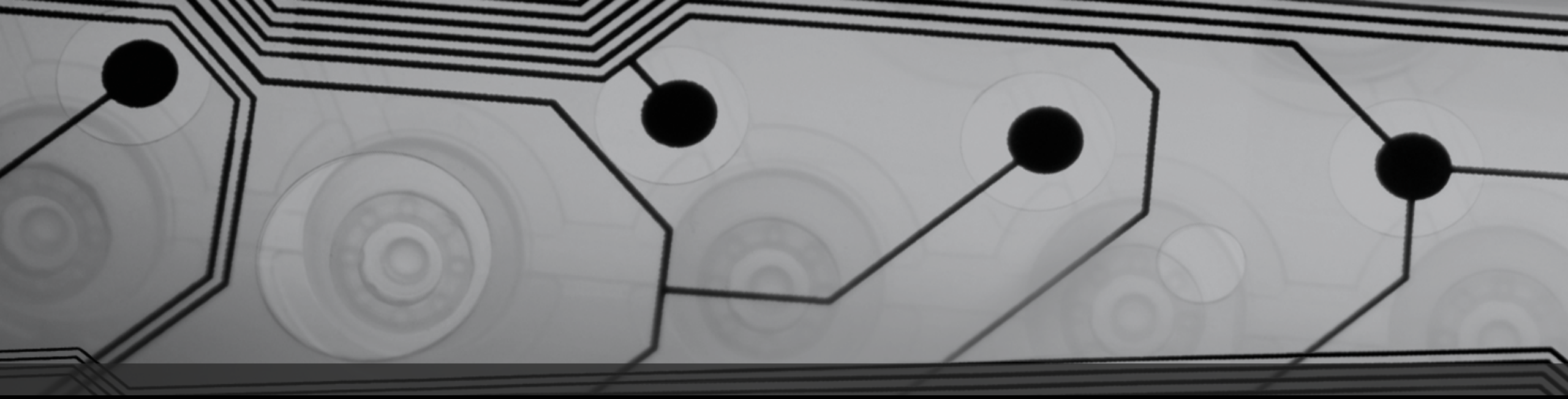


Machine Learning in R

A Primer on *stats::glm and glmnet* packages | Logistic Regression

Shaurya Jauhari, Mora Lab.

Agenda

- A word about *Docker*
- Background: Machine Learning
- Characterization: Machine Learning
- Linear Regression
- Logistic Regression
  - Theory
  - *Stats::glm()*
  - *glmnet* package
- General Concepts
- Dataset Description
- Hands-on session

Docker

- Provides for operating-system level virtualization
  - Repository name: shauryajauhari/lab_workshop_logistic_regression
- *Containers* hold the runtime binaries
  - Instances of an *Image*
- 3 essential components
  - Program script
  - Dockerfile ~ DESCRIPTION file in an R package
  - Base Image

