Data Science

Cross Validation

October 8th, 2014

Jarret Petrillo, jarret.petrillo@gmail.com

The First Premise of Statistics

Blind Experiments

sample mean = sum(x)/n

Two treatment groups
treatment effect = (mean 1) - (mean 2)

Under the assumption of Normalty
treatment effect ~ t_distribution

Cross Valication Reproduces Blind Experiment

- *Divide sample and for each model predict y values*
- *Decide on an evaluation metric - is the average response the right metric?*
- *Compare model treatments across groups*

Cross Validation Has No Benchmark

- *In our first blind experiment we began with a modelled distribution*
- *Under normalty two sample means are distributed asymptotically as a t distribution*
- *An expected distribution allows a researcher to form and test hypotheses*