生物统计学：
生物信息中的概率统计模型

2024年秋




Eisen MB,
Spellman PT,
Brown PO,
Botstein D.

有关信息

- 授课教师：宁康
  - Email: ningkang@hust.edu.cn
  - Office: 华中科技大学东十一楼606室
  - Phone: 18627968927

- 课程网页
  - http://www.microbioinformatics.org/Biostatistics.html
  - QQ群：717914581

课程安排

- 生物背景和课程简介
- 传统生物统计学及其应用
- 生物统计学和生物大数据挖掘
  - Hidden Markov Model (HMM)及其应用
    - Markov Chain
    - HMM理论
    - HMM和基因识别 (Topic I)
    - HMM和序列比对 (Topic II)
  - 进化树的概率模型 (Topic III)
  - Motif finding中的概率模型 (Topic IV)
    - EM algorithm
    - Markov Chain Monte Carlo (MCMC)
  - 基因表达数据分析 (Topic V)
    - 聚类分析-Mixture model
    - Classification-Lasso Based variable selection
  - 基因网络推断 (Topic VI)
    - Bayesian网络
    - Gaussian Graphical Model
  - 基因网络分析 (Topic VII)
    - Network clustering
    - Network Motif
    - Markov random field (MRF)
  - Dimension reduction及其应用 (Topic VIII)
- 面向生物大数据挖掘的深度学习

研究对象:
生物序列,
进化树,
生物网络,
基因表达
...

方法:
生物计算与生物统计

机器学习的分类

机器学习总共分为三类：

- **监督学习（supervised learning）**主要是从训练数据中学习输入x到输出y的映射关系，也称之为**预测**（predictive）。x我们一般称之为属性。当输出y是连续量是，此时问题称之为**回归（regression）**，当输出y是离散量时，此时问题称之为**分类（classification）**。当我们的标记空间是有一些自然的顺序的，比如成绩A-F，此时归为传统的分类问题就不合适，因为没有利用A>B>C>D>E>F的特性，该类问题称为有序回归（ordinal regression）或有序分类（ordinal classification），这是一个介于回归和分类中间的一个问题。

- **无监督学习（unsupervised learning）**主要是从训练数据中发现一些有趣的模式。有时也叫做**知识发现**。在这里要注意的是，无监督学习的数据是没有标签的。比较常见典型的无监督学习问题是**聚类（clustering）**。即给你一些离散的点，然后通过学习对这些点进行分类。

- **强化学习（reinforcement learning）**是对行为进行奖赏或者惩罚，通过自我学习，争取获得更多的奖赏而不是惩罚。

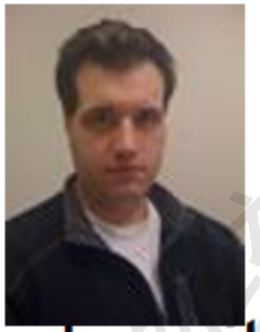
第6-1章: Clustering Analysis

1. Hierarchical clustering
2. Model-based clustering

References:

- M. Eisen et al.: Cluster analysis and display of genome-wide expression patterns. Proc.Natl.Acad.Sci.USA 95, 14863-8, 1998
- Wei Pan, Jizhen Lin and Chap T Le. Model-based cluster analysis of microarray gene-expression data. Genome Biology 3(2): research0009.1–0009.8, 2002.
- 2002. G.J. McLachlan, R.W. Bean, and D. Peel, A Mixture Model-Based Approach to the Clustering of Microarray Expression Data. Bioinformatics 18, 413-422,

- Eisen MB, Spellman PT, Brown PO, Botstein D. 1998. Cluster analysis and display of genome-wide expression patterns. *PNAS* 95: 14863-14868.

- Google scholar citation: 13,06 (04/25/2013), 17,191(12/26/2017) , 19,413(12/08/2020)

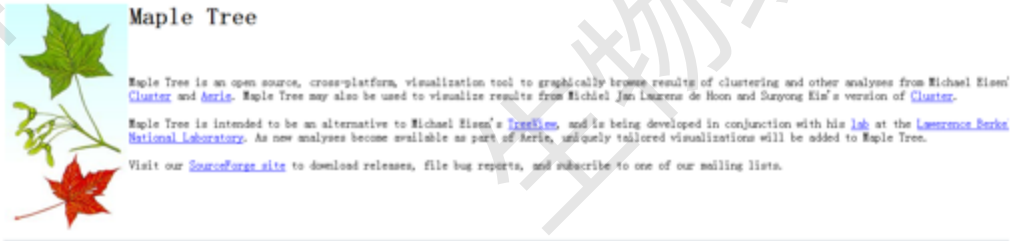
Cluster Analysis and Visualization Software

- Cluster 3.0

  http://bonsai.hgc.jp/~mdehoon/software/cluster/software.htm

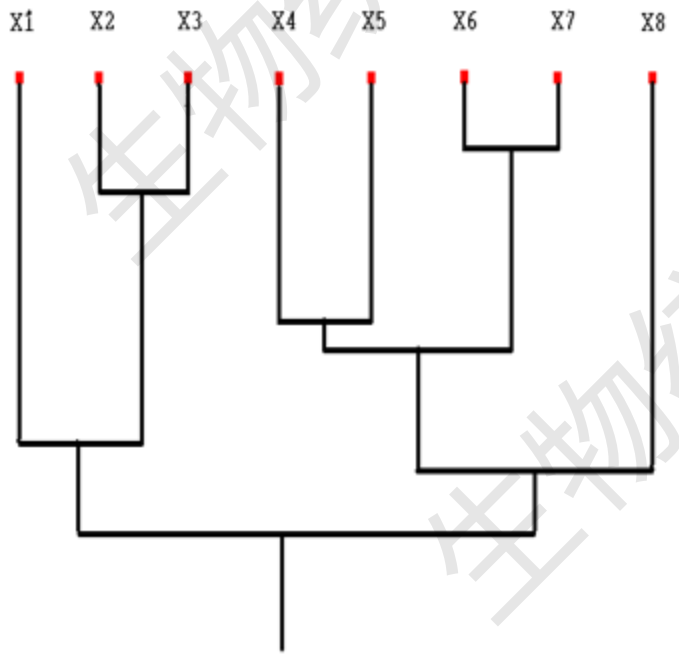
- TreeView

  http://www.eisenlab.org/eisen/?page_id=42

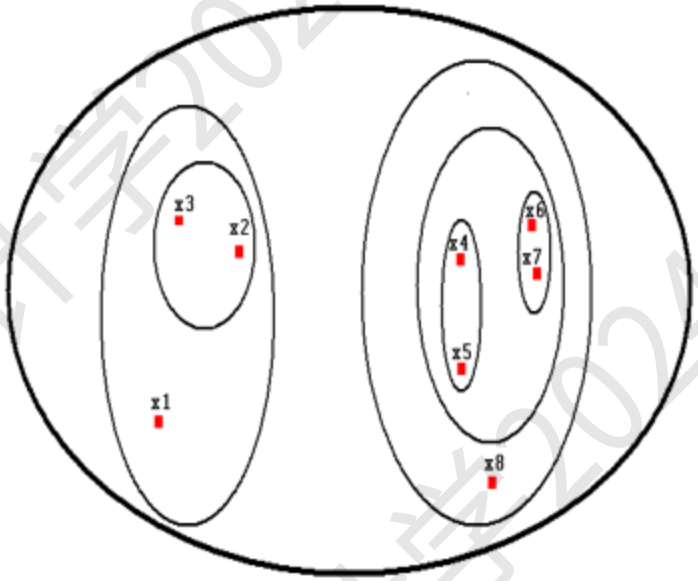


http://mapletree.sourceforge.net
/

Hierarchical Clustering



Dendrogram



Venn Diagram of Clustered Data

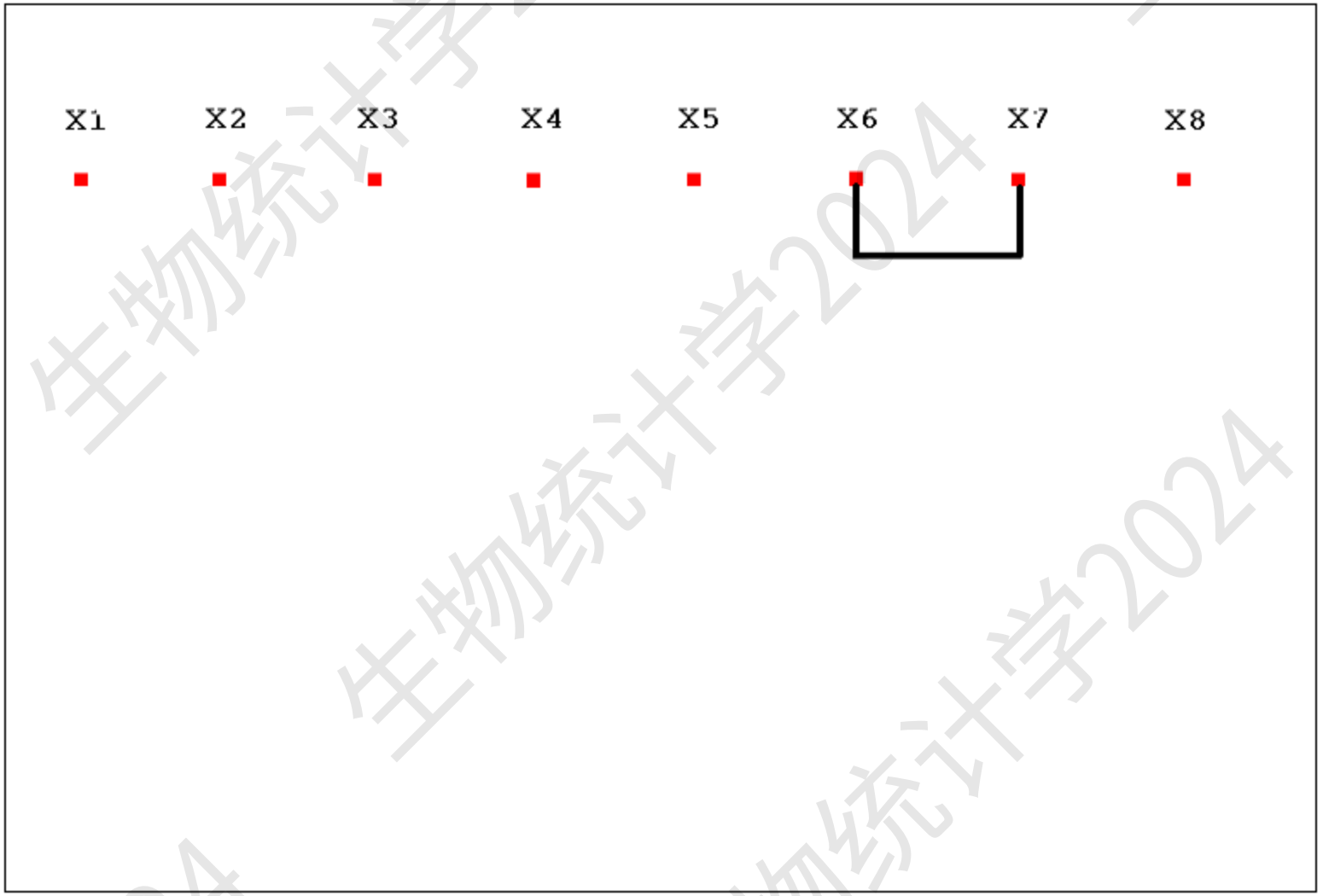
From http://www.stat.unc.edu/postscript/papers/marron/Stat321FDA/RimaIzempresentation.ppt