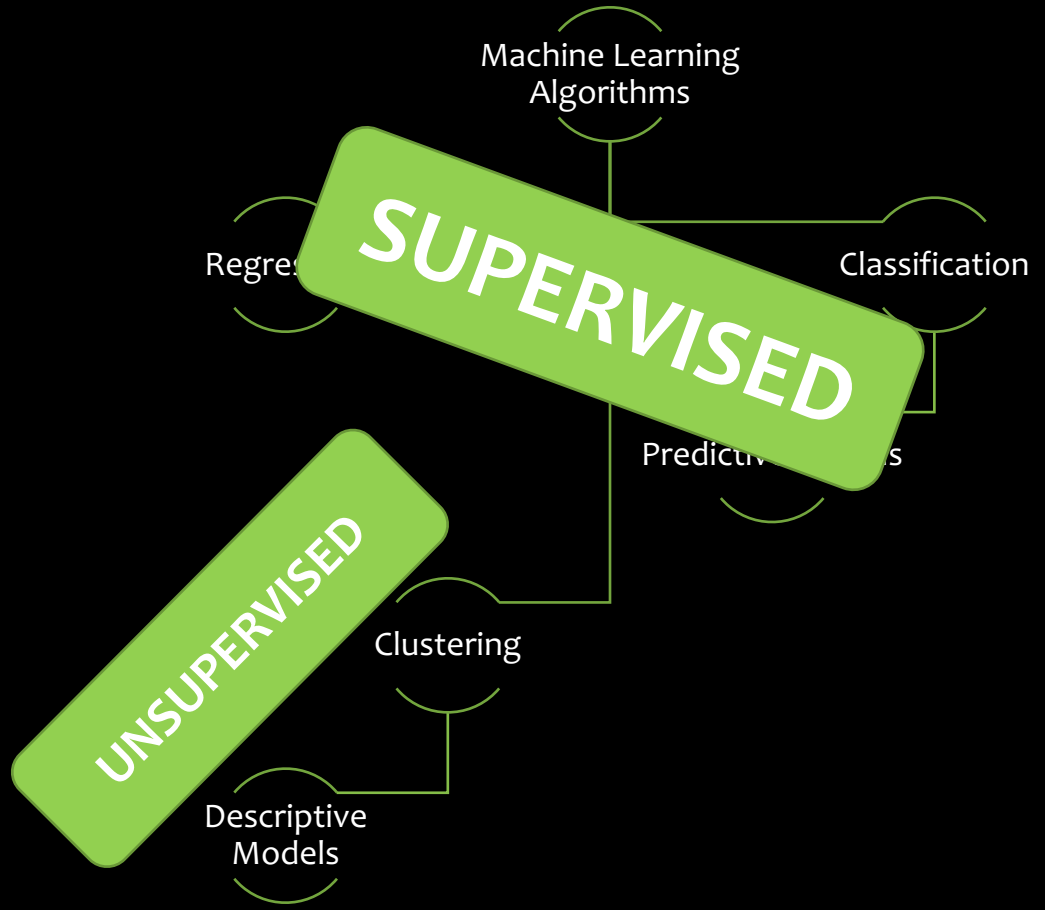
```
FROM rocker/r-base:latest
FROM rocker/verse:3.5.1
USER rstudio
WORKDIR /Users/mei/Desktop/Machine_Learning_Logistic_Regression/Logistic_Regression_Lab
RUN export PATH=$PATH:/Applications/RStudio.app/Contents/MacOS/pandoc
COPY ./glmnet_stats_glm_demo.Rmd glmnet_stats_glm_demo.Rmd
RUN R -e "rmarkdown::render('glmnet_stats_glm_demo.Rmd')"
```

Background

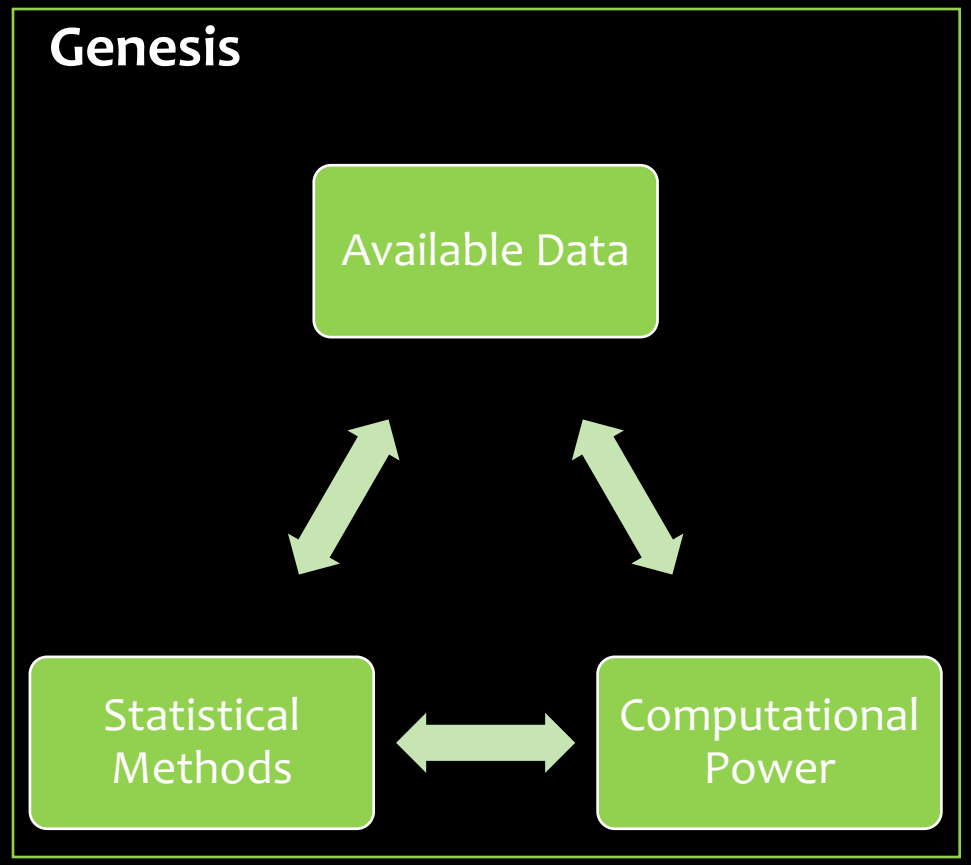
- Under the umbrella of Artificial Intelligence
- Inducing intelligence in the systems in a manner to perform, just like any human would
- Broadly for classification and clustering tasks
  - Fake emails, fraudulent bank transactions, weather prediction, etc.

