Lecture 05: Statistical Modeling

Dihui Lai

dlai@wustl.edu

September 23, 2019

Generalized Linear Model

Exponential family of probability density function

$$f(y) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right)$$

The distribution have the following properties

- $E(Y) = b'(\theta)$
- $Var(Y) = b''(\theta)a(\phi)$

Link Function

Assume a linear model that models the mean of the distribution

$$b'(\theta) = g(\eta) = g(\vec{x} \cdot \vec{\beta})$$

here $g^{-1}(\cdot)$ is called the link function.

Log Likelihood Function of GLM

The log likelihood function of GLM

$$\ell = \sum_i \frac{y^i \theta^i - b(\theta^i)}{a(\phi)} + c^i(y^i, \phi)$$

In the model, only θ is depedent on $\vec{x} \cdot \vec{\beta}$. Therefore, "maximize" the likelihood function is equivalent to minimize

$$\ell = -2 \sum_i \left[y^i \theta^i - b(\theta^i) \right]$$

Examples of GLM

- Normal: $\log f(y_i, \theta_i, \phi) = -\frac{(y_i - \mu_i)^2}{2\phi} + C$
  Canonical link function
  $$\mu_i = \vec{x^i} \cdot \vec{\beta}$$
- Poisson:$\log f(y_i, \theta_i, \phi) = y_i \log(\lambda_i) - \lambda_i + C$
  Canonical link function
  $$\log(\lambda_i) = \vec{x^i} \cdot \vec{\beta}$$
- Binomial: $\log f(y_i, \theta_i, \phi) = y_i \log(\frac{p_i}{1-p_i}) - \log(1 - p_i) + C$
  Canonical link function
  $$\log(\frac{p_i}{1 - p_i}) = \vec{x^i} \cdot \vec{\beta}$$

The Gradient and Hessian; Optimization quantity

The gradient can be derived as

$$\frac{\partial \ell}{\partial \beta_j} = -2\sum_i (y^i - b'(\theta^i))\frac{\partial \theta^i}{\partial \beta_j}$$

$$\frac{\partial \ell}{\partial \beta_j} = -2\sum_i (y^i - \mu^i)\frac{g'(\eta^i)}{V(\mu^i)}x_j^i$$

The hessian can be derived as

$$\frac{\partial^2 \ell}{\partial \beta_k \partial \beta_j} = 2\sum_i \left[\frac{g'(\eta^i)^2}{V(\mu^i)} - (y^i - \mu^i)\frac{g''(\eta^i)V(\mu^i) - g'(\eta^i)^2 V'(\mu^i)}{V(\mu^i)^2}\right] x_j^i x_k^i$$

Statistical Modeling Framework

- A quantity to be optimized: Sum of squared errors/Likelihood function/Loss function
- A model that explains the predictors and the target variable, directly or indirectly
- An optimization algorithm: Newton-Raphson; greedy search etc..

Statistical Modeling Practical Steps

- Data wrangling, visualization, basic statistic analysis
- Determine the property of the target variable and the loss function to use;
- Variable selection; feature engineer
- Model quality control; prevent overfitting;

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; unique values; single variable statistics
- Starting with something that you are familiar with; use the likelihood/loss function that is given to you (by your boss, a client or any stakeholders)
- Pentalty terms: L1/L2 norm
  - L1 regularization $\ell + \alpha \sum_{j=1}^{m} |\beta_j|$
  - L2 regularization $\ell + \alpha \sum_{j=1}^{m} \beta_j^2$

Goodness of Fit

The metrics to measure the goodness of a model fitting
In Ordinary Least Square (OLS), this could be the sum square errors.
In GLM this could be one of the followings