Nearest Neighbor Algorithm

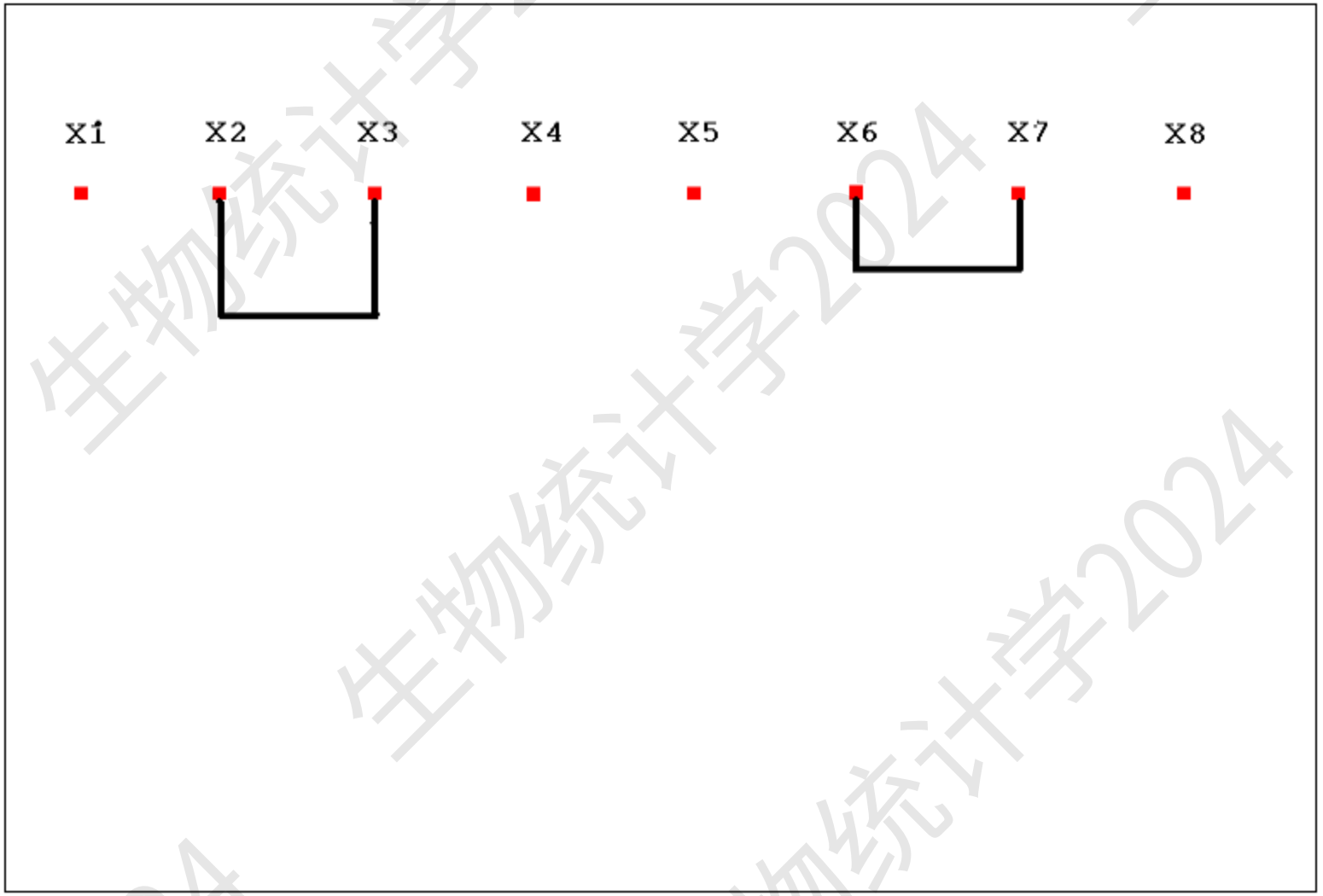
- Nearest Neighbor Algorithm is an agglomerative approach (bottom-up).

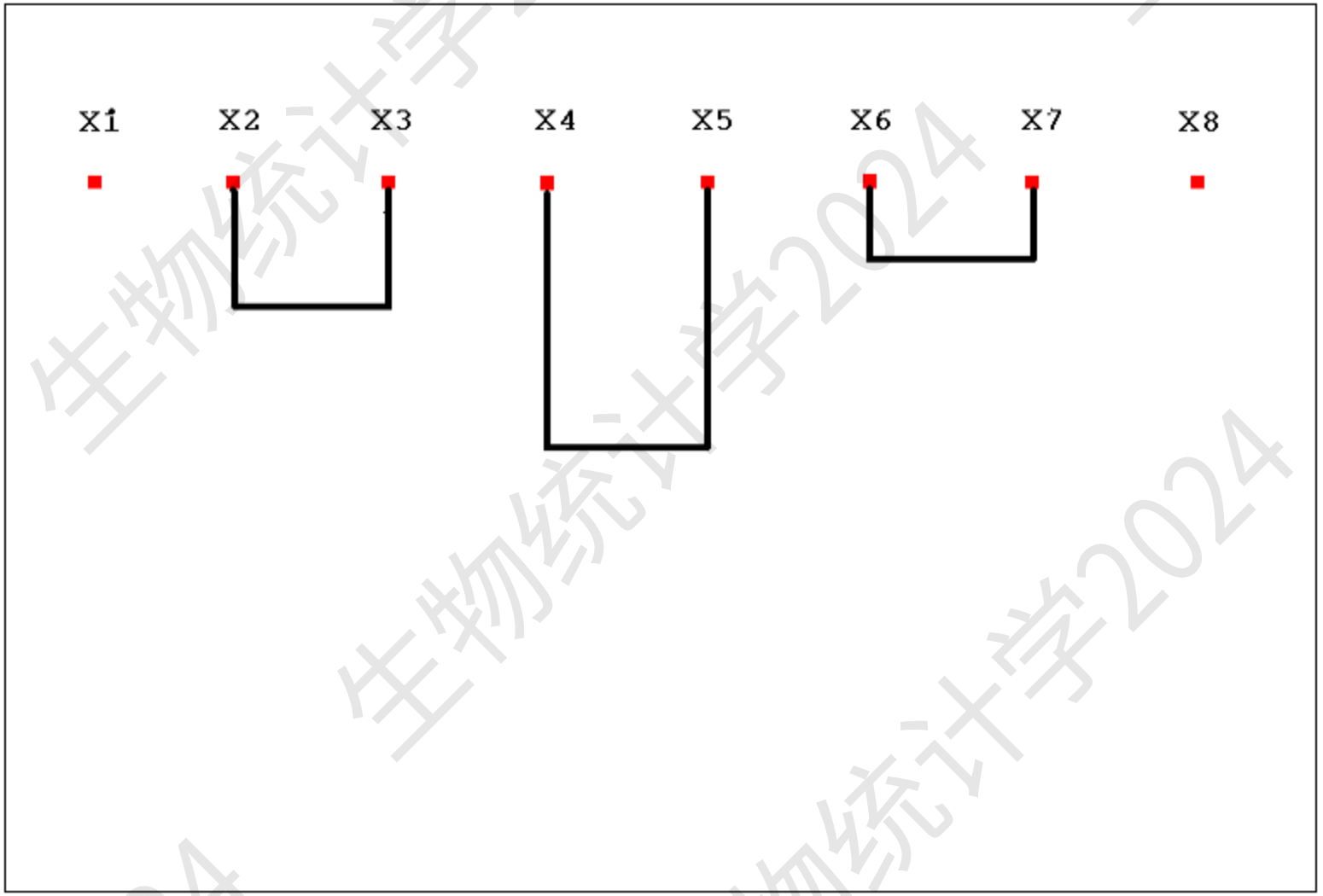
- Starts with *n* nodes (*n* is the size of our sample), merges the 2 most similar nodes at each step, and stops when the desired number of clusters is reached.