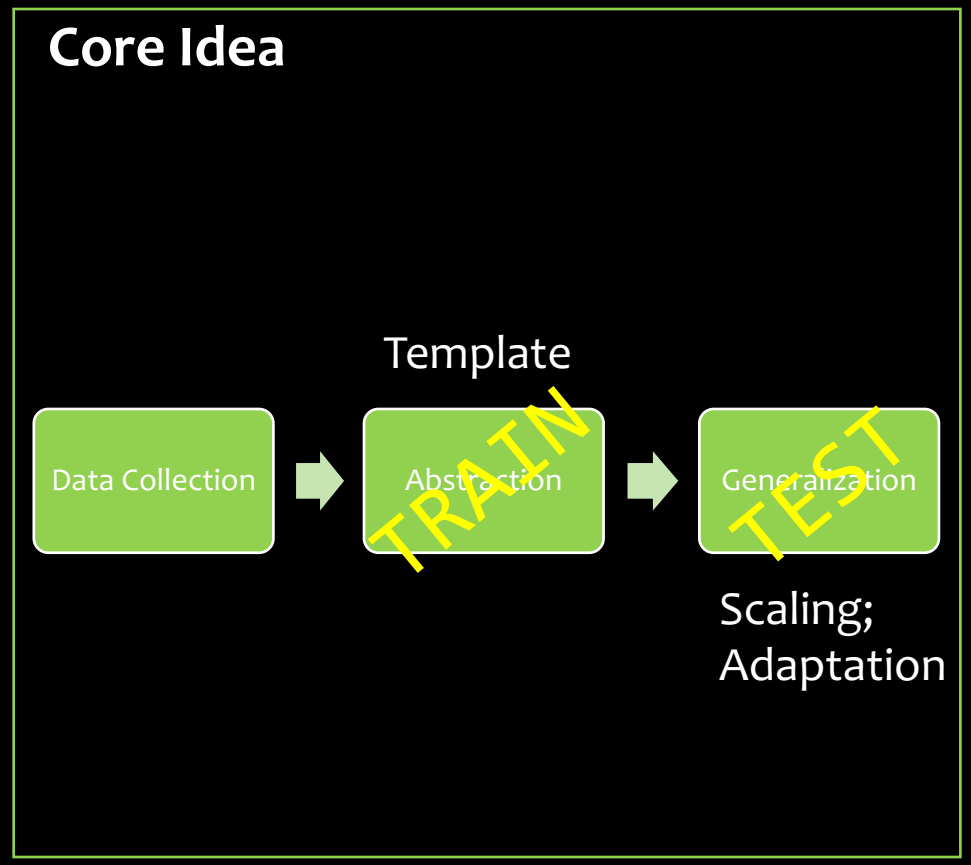
- **Classification** : mapping to Known labels
- **Clustering**: Exploratory analysis; finding patterns on the fly.

Characterizing Machine Learning



Postulates


Genesis
Available Data
Statistical Methods
Computational Power


Core Idea
Template
Data Collection
Abstraction
Generalization
TRAIN
TEST
Scaling;
Adaptation

Why Machine Learning?

