

# R: Advanced Topics

## Machine Learning and Big Data

Gang Chen  
chengangcs@gmail.com

November 26, 2016

# Outline

- 1 Overview of Machine Learning
- 2 Unsupervised Learning
- 3 Supervised Learning
- 4 Future
- 5 Big Data

# Next

## 1 Overview of Machine Learning

- What is Data Mining?
- Organization of the course
- QuickStarts of R

## 2 Unsupervised Learning

## 3 Supervised Learning

## 4 Future

## 5 Big Data

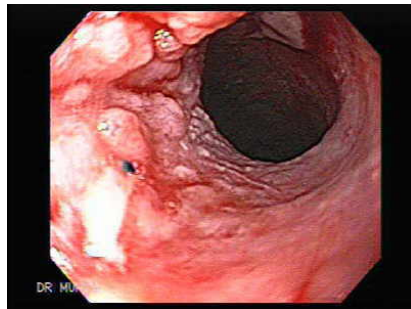
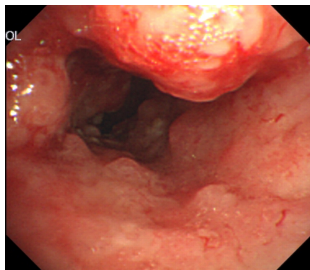
# Biomarker for Tumor Stage

```
## Error in library(curatedOvarianData): there is no  
package called 'curatedOvarianData'  
## Warning in data(TCGA_eset): data set 'TCGA_eset'  
not found  
## Error in boxplot(log(exprs(TCGA_eset)[, 1:2])), col  
= TCGA_eset$tumorstage): "exprs"
```

# Biomarker for Tumor Stage

```
## Error in boxplot(log(exprs(TCGA_eset)[, 1:30]),  
col = TCGA_eset$tumorstage):  "exprs"
```

```
## Error in boxplot(log(exprs(TCGA_eset)[, 1:90]),  
col = TCGA_eset$tumorstage):  "exprs"
```



How to identify tumor from genomics perspective?  
Is it possible to predict tumor stage?

**The answer is Data Mining.**





# What is Data Mining?

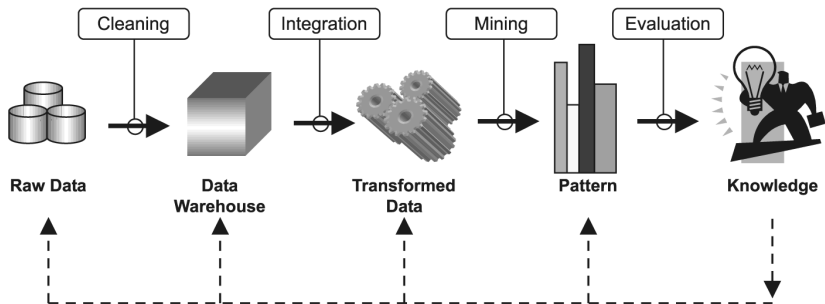
## Pang-Ning Tan, Introduction to Data Mining

Data Mining is the process of automatically discovering useful information in large data repositories.

## Knowledge Discovery in Databases

Data Mining is an integral part of knowledge discovery in databases(KDD).

# Data Mining and KDD



# Machine Learning and Bioinformatics

## Biological Experiments

Microarray  
Sequencing  
Mass Spectrum  
...



## Preprocssing

base calling  
alignment  
variants  
...



## Data Mining

Classification  
Clustering  
Regression  
Feature Engineering  
...



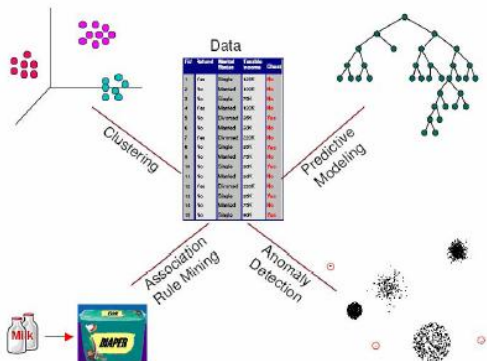
## Biological Knowledge

# Traditional Data Analysis

## Motivations

- Scalability
- High Dimensionality
- Heterogeneous and Complex Data
- Data Ownership and Distribution
- Non-traditional analysis

# Data Mining Tasks



# Data Mining and Machine Learning

## Machine Learning

Machine learning provides the technical basis of data mining.