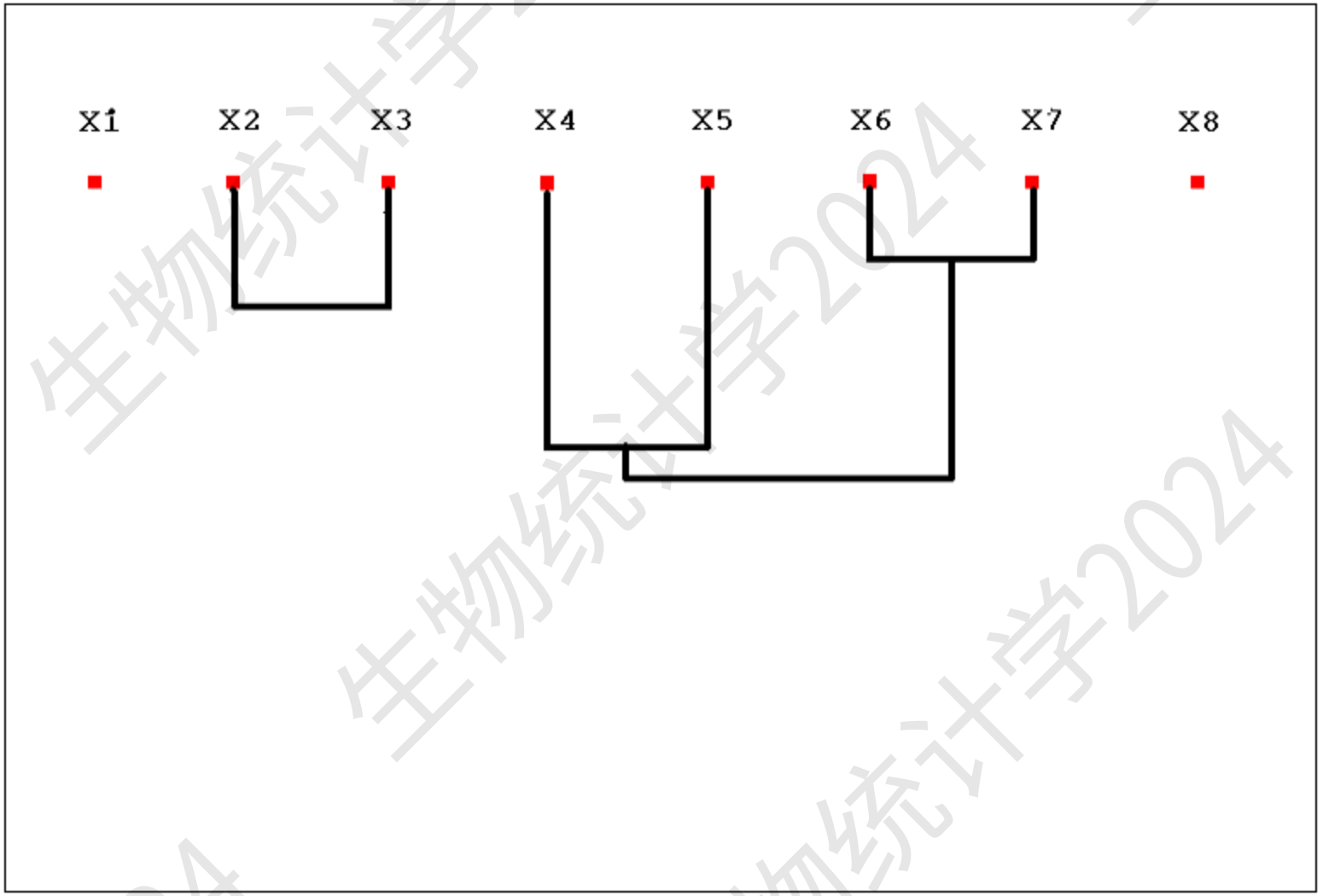
From http://www.stat.unc.edu/postscript/papers/marron/Stat321FDA/RimaIzempresentation.ppt

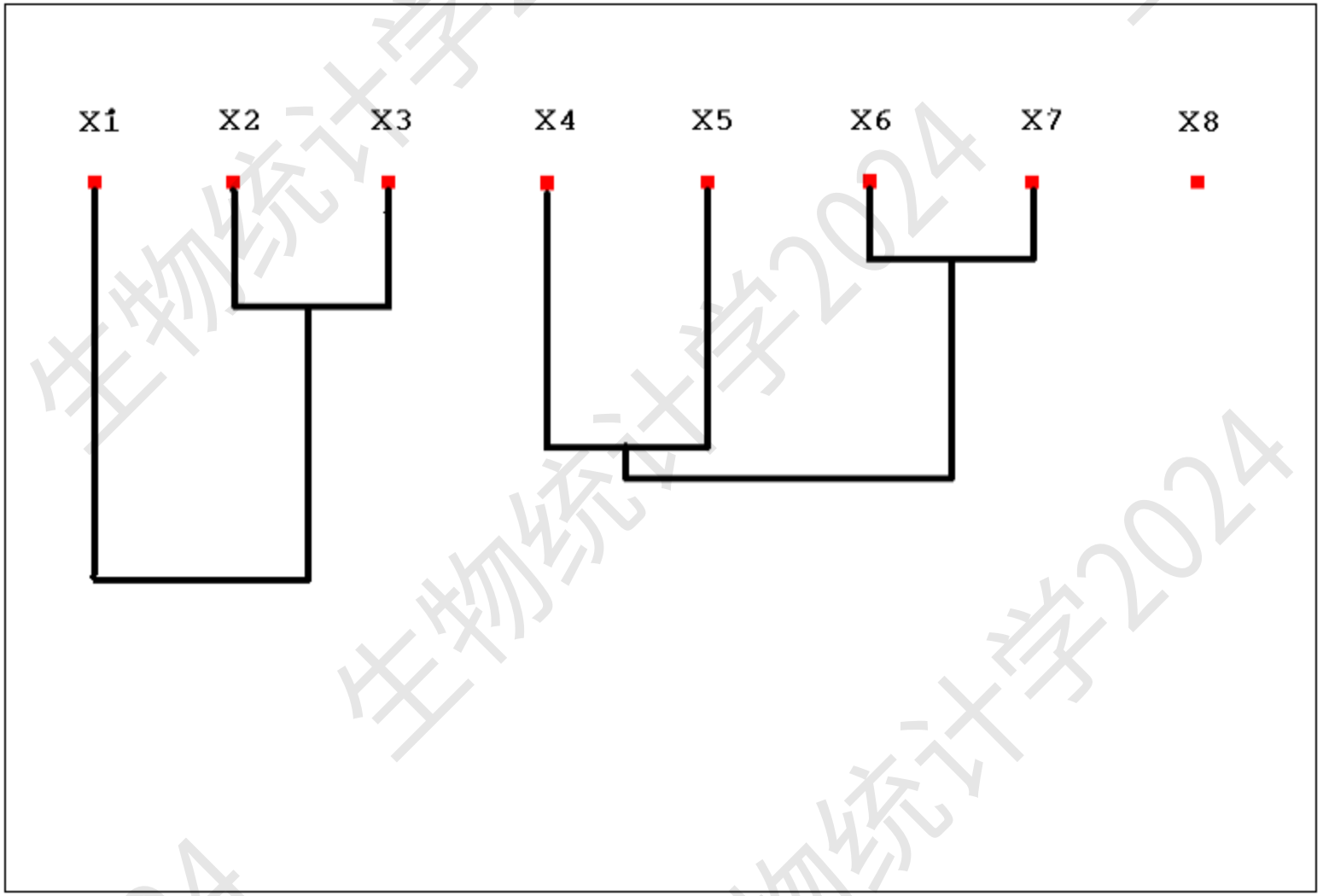


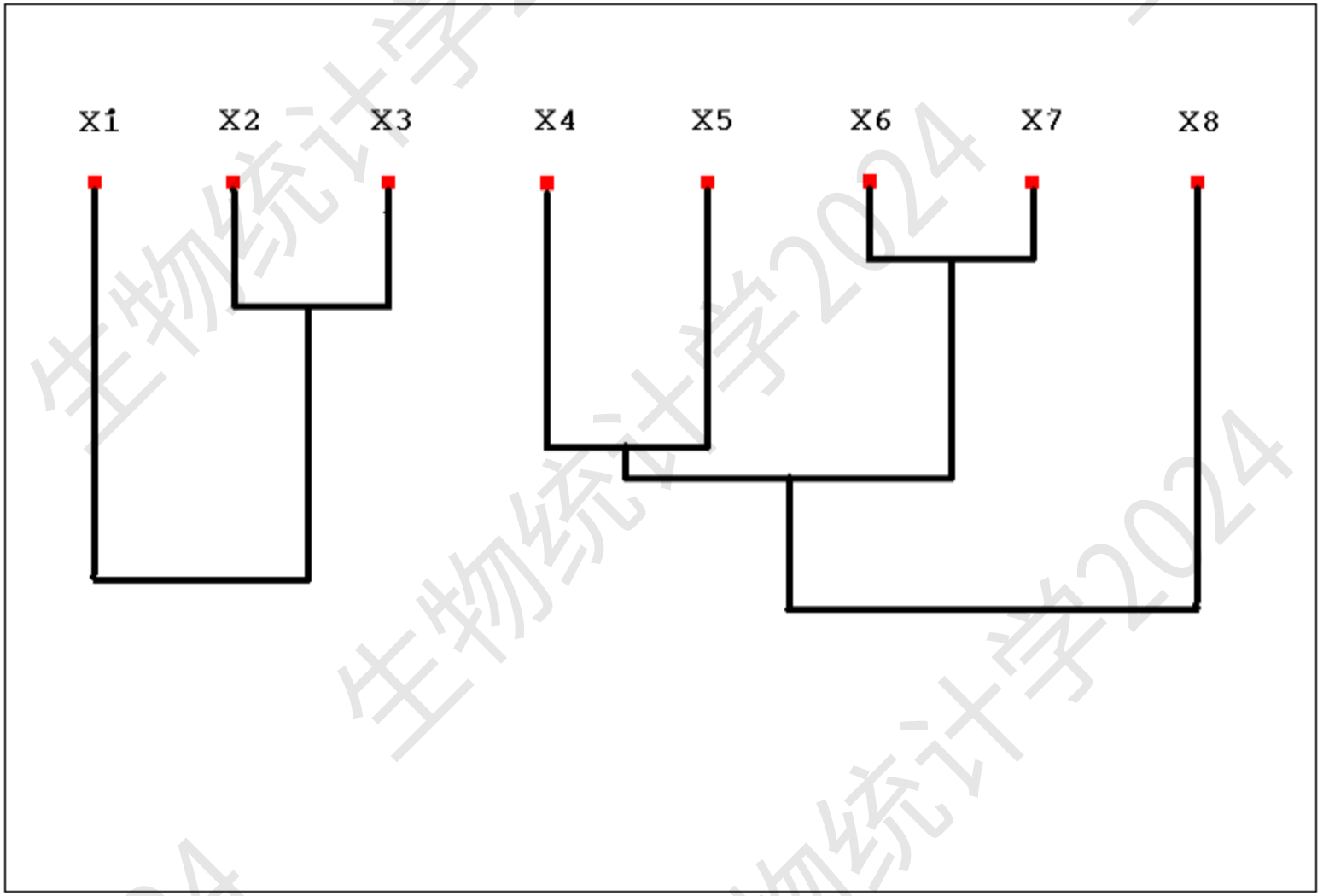
From http://www.stat.unc.edu/postscript/papers/marron/Stat321FDA/RimaIzempresentation.ppt

