IIT ROORKEE

NPTEL ONLINE CERTIFICATION COURSE

LOGISTIC REGRESSION

LECTURE 46

DR. GAURAV DIXIT

DEPARTMENT OF MANAGEMENT STUDIES

LOGISTIC REGRESSION

- Equivalent of linear regression for categorical outcome variable
  - Predictors can be categorical or continuous
- Applied in following tasks
  - Classification task
    - Predicting the class of a new observation
  - Profiling
    - Understanding similarities and differences among groups

LOGISTIC REGRESSION

- Steps for logistic regression
  - Estimate probabilities of class memberships
  - Classify observations using probabilities values
    - Most probable class method: assign the observation to the class with highest probability value
      - Equivalently, for a two-class case, cutoff value of 0.5 can be used
    - Class of interest: user specified cutoff value
      - For a two-class case, typically a value greater than average probability value for class of interest, but less than 0.5 can be used

LOGISTIC REGRESSION

- Logistic Regression Model
  - Used typically in cases when structured model is preferred over data-driven models for classification tasks
  - Categorical outcome variable cannot be directly modeled as a linear function of predictors
    - Inability to apply various mathematical operators
    - Variable type mismatches
    - Range reasonability issues
      - LHS range={0, …, m-1}
      - RHS range=(-∞, ∞)

LOGISTIC REGRESSION

- Logistic Regression Model
  - Instead of using outcome variable (Y) in the model, a function of Y, called *logit* is used
- Logit
  - Think about modeling probability value as a linear function of predictors, specifically in a two-class case

If P is the probability of class 1 membership

$$P = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

Where p is the no. of predictors

LOGISTIC REGRESSION

- Logit
  - LHS range improves from {0, 1} to [0, 1], however still cannot match RHS
  - Can we bring RHS range to [0,1]?
    - Nonlinear approach
  - Typically, a nonlinear function of the following form is used to perform the required transformation

$$P = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)}}$$

This function is called *logistic response function*

LOGISTIC REGRESSION

- Logit
  - Rearrange the previous equation as below:

$$\frac{P}{1-P} = e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}$$

LHS is expression for *odds*, another measure of class membership

$$odds = \frac{P}{1-P}$$

  - Odds of belonging to a class is defined as ratio of probability of class 1 membership to probability of class 0 membership
    - This metric is popular in sports, horse racing, gambling, and many other areas

LOGISTIC REGRESSION

- Logit
  - Previous equation can be rewritten as

$$odds = e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}$$

    - Range is now (0, ∞)
  - Take log on both sides of previous equation

$$\log(odds) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

    - Standard logistic model
    - Now, LHS and RHS both have same range (-∞, ∞)
  - Log(odds) is called logit
    - Logit is used as the outcome variable in the model instead of categorical Y

LOGISTIC REGRESSION

- Odds and logit can be written as a function of probability of class 1 membership
  - Open RStudio

- In logistic regression model, we predict the logit values and therefore corresponding probability of a categorical outcome
  - Predicted probabilities values become the basis for classification
  - A prediction model for classification task

Key References

- Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data by EMC Education Services (2015)
- Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner by Shmueli, G., Patel, N. R., & Bruce, P. C. (2010)

Thanks…