---Data Mining: Practical Machine Learning Tools and Techniques

# Schedule

## Schedule

- Introduction
- Unsupervised Learning: Clustering
- Supervised Learning: Classification
- Discussion

# Softwares

## Softwares

- R: R is an free platform for data analysis and visuaztion.
- R packages:
  - e1071 SVM
  - curatedOvarianData Microarry data of tumor. [Database 2013]
  - mtcars A small dataset for concept demonstration.
- Rgui
- Emacs + ESS
- Vim + R-Plugin
- RStudio
- RStudio in the Cloud



# R

- Download from [www.r-project.org](http://www.r-project.org).
- Installation and start;
- Install e1071 package.

# Next

- 1 Overview of Machine Learning
- 2 Unsupervised Learning**
  - Unsupervised Learning in Bioinformatics
  - Hierarchical Clustering and its Applications in Bioinformatics
  - Summary
- 3 Supervised Learning
- 4 Future
- 5 Big Data

# Heat Map

```
## Error in library(colorRamps): there is no package  
called 'colorRamps'  
## Error in eval(expr, envir, enclos):  "exprs"  
## Error in eval(expr, envir, enclos):  'exprData'  
## Error in heatmap(subdata, col = green2red(4),  
scale = "none"):      'subdata'
```

# Unsupervised Learning: Clustering

## What is clustering?

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters).

---Wikipedia

# Applications

## Applications of Clustering

- Clustering for Understanding
  - Biology
  - Information Retrieval
  - Climate
  - Psychology and medicine
  - business
  - .....
- Clustering for Utility
  - Summarization
  - compression
  - Efficiently finding nearest neighbors
  - .....

# Common Clustering Methods

## Clustering Methods

- Density-based clustering
- K-means
- Hierarchical Clustering
- Semi-supervised clustering
- .....

# Unsupervised Learning in Bioinformatics

- Discovery of tumor subtypes by clustering gene expression, CNV, miRNA or integrated data.
- Clonal evolution analysis of tumor
- Mutation spectrum clustering
- Pathway or functional annotation based clustering
- Graph clustering for identification of protein functional module or protein complex
- Clustering metagenomic sequences
- Metabolomics
- .....

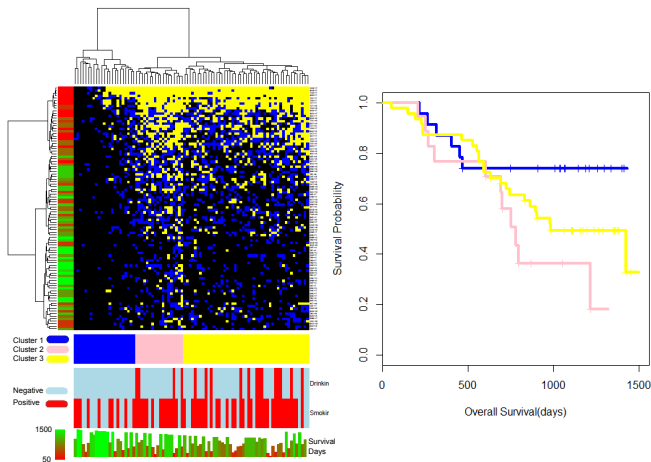
# Hierarchical Clustering

## Steps

- Calculating distance between individuals
- Combine closest individuals (optional, recalculate distance)
- Visualization and annotation



# Clustering in Bioinformatics



# Hierarchical Clustering in R

```
help(dist)
help(hclust)
help(heatmap)
```

# Distance Calculation

```
dist(x, method = "euclidean",  
diag = FALSE,  
upper = FALSE,  
p = 2)
```

# Distance Calculation

```
exprDist = dist(t(subdata))  
  
## Error in t(subdata):    'subdata'  
  
exprDist  
  
## Error in eval(expr, envir, enclos):    'exprDist'
```

# Clustering

```
hclust(d, method = "complete", members = NULL)
```

# Clustering

```
exprDist = dist(t(subdata))  
  
## Error in t(subdata):      'subdata'  
  
exprClust = hclust(exprDist)  
  
## Error in hclust(exprDist):  'exprDist'  
  
exprClust  
  
## Error in eval(expr, envir, enclos):  'exprClust'
```