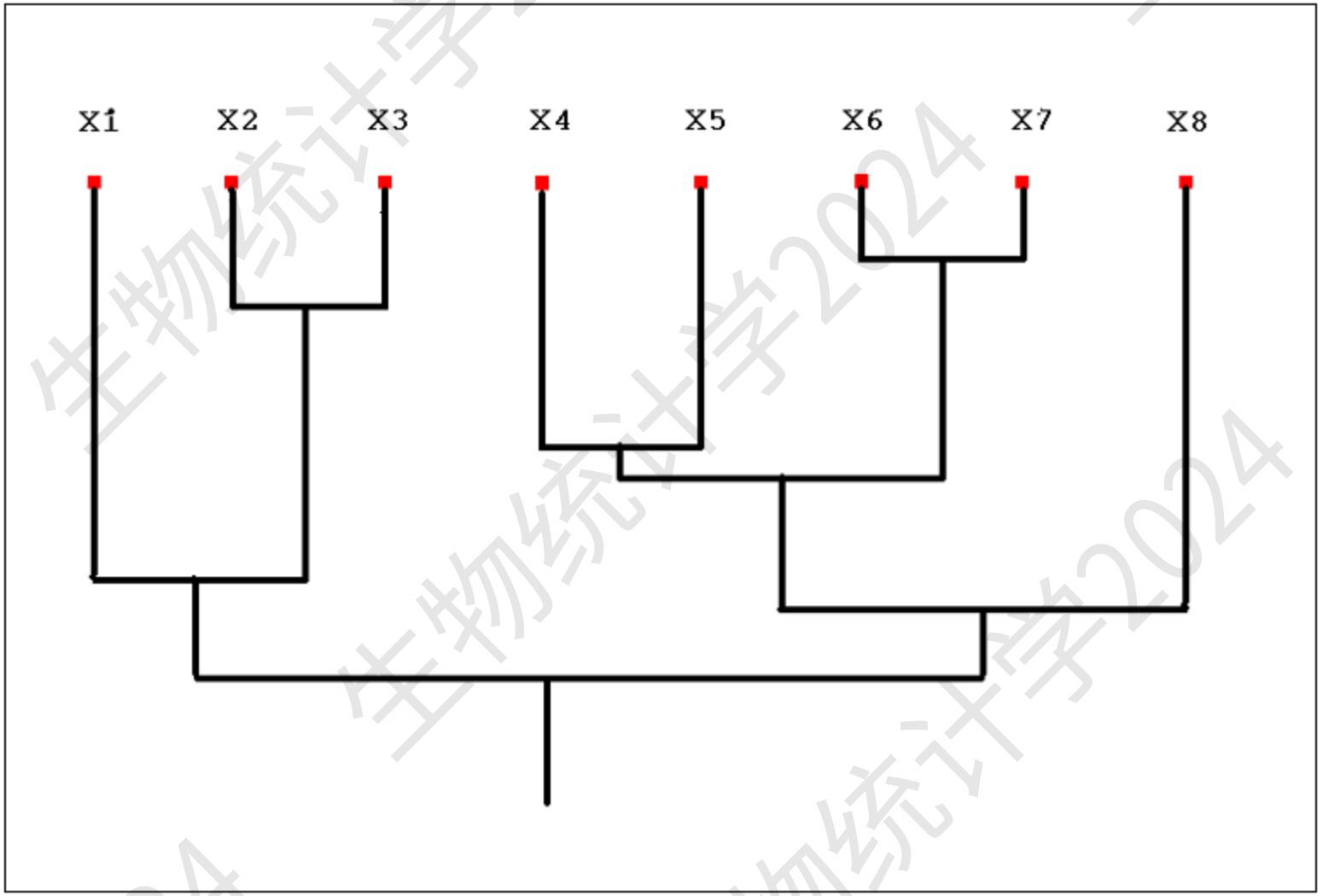









X1
X2
X3
X4
X5
X6
X7
X8


X1
X2
X3
X4
X5
X6
X7
X8

Similarity Measurements

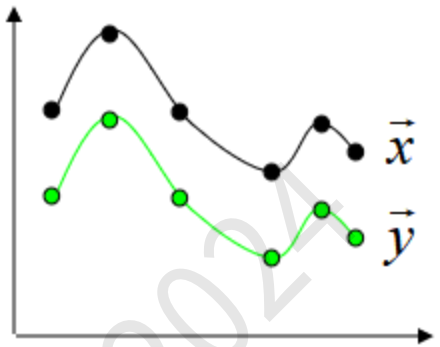
- Pearson Correlation

Two profiles (vectors) $\vec{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix}$ and $\vec{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}$

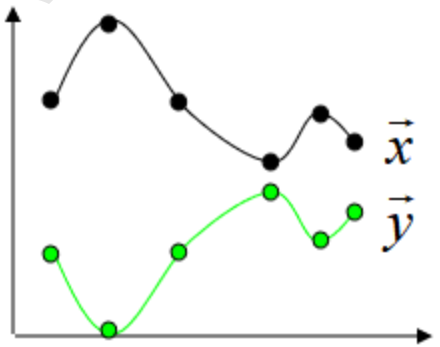
$$C_{pearson}(\vec{x},\vec{y}) = \frac{\sum_{i=1}^{N}(x_i - m_x)(y_i - m_y)}{\sqrt{[\sum_{i=1}^{N}(x_i - m_x)^2][\sum_{i=1}^{N}(y_i - m_y)^2]}}$$

$$m_x = \frac{1}{N}\sum_{n=1}^{N} x_n$$

$$m_y = \frac{1}{N}\sum_{n=1}^{N} y_n$$



+1 ≥ Pearson Correlation ≥ – 1



Similarity Measurements

- Euclidean Distance

$$\vec{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix} \quad \vec{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}$$

$$d(\vec{x}, \vec{y}) = \sqrt{\sum_{n=1}^{N} (x_n - y_n)^2}$$

Similarity Measurements

- Cosine Correlation

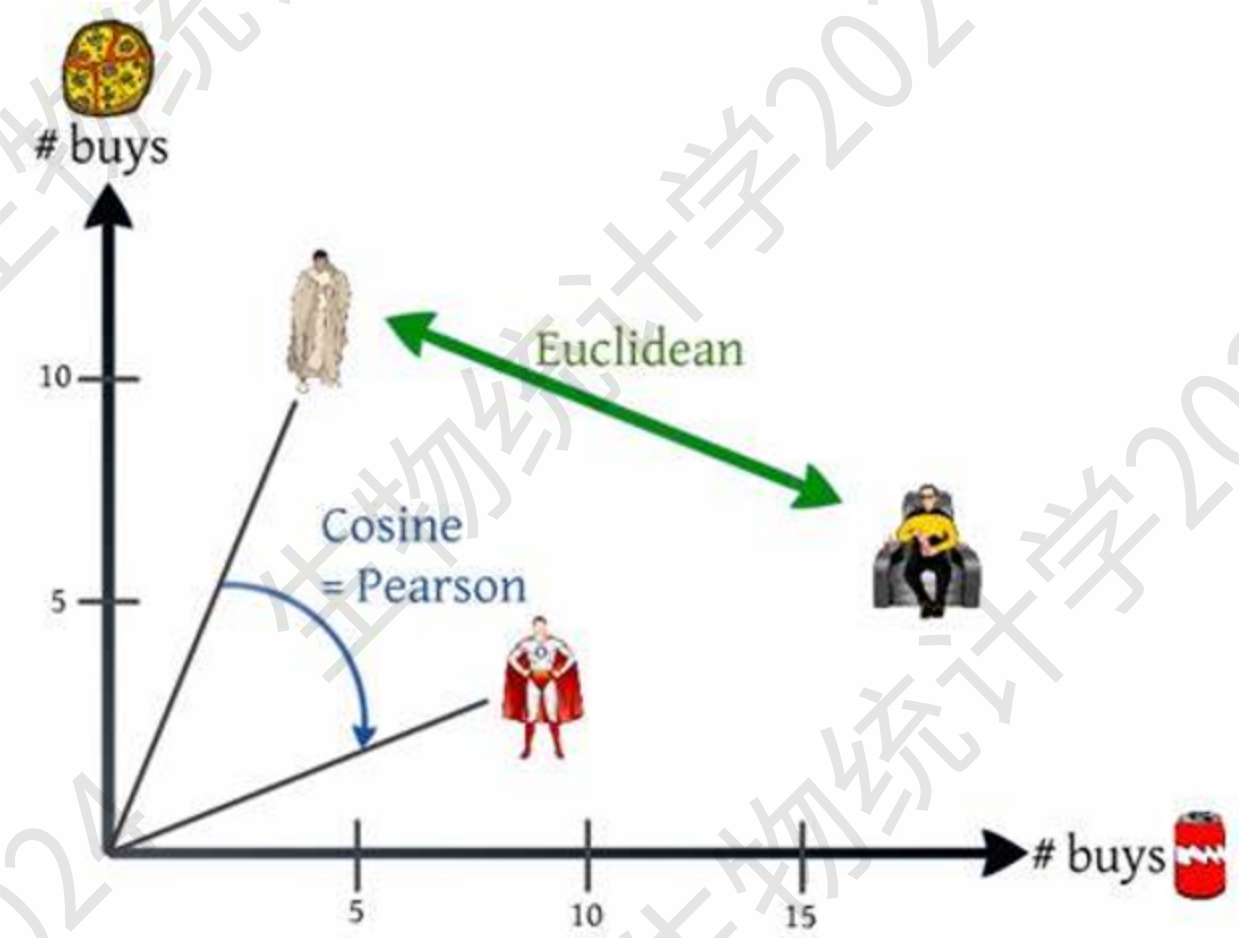
$$\vec{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix} \qquad \vec{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}$$

$$C_{\text{cosine}}(\vec{x}, \vec{y}) = \frac{\sum_{i=1}^{N} x_i \times y_i}{\|\vec{x}\| \times \|\vec{y}\|}$$