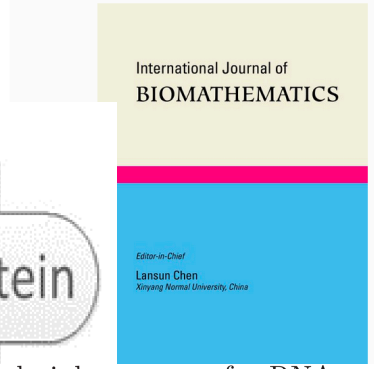
- Relationships
  - Quantification and Mechanization
- Highlighting variables that best describe the relationships within data
  - Normalization
  - Dimensionality reduction
- Aids inference
  - Inference is the basis of statistics
  - Population -> Sample

International Journal of Biomathematics | Vol. 10, No. 01, 1750006 (2017) | Research Articles

A priori, de novo mathematical exploration of gene expression mechanism via regression viewpoint with briefly cataloged modeling antiquity

Shaurya Jauhari and S. A. M. Rizvi

https://doi.org/10.1142/S1793524517500061



Abstract

Various algorithms have been devised to

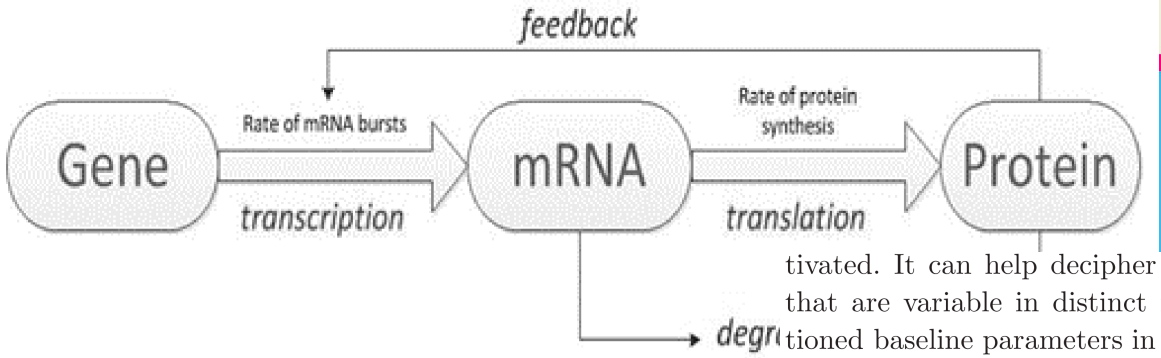


Fig. 3. Kinetics of the central dogma of m

tivated. It can help decipher range of mRNAs and synthesized protein amounts that are variable in distinct metabolic settings. After considering the aforementioned baseline parameters in Sec. 2.3, the following mathematical expressions can be formulated.

$$\text{Gene Expression} = \text{Transcription} + \text{Translation}. \tag{1}$$

We can symbolize the mathematical characterization of the above-terms leveraging Greek symbols. Let T_c denote transcription and T_l represent translation for our mathematical convenience. Also,

$$\text{Gene Expression} = \frac{\text{Treated Gene}_{\text{Expression}}}{\text{Control Gene}_{\text{Expression}}}, \tag{2}$$

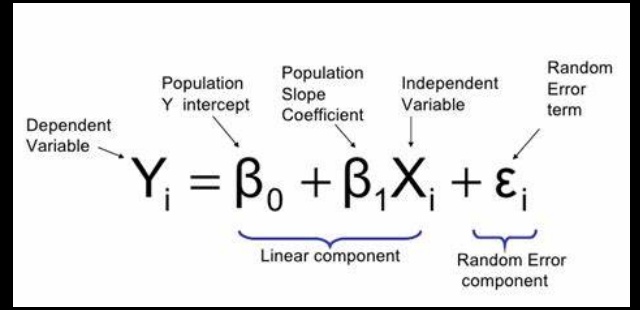
since microarrays represent distinct values that are hybridized.

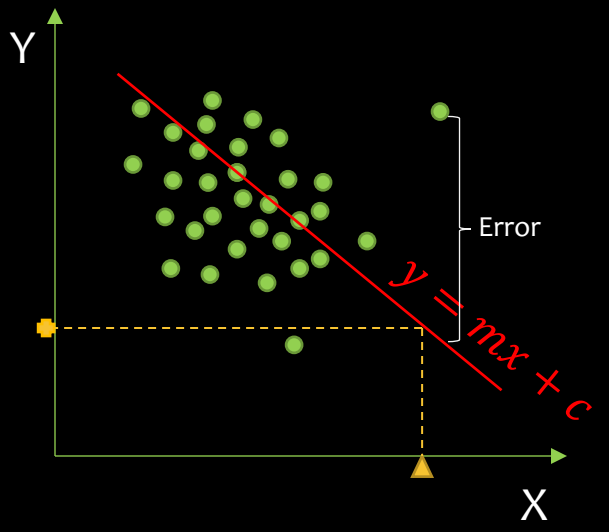
Comparing Eqs. (1) and (2), we have

$$\text{Gene Expression} = \frac{(T_c + T_l)\text{Treated Gene}_{\text{Expression}}}{(T_c + T_l)\text{Control Gene}_{\text{Expression}}}. \tag{3}$$