# Visualization

## Visualization

- Dendrogram: `plot.hclust`
- Heat Map: `heatmap`

# Dendrogram

```
plot(exprClust, cex=0.6, main="Dendrogram")  
  
## Error in plot(exprClust, cex = 0.6, main =  
"Dendrogram"):      'exprClust'
```



# plot.hclust

```
plot(x, labels = NULL,  
     hang = 0.1,  
     axes = TRUE,  
     frame.plot = FALSE,  
     ann = TRUE,  
     main = "Cluster Dendrogram",  
     sub = NULL,  
     xlab = NULL, ylab = "Height", ...)
```

# Heat Map

```
heatmap(subdata)
```

```
## Error in heatmap(subdata):      'subdata'
```

# heatmap

```
heatmap(x, Rowv = NULL, Colv = if(symm)"Rowv" else NULL,  
distfun = dist, hclustfun = hclust,  
reorderfun = function(d, w) reorder(d, w),  
add.expr, symm = FALSE, revC = identical(Colv, "Rowv"),  
scale = c("row", "column", "none"), na.rm = TRUE,  
margins = c(5, 5), ColSideColors, RowSideColors,  
cexRow = 0.2 + 1/log10(nr), cexCol = 0.2 + 1/log10(nc),  
labRow = NULL, labCol = NULL, main = NULL,  
xlab = NULL, ylab = NULL,  
keep.dendro = FALSE, verbose = getOption("verbose"), ...)
```

# How to read Heat Map

```
## Error in heatmap(subdata, col = green2red(4),  
scale = "none"):      'subdata'
```

# Summary

## Clustering

- Clustering is widely used in bioinformatics
- Clustering can be implemented by using built-in functions of R
- Clustering can be visualized as heat map and dendrogram

# Next

- 1 Overview of Machine Learning
- 2 Unsupervised Learning
- 3 **Supervised Learning**
  - Supervised Learning in Bioinformatics
  - Support Vector Machine and its Applications in Bioinformatics
  - Summary
- 4 Future
- 5 Big Data

# Supervised Learning: Classification

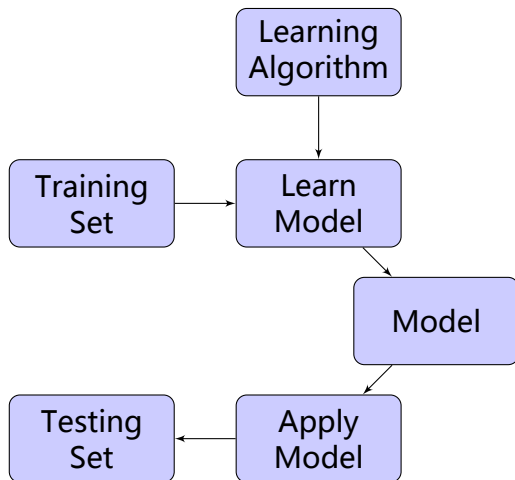
## Classification

Assigning objects to one of several predefined categories.

## Definition

Classification is the task of learning a target function  $f$  that maps each attribute set  $x$  to one of the predefined class labels  $y$ .

# How to solve a classification problem?



General approach for building a classification model



# Evaluation

Confusion Matrix for a 2-class problem

	Prediction=1	Prediction=0
Class=1	$f_{11}$	$f_{10}$
Class=0	$f_{01}$	$f_{00}$

# Evaluation

## Accuracy and Error Rate

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

$$= \frac{f_{11} + f_{00}}{f_{11} + f_{10} + f_{00} + f_{01}}$$

$$\text{ErrorRate} = \frac{\text{Number of wrong predictions}}{\text{Total number of predictions}}$$

$$= \frac{f_{10} + f_{01}}{f_{11} + f_{10} + f_{00} + f_{01}}$$

# Classification in Bioinformatics

## Applications

- Classification of diseases, especially cancer.
- Prediction of clinical outcome.
- Prediction of the function of gene or proteins.
- Prediction of the structure of proteins.
- .....