$\vec{x} = \vec{y}$ +1 ≥ Cosine Correlation ≥ – 1 $\vec{x} = -\vec{y}$

Similarity Measurements

- Cosine Correlation



Similarity Measurements

- Cosine Correlation

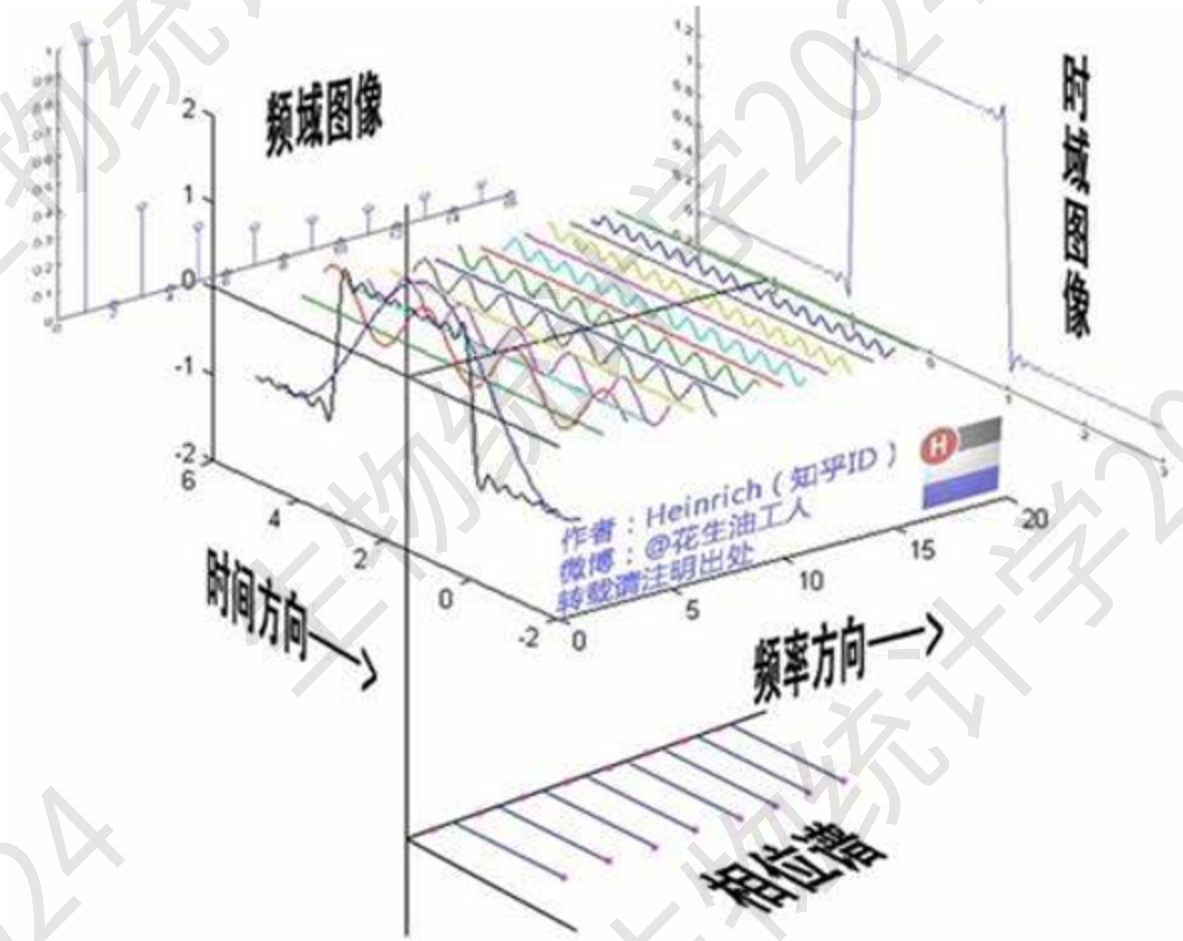
词语	新闻1	新闻2	新闻3
股市	15	30	0
指数	10	20	0
投资	5	10	1
涨幅	6	12	0
比赛	0	0	16
篮球	0	0	30
NBA	0	0	12

$$\cos(\theta) = \frac{\sum_{i=1}^{n}(X_i \times Y_i)}{\sqrt{\sum_{i=1}^{n}(X_i)^2} \times \sqrt{\sum_{i=1}^{n}(Y_i)^2}}$$

文件	余弦相似度
finance1	1
finance2	0.339534
finance4	0.231394
finance5	0.220267
finance3	0.191725
finance7	0.148047
finance6	0.141999
finance10	0.123854
tech5	0.10717
tech6	0.078086
finance9	0.077941
…	…

Similarity Measurements

- Fourier Transformation

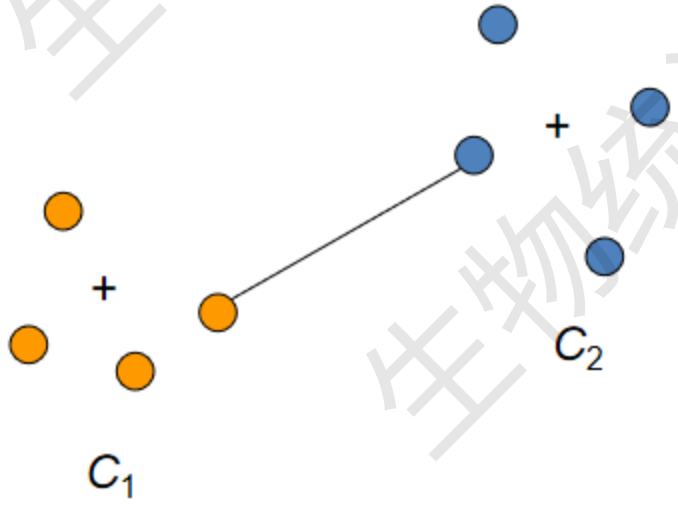


Group Similarity

- Single linkage
- Complete linkage
- Average linkage
- Average group linkage

Clustering (聚类)

Single Linkage

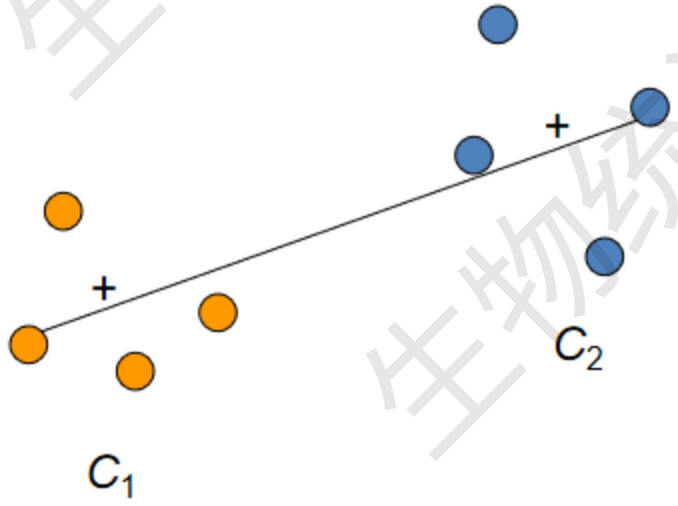


Dissimilarity between two clusters = Minimum dissimilarity between the members of two clusters

Tend to generate “long chains”

Clustering

Complete Linkage

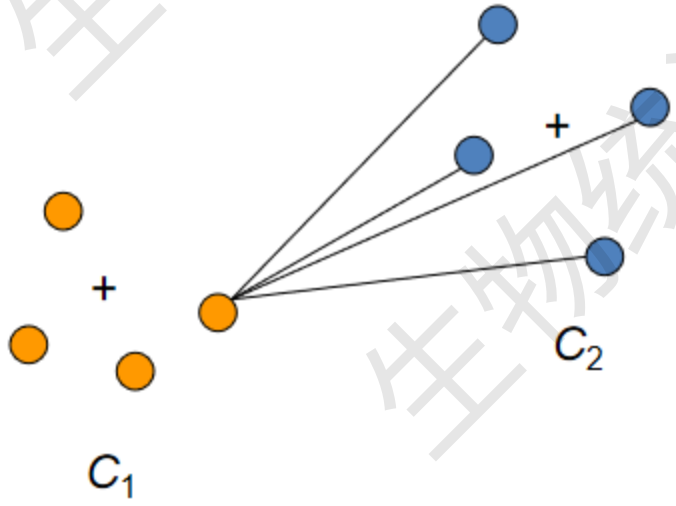


Dissimilarity between two clusters = Maximum dissimilarity between the members of two clusters

Tend to generate “clumps”

Clustering

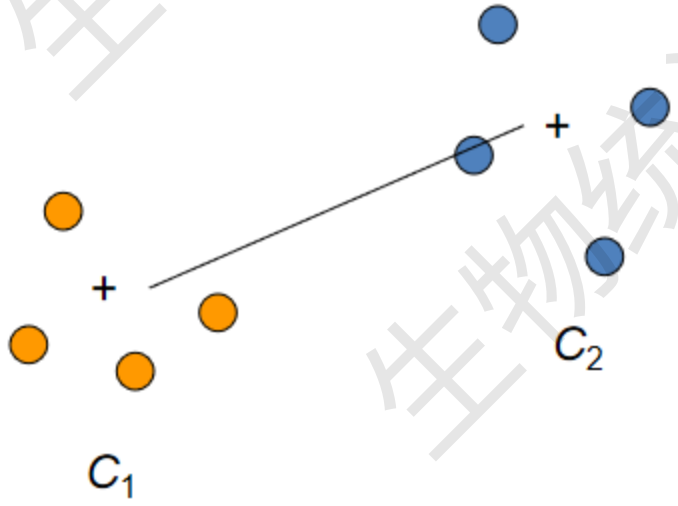
Average Linkage



Dissimilarity between two clusters = Averaged distances of all pairs of objects (one from each cluster).

Clustering

Average Group Linkage



Dissimilarity between two clusters =
Distance between two cluster means.