Linear Regression

- In its simplest form, regression helps ascertain the value for dependent/response variable, given predictor(s) variables.
- Correlation can only quantify the scale of relationship
  - -1 (negative),0 (no correlation),1 (positive correlation)
- Simple Linear Regression (one predictor variable); Multiple Linear Regression (two or more predictor variables)
- No direct Machine Learning linkage.



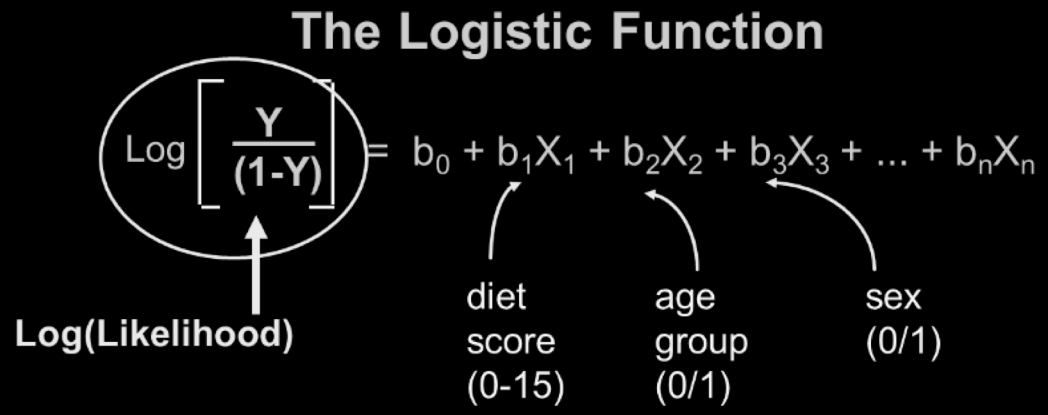


Logistic Regression

- Classification Technique | Categorical, response variable
- Discrete Dependent Variable
  - ***dummy variable*** coded as “yes” or “no”, “0” or “1”, “true” or “false”, etc.
- Logistic regression is a methodology used for ascertaining binary outputs.
  - Rain or not
  - Head or tails
- Binomial Models
  - Two dimensional data (two predictive variables)
- Multinomial Models
  - Multidimensional data (many predictive variables)

Mathematics of Logistic Function

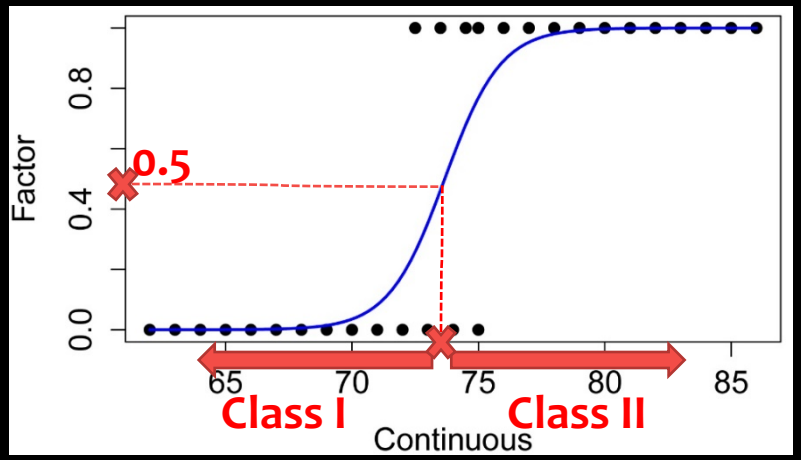
- $\ln\left(\frac{p}{1-p}\right)=\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_n x_n$, where *n* is the number of variables/ features, p is the probability of acceptance/ success and $(1-p)$ is the probability of rejection/ failure.
- Let $\ln\left(\frac{p}{1-p}\right)=y$, then $p=\frac{e^y}{1+e^y}$.