# Objective

## Two Objectives

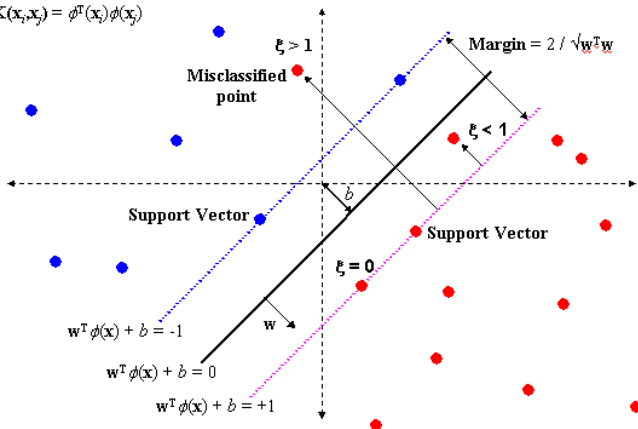
- ① To build accurate classifiers or predictors
- ② To derive inferences from the results obtained

## Challenges

- data inconsistency and missing values
- noise
- normalization
- Dimensionality reduction

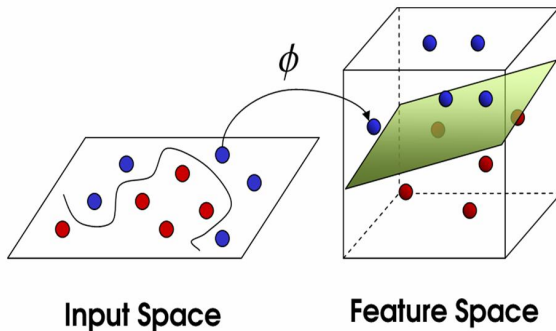
## SVM

$$K(x_i, x_j) = \phi^T(x_i) \phi(x_j)$$



# Kernal

## Principle of Support Vector Machines (SVM)



# SVM in R

```
## Error in eval(expr, envir, enclos):      'TCGA_eset'  
## Error in stages[stages < 3] <- 0:      'stages'  
## Error in stages[stages >= 3] <- 1:      'stages'  
## Error in is.factor(x):                  'stages'
```

```
library(e1071)
```

# SVM

```
library(e1071)
```

```
## Error in library(e1071): there is no package  
called 'e1071'
```

```
model = svm(x = exprData[1:400,], y = stages[1:400], cross=5)
```

```
## Error in eval(expr, envir, enclos):  "svm"
```



# SVM model

```
model
```

```
## Error in eval(expr, envir, enclos):  'model'
```

# SVM

```
ret = predict(model, exprData[401:500,])  
  
## Error in predict(model, exprData[401:500, ]):  
  'model'  
  
table(ret, stages[401:500])  
  
## Error in table(ret, stages[401:500]):      'ret'
```

?

# Summary

## Classification

- Classification is an important technique for bioinformatics
- SVM is powerful
- Classification algorithm can be implemented in R easily

# Next

- 1 Overview of Machine Learning
- 2 Unsupervised Learning
- 3 Supervised Learning
- 4 Future**
- 5 Big Data

# Future Reading

- Data Mining for Bioinformatics
- Introduction to Machine Learning
- CRAN Task View: Machine Learning & Statistical Learning: <http://cran.r-project.org/web/views/MachineLearning.html>

# Next

- 1 Overview of Machine Learning
- 2 Unsupervised Learning
- 3 Supervised Learning
- 4 Future
- 5 **Big Data**

# High Performance Computing

- Task View: High-Performance and Parallel Computing with R

<http://cran.r-project.org/web/views/HighPerformanceComputing.html>

- compiler package and JIT
- Revolution R Enterprise
- Rcpp
- Multi-core
- GPU
- MPI

# compiler package

These functions provide an interface to a byte code compiler for R.

```
cmpfun(f, options = NULL) compile(e, env = .GlobalEnv,  
options = NULL) cmpfile(infile, outfile, ascii = FALSE,  
env = .GlobalEnv, verbose = FALSE, options = NULL)  
enableJIT(level)
```

see compiler.R for example



# Big Data Framework

- Hadoop: RHIPe
- Spark: SparkR, AMPLab UC BERKELEY  
<http://amplab-extras.github.io/SparkR-pkg/>
- Storm:
  - <https://github.com/allenday/R-Storm>
  - <https://github.com/quintona/storm-r>