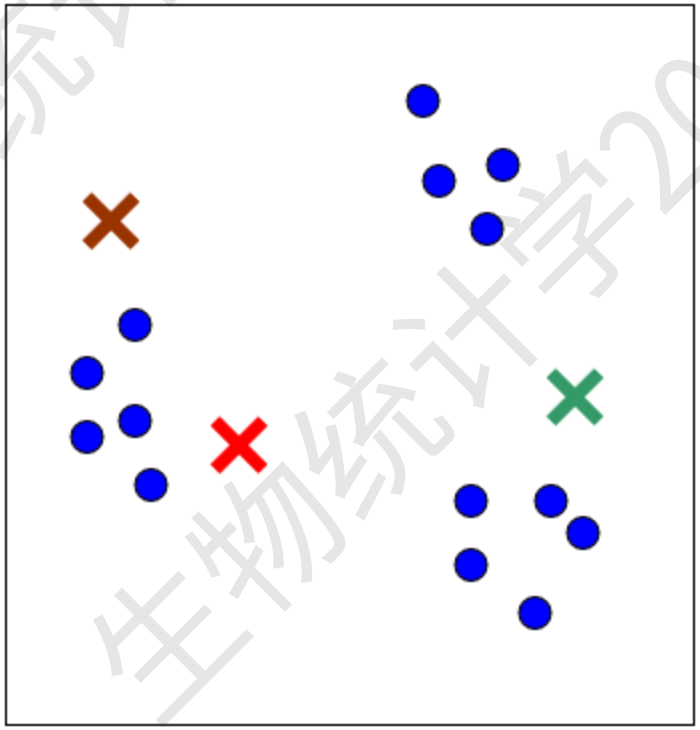
Other Clustering Methods

- K-means, fuzzy k-means
- Self-organization mapping (SOM)
- Gaussian mixture model, Bayesian clustering algorithms
- Nonnegative Matrix factorization
- Iterative signature algorithm (ISA), progressive iterative signature algorithm (PISA)…
- Biclustering

K-means Algorithm

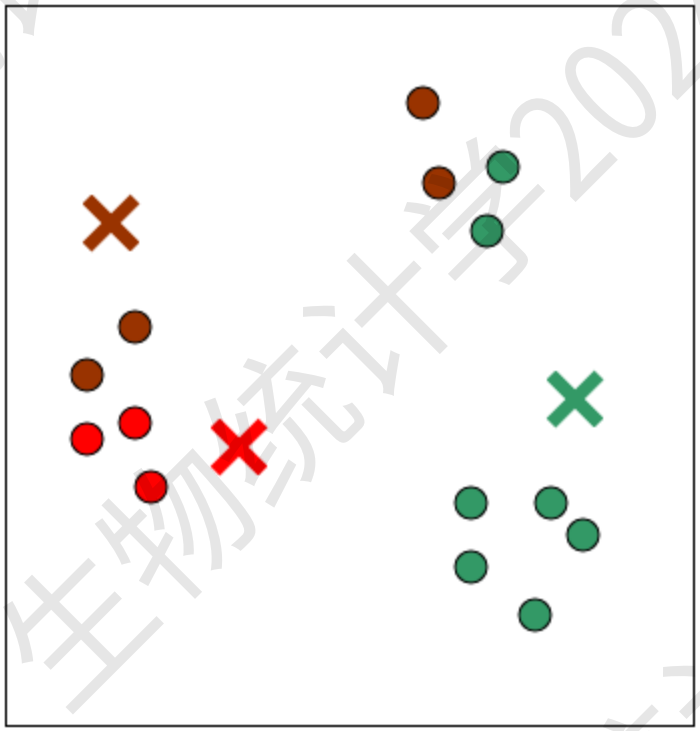
1. Choose K centroids at random
2. Make initial partition of objects into k clusters by assigning objects to closest centroid
3. Calculate the centroid (mean) of each of the k clusters.
4. a. For object i, calculate its distance to each of the centroids.
   b. Allocate object i to cluster with closest centroid.
   c. If object was reallocated, recalculate centroids based on new clusters.
4. Repeat 3 for object i = 1,….N.
5. Repeat 3 and 4 until no reallocations occur.
6. Assess cluster structure for fit and stability

K-means Algorithm



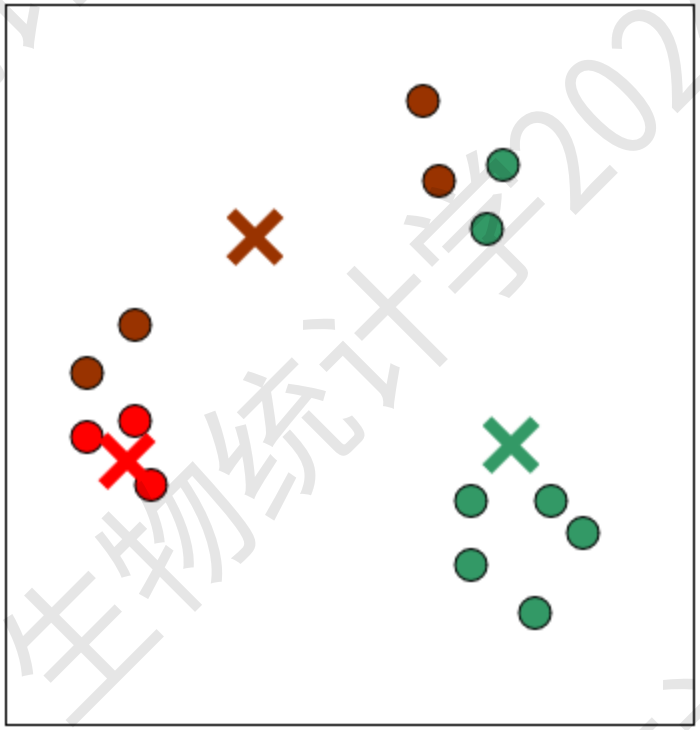
Iteration = 0

K-means Algorithm



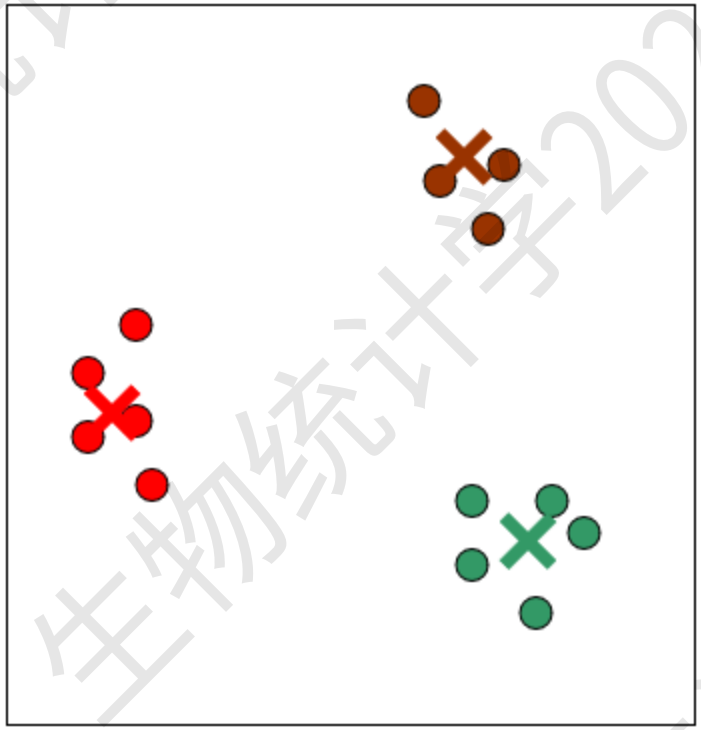
Iteration = 1

K-means Algorithm



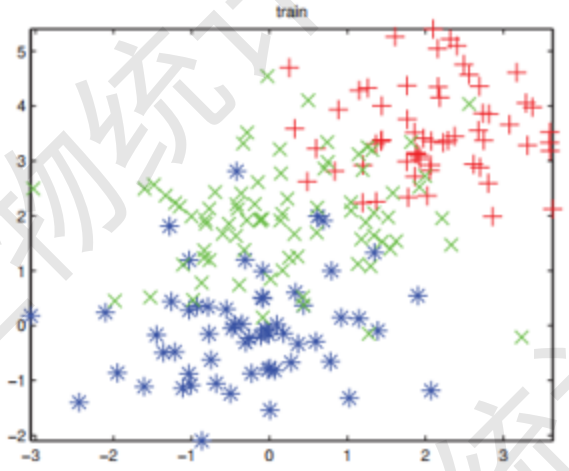
Iteration = 2

K-means Algorithm

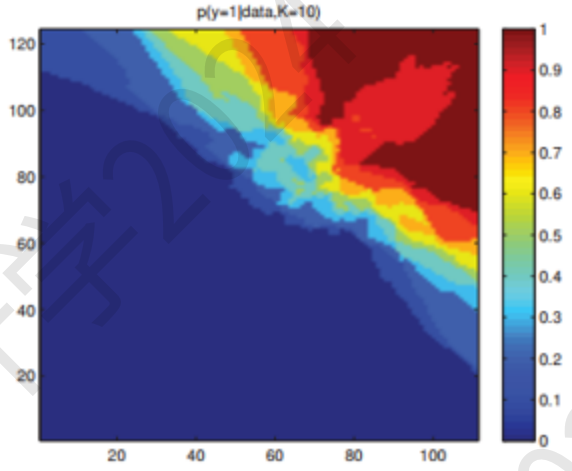


Iteration = 3

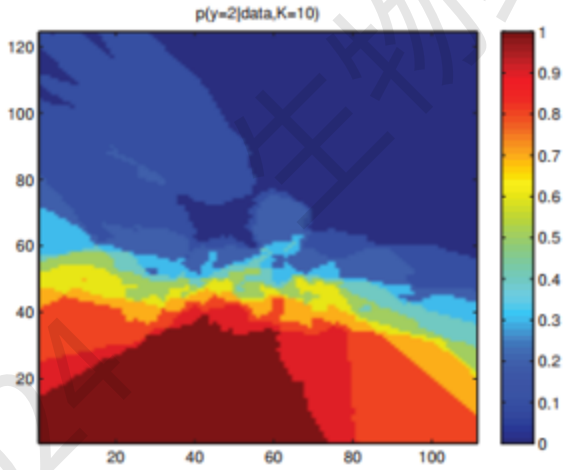
K-means Algorithm



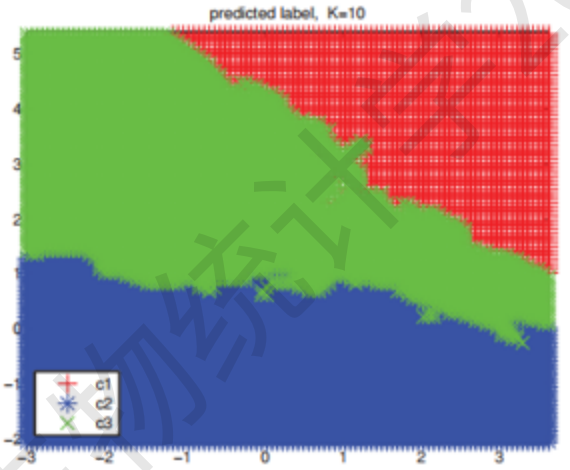
(a)