Model

- Suppose the response variable takes two values: 0 and 1.
  - $P(Y = 1 \mid X=x) = P(Y = 0 \mid X=x) + \frac{e^{\beta_0 + \beta^T x}}{1+e^{\beta_0 + \beta^T x}}$
  - Alternatively, $\log \frac{P(Y=1 \mid X=x)}{P(Y=0 \mid X=x)} = \beta_0 + \beta^T x$
  - **Logistic/ Log-odds transformation**

  - **Multinomial Variant:**
    - $P(Y = k \mid X=x) = \frac{e^{\beta_{0k} + \beta_k^T x}}{\sum_{l=1}^{K} e^{\beta_{0l} + \beta_l^T x}}$
    - Different than having continuous response variable



Penalized Logistic Regression: Mitigating high-dimensionality evil

- Also termed *Regularization* of *Bias-Variance Trade-Off Problem*
- *(Multi)Collinearity Problem:*
  - *Overfitting*
  - *Unstable Estimates*
  - *Inaccurate Prediction Model*
- Imposing penalty due to large number of variables
- Results in shrinking coefficients of less contributive variables
- Three methods adopted:
  - **Ridge** (coefficients pushed closer to zero)
  - **Lasso** (coefficients reduced to zero)
  - **Elastic Net** (hybrid approach)

Sum of Squared Errors

- $SSE_{Ridge}=\sum(Y-\hat{Y})^2+\lambda\ \sum\beta^2$
- $SSE_{Lasso}=\sum(Y-\hat{Y})^2+\lambda\ \sum|\beta|$
- $SSE_{Elastic\ Net}=\sum(Y-\hat{Y})^2+\lambda\ [(1-\alpha)\sum\beta^2+\ \alpha\sum|\beta|]$
- , where λ
  - is the degree of penalty to be imposed on the coefficients
  - is derived via *cv.glmnet()* function in the *glmnet* package

- , where α
  - is the mixing parameter (*alpha* in *glmnet()*)
  - α=0 -> Ridge
  - α=1 -> Lasso

stats::glm

glm {stats}

Fitting Generalized Linear Models

Description

`glm` is used to fit generalized linear models, specified by giving a symbolic description of the linear predictor and a description of the error distribution.

Usage

"binomial"

```
glm(formula, family = gaussian, data, weights, subset,
    na.action, start = NULL, etastart, mustart, offset,
    control = list(...), model = TRUE, method = "glm.fit",
    x = FALSE, y = TRUE, singular.ok = TRUE, contrasts = NULL, ...)

glm.fit(x, y, weights = rep.int(1, nobs),
        start = NULL, etastart = NULL, mustart = NULL,
        offset = rep.int(0, nobs), family = gaussian(),
        control = list(), intercept = TRUE, singular.ok = TRUE)

## S3 method for class 'glm'
weights(object, type = c("prior", "working"), ...)
```

glmnet

Glmnet Vignette

Trevor Hastie and Junyang Qian

Stanford June 26, 2014

Introduction

Glmnet is a package that fits a generalized linear model via penalized maximum likelihood. The regularization path is computed for the lasso or elasticnet penalty at a grid of values for the regularization parameter lambda. The algorithm is extremely fast, and can exploit sparsity in the input matrix `x`. It fits linear, logistic and multinomial, poisson, and Cox regression models. A variety of predictions can be made from the fitted models. It can also fit multi-response linear regression.

The authors of glmnet are Jerome Friedman, Trevor Hastie, Rob Tibshirani and Noah Simon, and the R package is maintained by Trevor Hastie. The matlab version of glmnet is maintained by Junyang Qian. This vignette describes the usage of glmnet in R.

`glmnet` solves the following problem

$$\min_{\beta_0,\beta} \frac{1}{N} \sum_{i=1}^{N} w_i l(y_i, \beta_0 + \beta^T x_i) + \lambda \left[(1-\alpha)||\beta||_2^2/2 + \alpha||\beta||_1\right]$$

over a grid of values of λ covering the entire range. Here $l(y,\eta)$ is the negative log-likelihood contribution for observation i; e.g. for the Gaussian case it is $\frac{1}{2}(y-\eta)^2$. The *elastic-net* penalty is controlled by α, and bridges the gap between lasso ($\alpha = 1$, the default) and ridge ($\alpha = 0$). The tuning parameter λ controls the overall strength of the penalty.

