Putting the CART before the horse

Luke Hayden

26th September 2018

Classic modelling

Multiple Regression approach

Simple linear regression:

Age = X(marker1) + c

We try to find values for x & c that come as close as possible to solving the equation for each set of values for *Age* and *marker1* we have.

Two predictors:

Age = X(marker1) + Y(marker2) + c

Many predictors

Age = X(marker1) + Y(marker2) + Z(marker3) + W(marker4) + ... + c

Where we have many different markers, we can find values of x,y,z,w, etc that solve this equation very well but don't provide

How do we avoid overfitting?

We want:

Modelling approach that can capture the signal without simply reproducing all the noise present in our dataset
To maximise predictive power

Approaches:

Data partitioning:

train-test split
cross-validation)

Model type

Ensemble methods!

Model parameters

Exploring parameter space

Machine Learning terminology

Supervised vs unsupervised learning

Unsupervised learning: find the shape of the data (
(eg: PCA, kmeans clustering)
Supervised learning: train an algorithm to recapitulate the examples it sees in a dataset
(eg: linear regression)

Classification vs Regression

Classification: categorise examples into one of a number of discrete categories
Regression: determine value along range

Tree ensemble approaches

Decision tree

Classify or perform regression by asking binary questions of data: whether value of marker X is above or below key value Y, whther marker Z is above or below. . . ..

Random Forest

Ensemble of decision trees, each using a random subset of the predictors to classify/perform regression on a random subset of the data

Resists overfitting

Gradient Boosting Machine

Start with simple model (eg: mean of values in training dataset)

Random Forest parameters

ntree: number of trees

mtry: Number of variables randomly sampled as candidates at each split

min.node.size: sets depth of trees

cross-validation folds: number of repartitions of data for testing

splitting model: variance or “extratrees”

My project as example

Project

Examine the effect of regeneration on the molecular age profile of *Parhyale* limbs

Designing codeset

*Nanostring as method to quantify gene expression
*200 genes in codeset
-195 genes chosen on the basis of differential expression analysis
-5 control genes: do not vary in expression between conditions