(b)

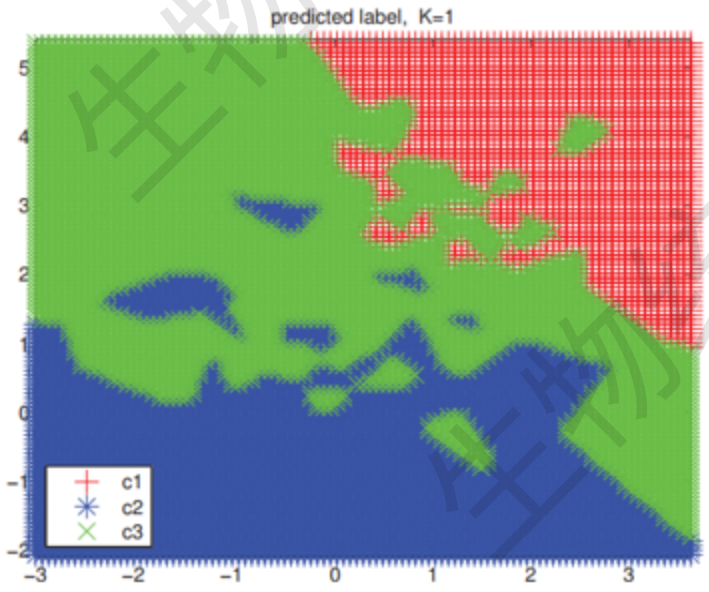


(c)

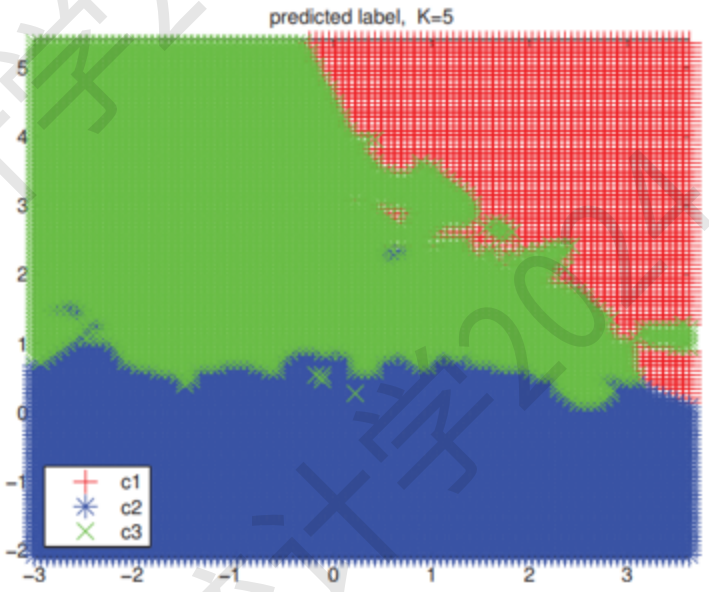


(d)

K-means Algorithm



(a)



(b)

Gaussian Mixture Model

- Each class corresponding to a normal distribution
- The data point y is taken to be a realization from a Gaussian mixture model

$$f(y;\Theta) = \sum_{i=1}^{g} \pi_i \phi(y;\mu_i, V_i)$$

Learning the Parameters

- Maximum likelihood estimation. Given data points $y_1, \ldots, y_n$,

$$l(\Theta) = \sum_{i=1}^{n} \log f(y_i; \Theta)$$

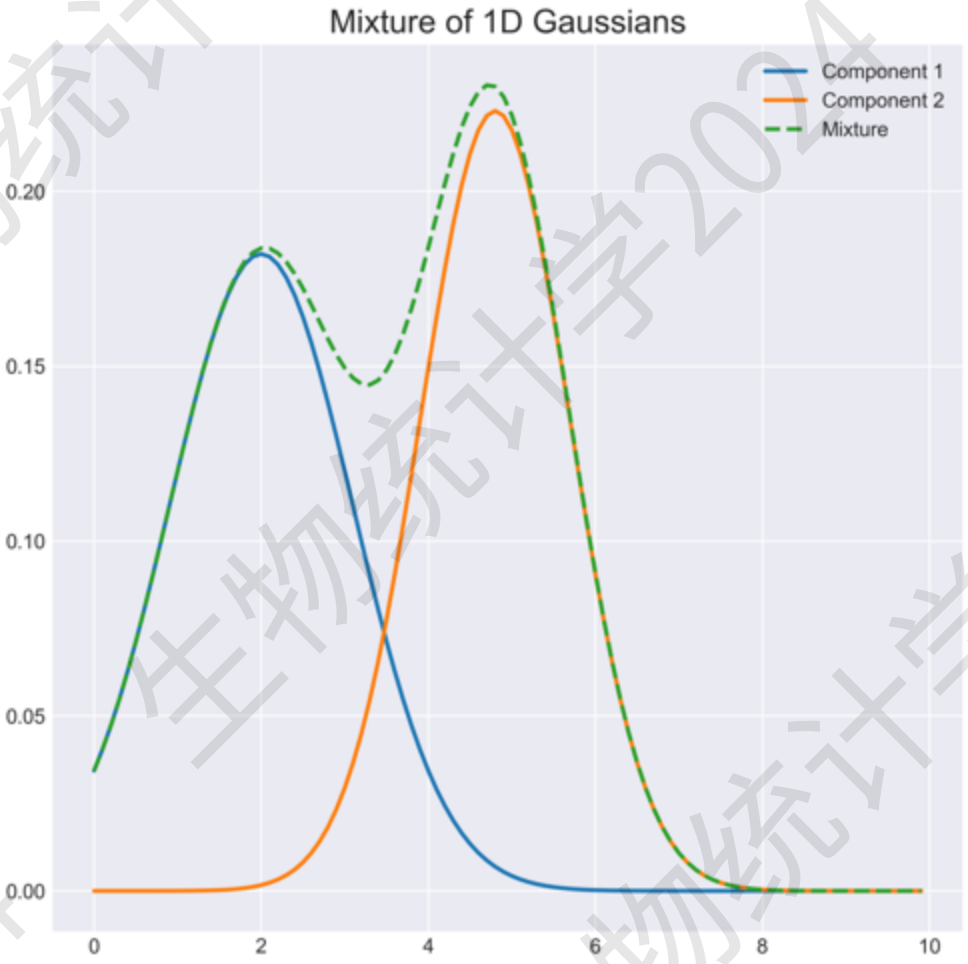
- Missing data problem, the class label of each data point.

EM Algorithm

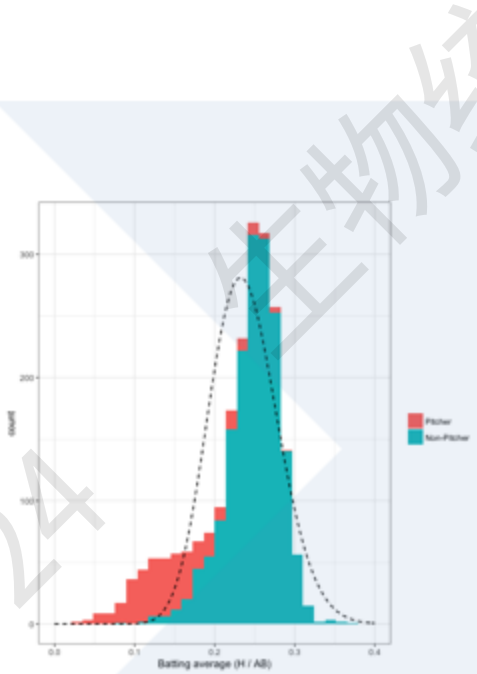
- Iteratively update

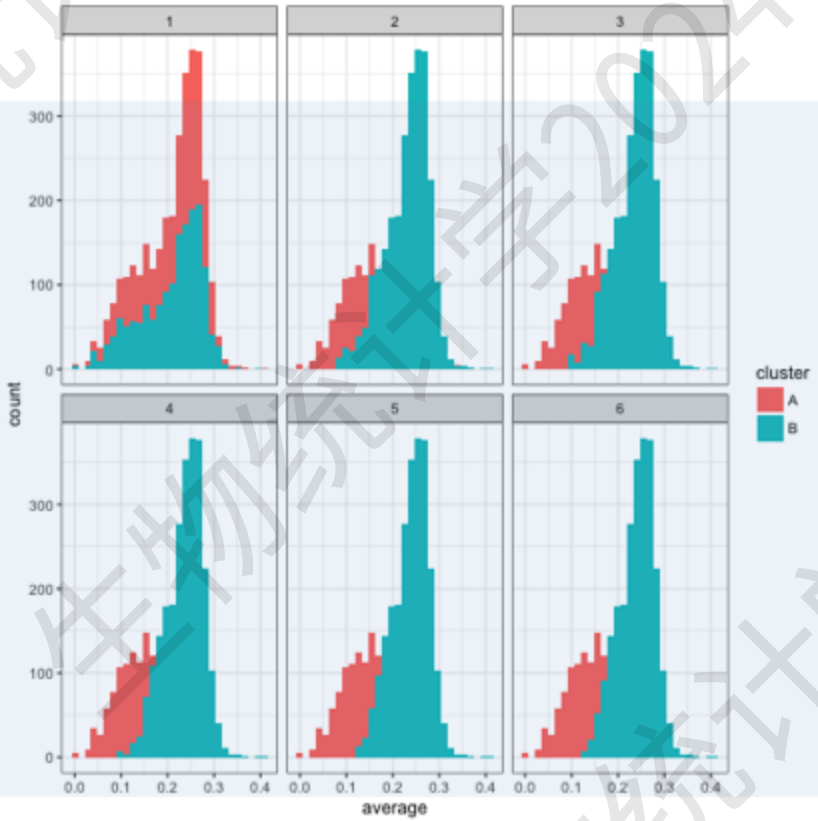
$$
\begin{cases}
\tau_{ij}^{(k)} = \dfrac{\pi_i^{(k)} \phi(y_j; \mu_i^{(k)}, V_i^{(k)})}{f(y_j; \Theta_k)} \\
\pi_i^{(k+1)} = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} \tau_{ij}^{(k)} \\
\mu_i^{(k+1)} = \dfrac{\sum_{i=1}^{n} \tau_{ij}^{(k)} y_j}{\sum_{i=1}^{n} \tau_{ij}^{(k)}} \\
V_i^{(k+1)} = \dfrac{\sum_{i=1}^{n} \tau_{ij}^{(k)} (y_j - \mu_i^{(k+1)})(y_j - \mu_i^{(k+1)})^T}{\sum_{i=1}^{n} \tau_{ij}^{(k)}}
\end{cases}
$$

