

BUAD 5082

Machine Learning 2

Generalized Linear Models

Agenda

- Brief word on learning algorithms covered so far.
- Introduction to Generalized Linear Models (glm's)
- Poisson Regression example

Modeling methods discussed so far...

1. KNN and KNN.Reg
2. ?
3. ?
4. ?
5. ?
6. ?
7. ?
8. ?
9. ?
10. ?
11. ?
12. ?
13. ?
14. ?
15. ?
16. ?
17. ?
18. ?
19. Local Regression

Modeling Methods discussed so far...

1. KNN and KNN.Reg
2. Simple Linear Regression
- 3.
4. Logistic Regression
- 5.
- 6.
7. Subset Selection (Best, Forward, Backward)
- 8.
9. The Lasso
- 10.
11. Partial Least Squares Regression
12. Maximal-Margin Separating Hyperplanes
- 13.
- 14.
15. (a Basis Function method)
16. Step Function Regression (also a Basis Function method)
17. (also a Basis Function method)
18. Smoothing Splines (**not** a Basis Function method)
19. (**not** a Basis Function method)

Modeling Methods discussed so far...

1. KNN and KNN.Reg
2. Simple Linear Regression
3. Multiple Linear Regression
4. Logistic Regression
5. LDA
6. QDA
7. Subset Selection (Best, Forward, Backward)
8. Ridge Regression
9. The Lasso
10. Principal Component Regression
11. Partial Least Squares Regression
12. Maximal-Margin Separating Hyperplanes
13. Support Vector Classifiers
14. Support Vector Machines
15. Polynomial Regression (a Basis Function method)
16. Step Function Regression (also a Basis Function method)
17. Regression Splines (also a Basis Function method)
18. Smoothing Splines (**not** a Basis Function method)
19. Local Regression (**not** a Basis Function method)

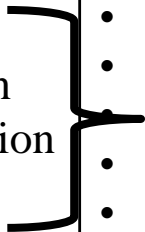
Summary So Far

| Dimensionality | | | |
|--------------------------|--|-------------|----------------------|
| | | 1 Predictor | 1 or More Predictors |
| Parametric Methods | | | |
| Nonparametric Methods | | | |

Summary So Far

| Dimensionality | | |
|-----------------------|--|---|
| | 1 Predictor | 1 or More Predictors |
| Parametric Methods | <ul style="list-style-type: none">• Simple Linear Regression• Basis Function Methods<ul style="list-style-type: none">• Polynomial Regression• Step Function Regression• Regression Splines | <ul style="list-style-type: none">• Multiple Linear Regression• Logistic Regression• LDA and QDA• Subset Selection• Ridge Regression and Lasso• MM Separating Hyperplanes, Support Vector Classifiers and Support Vector Machines• PCR and PLS Regression |
| Nonparametric Methods | <ul style="list-style-type: none">• Smoothing Splines• Local Regression | <ul style="list-style-type: none">• KNN and KNN.Reg |

GAM's Do This...

| | | Dimensionality | |
|-----------------------|--|---|---|
| | | 1 Predictor | 1 or More Predictors |
| Parametric Methods | <ul style="list-style-type: none">• Simple Linear Regression• Basis Function Methods<ul style="list-style-type: none">• Polynomial Regression• Step Function Regression• Regression Splines |  | <ul style="list-style-type: none">• Multiple Linear Regression• Logistic Regression• LDA and QDA• Subset Selection• Ridge Regression and Lasso• MM Separating Hyperplanes, Support Vector Classifiers and Support Vector Machines• PCR and PLS Regression |
| | <ul style="list-style-type: none">• Smoothing Splines• Local Regression | | <ul style="list-style-type: none">• KNN and KNN.Reg |

GENERALIZED LINEAR MODELS (GLM's)

The Generalized Linear Model (glm's)

- Ordinary linear regression predicts the expected value of a given unknown quantity (the *response variable*, a random variable) as a linear combination of a set of observed values (*predictors*).
- This implies that a constant change in a predictor leads to a constant change in the response variable (i.e. a *linear-response model*).
- This follows from an assumption that ε , the irreducible error in the population, is independently and identically distributed around the expected value of the response variable, and has mean 0 and constant variance σ^2 .

The Generalized Linear Model (glm's)

- However, these assumptions are inappropriate for some types of response variables.
 - For example, in cases where the response variable is expected to take on values that are:
 - counts
 - always positive (e.g. probabilities)
 - such that constant input changes lead to geometrically varying, rather than constantly varying, output changes (e.g. heteroscedasticity)

The Generalized Linear Model (glm's)

- As an example...
 - The problem with this kind of prediction model is that it would imply a temperature drop of 10 degrees would lead to 1,000 fewer people visiting the beach.
 - Therefore, a beach whose expected attendance was 50 at a higher temperature would now be predicted to have the impossible attendance value of -950.
 - Logically, a more realistic model would instead predict a constant *rate* of increased beach attendance (e.g. an increase in 10 degrees leads to a doubling in beach attendance, and a drop in 10 degrees leads to a halving in attendance). Such a model is termed an *exponential-response model* (or log-linear model, since the logarithm of the response is predicted to vary linearly).

The Generalized Linear Model (glm's)

- Similarly, a model that predicts a probability of making a yes/no choice (a Bernoulli variable) is even less suitable as a linear-response model, since probabilities are bounded on both ends (they must be between 0 and 1).
- For example...
 - Imagine a model that predicts the probability of a given person going to the beach as a function of temperature.
 - A reasonable model might predict that a change in 10 degrees makes a person two times more or less likely to go to the beach.
 - But what does "twice as likely" mean in terms of a probability? It cannot literally mean to double the probability value (e.g. 50% becomes 100%, 75% becomes 150%, etc.).
 - Rather, it is the odds that are doubling: from 2:1 odds, to 4:1 odds, to 8:1 odds, etc. Such a model is a *log-odds model*.

The Generalized Linear Model (glm's)

- Generalized linear models cover all these situations by allowing the expected value of response variables to have an arbitrary distribution, and for an arbitrary function of the response variable (the *link function*) to vary linearly with the predicted values (rather than assuming that the response itself must vary linearly).
- For example, the case above of predicted number of beach attendees would typically be modeled with a Poisson distribution and a log link, while the case of predicted probability of beach attendance would typically be modeled with a binomial distribution and a log-odds (or *logit*) link function.

The Generalized Linear Model (glm's)

- The `glm()` function in R supports the following distributions (and associated link functions):

| Family | Default Link Function |
|------------------|-----------------------|
| binomial | (link = "logit") |
| gaussian | (link = "identity") |
| Gamma | (link = "inverse") |
| inverse.gaussian | (link = "1/mu^2") |
| poisson | (link = "log") |
| quasibinomial | (link = "logit") |
| quasipoisson | (link = "log") |

- Response variables that are counts normally use one of the Poisson or binomial distributions (depending on the degree of variance).
- Gamma distributions can address heteroscedasticity in non-negative data. The inverse-Gaussian distribution also relaxes homoscedastic assumptions.
- See **GLMs.R** for a simple example of a typical Poisson Regression model.