Statistical Modeling Framework

Dihui Lai

dlai@wustl.edu

March 17, 2019

Building Statistical Models: Likelihood/Loss Function

- We are planting some seeds in your garden. What kind of distribution shall we use if we want to know the number of seeds germinate depending on the amount of water and fertilizer (Poisson/Logistic).
- You were asked to develop an algorithm that makes predictions about the future sale prices of homes (Gaussian).
- A researcher is interested in how variables, such as GRE (Graduate Record Exam scores), GPA (grade point average) and prestige of the undergraduate institution, effect admission into graduate school (Logistic).
- A health-related researcher is studying the number of hospital visits in past 12 months by senior citizens in a community based on the characteristics of the individuals and the types of health plans under which each one is covered (Poisson).

How to select a Likelihood/Loss function?

- Determine the distribution of the target variable e.g. histogram
- Make a guess and verify
- Use the likelihood/loss function that is given to you (by your boss, a client or a stakeholder)

How to select a Likelihood/Loss function ... More...?

- MNIST handwriting data set classification (soft-max)
- Speech recognition/Optical character recognition (connectionist temporal classification)

Goodness of Fit

The metrics to measure the goodness of a model fitting

- Liklihood function

$$\ell = \sum_j \left[y^j \theta^j - b(\theta^j) \right]$$

- Deviance

$$D = \ell_{max} - \ell(\theta(\hat{\beta}))$$

  where ell_{max} is the log likelihood of the saturated model.

- AIC:

$$\mathrm{AIC} \; = \; 2k - 2\ln(\ell)$$

  Here, k is the number of estimated parameters in the model. Besides maximizing the likelihood, AIC also penalize the complexity of a model.

Goodness of Fit: Continued

- BIC:

  $$\mathrm{BIC} = \log(n)k - 2\ln(\ell)$$

  Here, n is the number of data points. Similar to AIC but penalize more on the model compelxity weighted by the amount of data.

Variable Selection

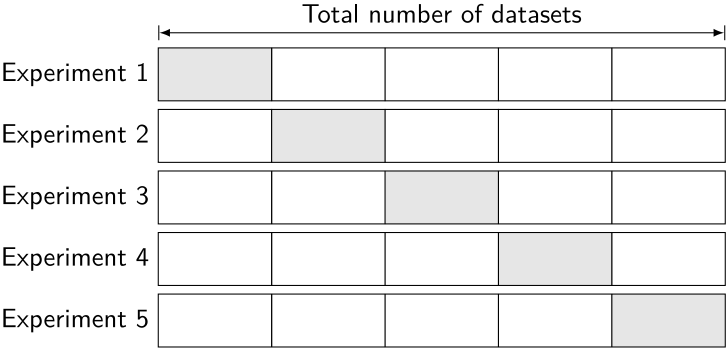
- Determine the contribution of a variable to the following metrics: Likelihood/Loss function, Deviance, AIC and BIC
- Beware of target leak.
- Consider the context of application. Is the variable available in application scenario?
- Does it make intuitive sense?
- Variable's statistic significance, p-value (reject the variable if p-value is above certain threshold e.g. > 0.05, > 0.01 etc.).

Feature Engineer

- Categorical variables
  - Regrouping
  - Converting to numeric version
- Numeric variables
  - Function transformation: polynomial, spline, power function, exponential, log etc.
- Variable interactions
- Features from sub-models

Overfitting and Cross validation

- When a model becomes over-complex, it starts to fit noises rather than the true pattern
- Training data v.s. validation data: one way to prevent overfitting is to split data into training set and validation set. If the model peforms significantly worse in the validationd data, it shows a sign of overfitting.
- Cross validation.



Total number of datasets

Experiment 1

Experiment 2

Experiment 3

Experiment 4

Experiment 5