- Liklihood function

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

- Deviance

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

  where ℓ_{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

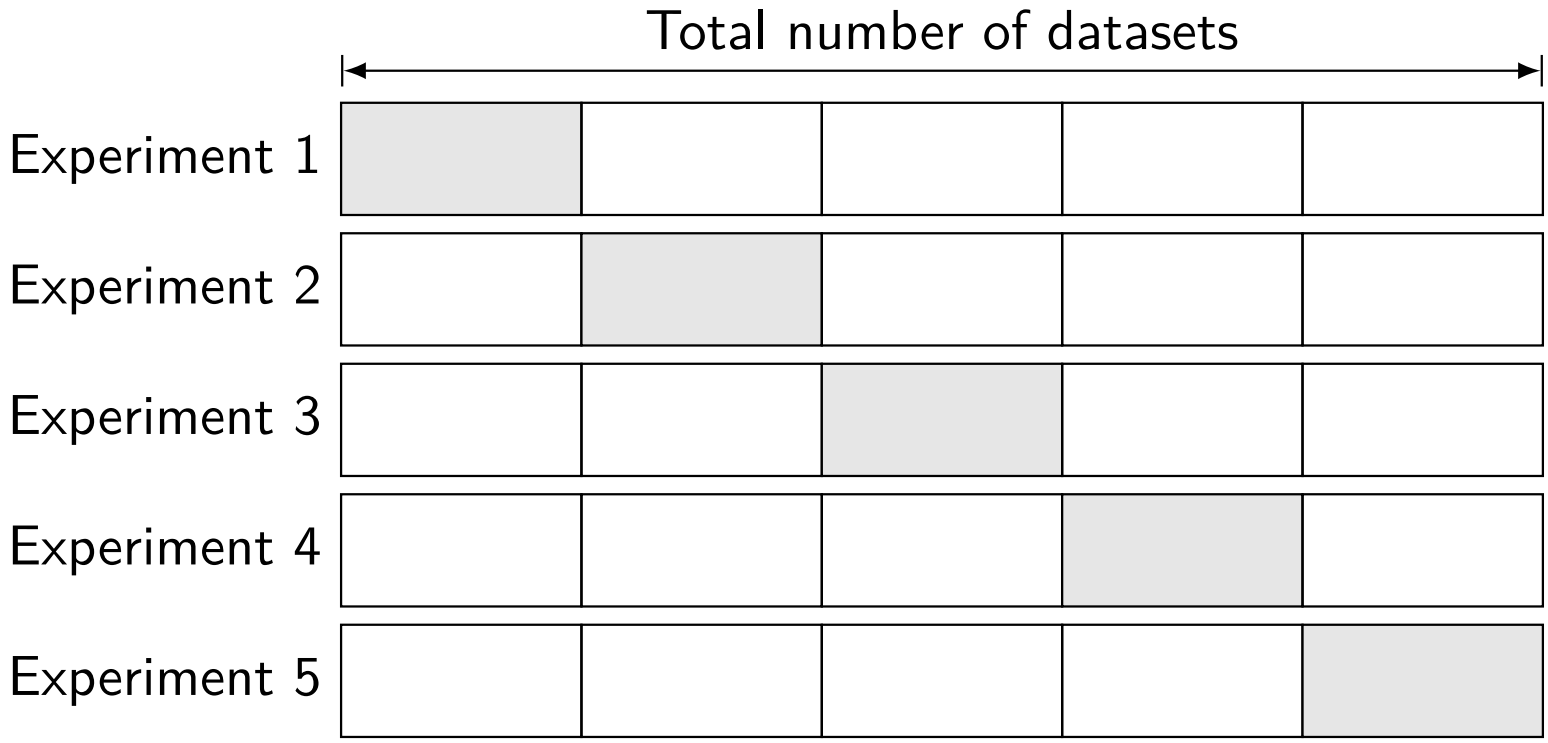
- Determine the contribution of a variable to the following metrics: Likelihood/Loss function, Deviance, AIC and BIC
- Beware of target leak.
- Consider the context of application. Is the variable available in application scenario?
- Does it make intuitive sense?
- Variable's statistic significance, 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 & 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