FEATURE SELECTION

THE 'REAL' DATASET

- The dataset we work on so far cannot be compared to the real dataset we will face from now on.

- The world nowadays is the world of "Big Data".
  - It means the data is really big
  - How big?
    - Millions of rows
    - Hundreds or Thousands of columns

DEALING WITH BIG DATA

- What can we do with that?
  - Obviously we can just do the same thing we did.
    - However, it will take too much time and resources
    - aka wasting money
  - Or we can work on just some features.
    - Not all features are equally important.
    - But how do we know which features are important?

> Feature Selection: A process of selecting the most important features of the dataset

FEATURE SELECTION

- How do we select such features?
  - What are “good” features?

FEATURE SELECTION

- How do we select such features?
  - We need some statistics here.
  - We need to find “correlation”
  - Correlation measures how likely one attribute relates to another attribute.
  - For example, $y = ax + b$
    - x correlates with y

ANALYSIS OF VARIANCE (ANOVA)

- We will explore 2 ways, using an analysis of variance, to perform feature selection.
  - f-regression
  - chi^2

F-REGRESSION

- Univariate linear regression test
  - Calculate correlation of features and class attribute. (F-Score)
- http://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.f_regression.html

CHI-SQUARE

- Compute chi^2 value between each non-negative attributes and class attribute.
  - Only non-negative
  - Testing of goodness of fit
  - Testing of homogeneity

  - Chi^2 can be used to compute p-value for null hypothesis testing.
  - http://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.chi2.html