It is known that the ridge penalty shrinks the coefficients of correlated predictors towards each other while the lasso tends to pick one of them and discard the others. The elastic-net penalty mixes these two; if predictors are correlated in groups, an $\alpha = 0.5$ tends to select the groups in or out together. This is a higher level parameter, and users might pick a value upfront, else experiment with a few different values. One use of α is for numerical stability; for example, the elastic net with $\alpha = 1 - \epsilon$ for some small $\epsilon > 0$ performs much like the lasso, but removes any degeneracies and wild behavior caused by extreme correlations.

The `glmnet` algorithms use cyclical coordinate descent, which successively optimizes the objective function over each parameter with others fixed, and cycles repeatedly until convergence. The package also makes use of the strong rules for efficient restriction of the active set. Due to highly efficient updates and techniques such as warm starts and active-set convergence, our algorithms can compute the solution path very fast.

The code can handle sparse input-matrix formats, as well as range constraints on coefficients. The core of `glmnet` is a set of fortran subroutines, which make for very fast execution.

The package also includes methods for prediction and plotting, and a function that performs K-fold cross-validation.

General Concepts

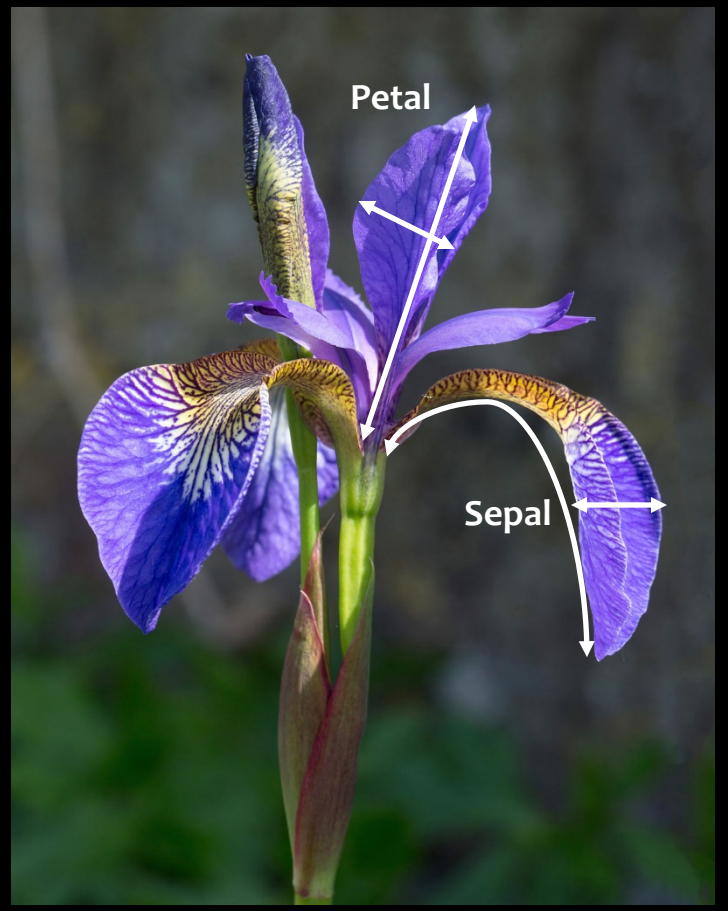
- Mis-classification Error
  - An example/row is mapped to inaccurate class.
  - `misclassfication_error <- 1- sum(diag(confusion_matrix))/sum(confusion_matrix)`

(TN+TP)

- K-fold Cross Validation
  - Divide dataset into k groups (of examples/rows)
  - Use kth group for testing and (k-1) groups for training; repeat process for each group and average out the results.
- Multicollinearity
  - When two or more variables exhibit high correlation
- Bias-Variance Tradeoff
  - Bias is the difference between estimated and actual values
  - Variance is deviation of the estimates
  - Inversely proportional

Dataset Description

- Ronald Fischer, 1936
- Iris family of flowers
  - Setosa
  - Versicolor
  - Virginica
- Numerical characterizations of sepals and petals
- 150 observations
- E. Coli. of datasets!



http://en.wikipedia.org

Thank You

Over to hands-on now!