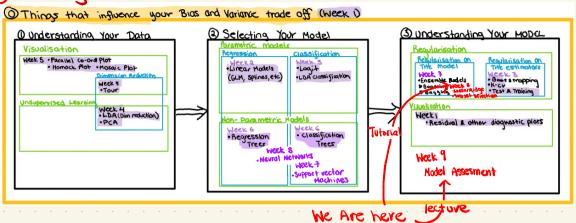
We Are here   lecture

② Subset Selection

Select a Subset of Variables using an algorithm.

Best: Fit EVERY possible model & pick the one with the lowest RSS for each number of variables. *Model can completely change between # of Pred

Forward: Start with the null model (only the intercept) and add the variable that decreases the RSS the most. Select the number of variables using Cross Validation.

e.g. M_0 (null model): $\hat{y} = C$

M_1 : $\hat{y} = C + B_1 x_5$ (doesn't leave)

M_2 : $\hat{y} = C + B_{12} x_5 + B_{22} x_1$

M_p $\hat{y} = C + B_{1p} x_5 + B_{2p} x_1 + \ldots B_{pp} x_2$

Backward: Start with the full model (all variables M_p) and remove the variable that increases the RSS by the least to get M_{p-1}. Continue until you have all $M_0 - M_p$ models. Select the final model using Cross validation.

Note: RSS CANNOT be used to compare models with different # of predictors, which is why we compare final models with Cross validation.