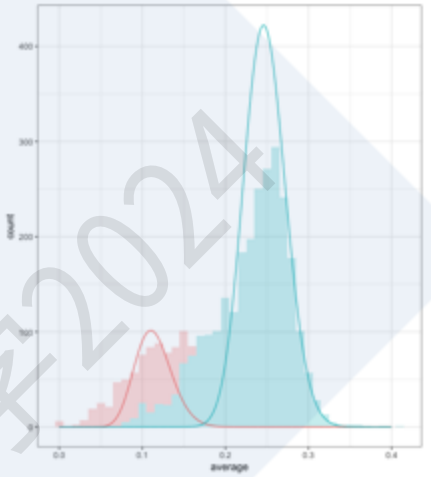
Gaussian Mixture Model



GMM-EM

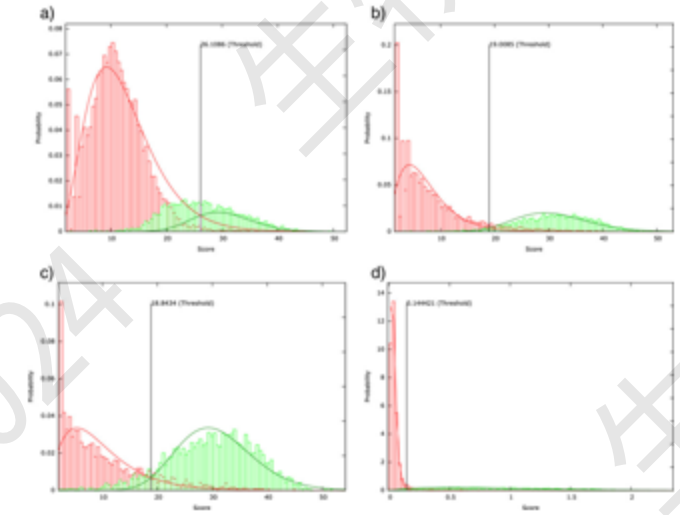


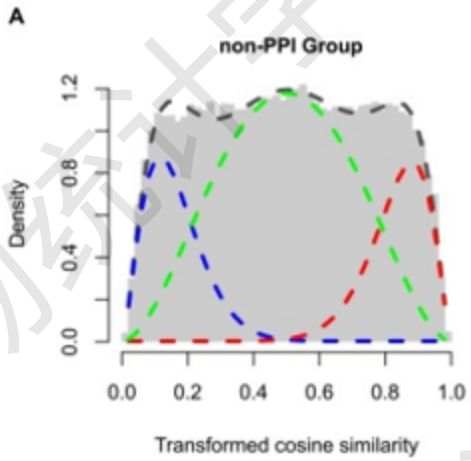


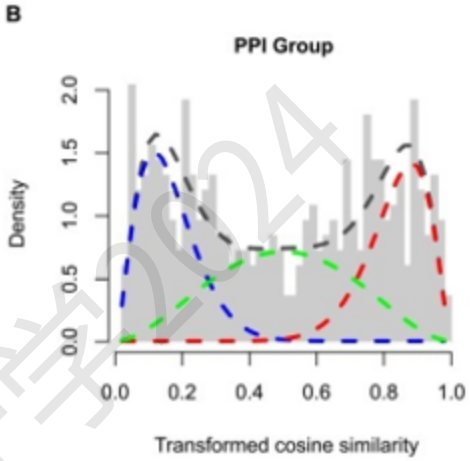


Reference: http://varianceexplained.org/r/mixture-models-baseball/

Other mixture models





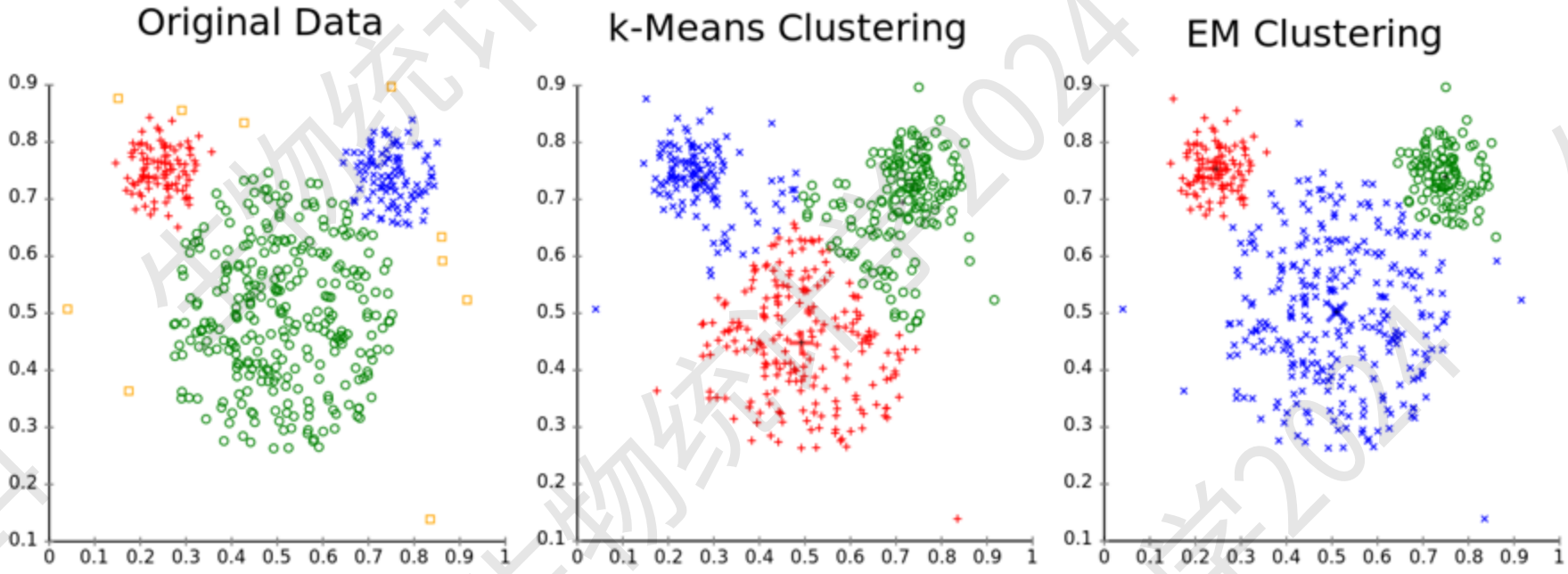


GMM: Gamma-Mixture Model

BMM: Beta-Mixture Mode

K-means VS. EM

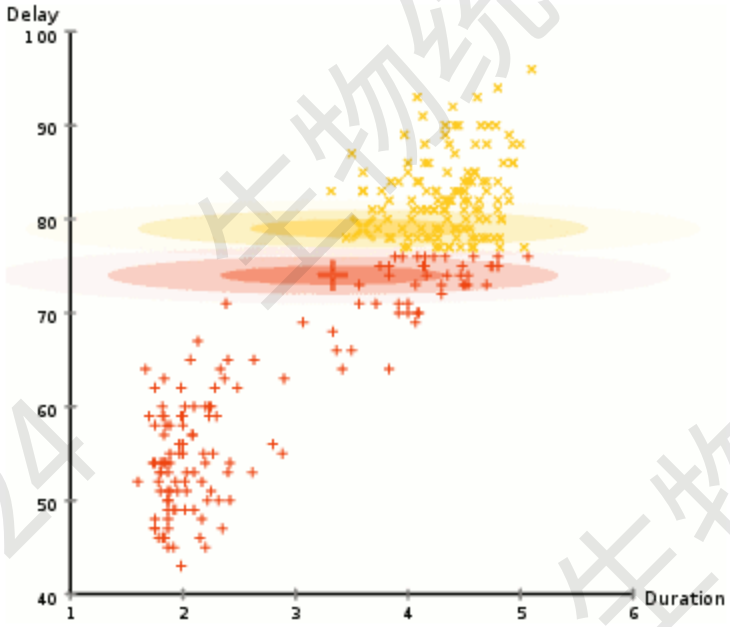
Different cluster analysis results on "mouse" data set:

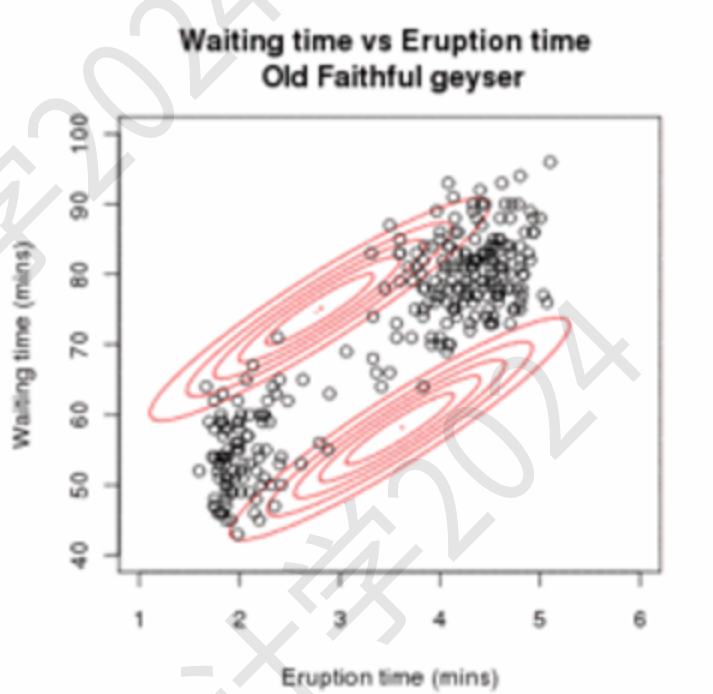


EM聚类与K-Means不同之处：并不计算距离，而是计算概率（并且明显要比K-Means复杂的多），用一个给定的多元高斯概率分布模型来估计出一个数据点属于一个聚类的概率，即将每一个聚类看作是一个高斯模型。

Reference:

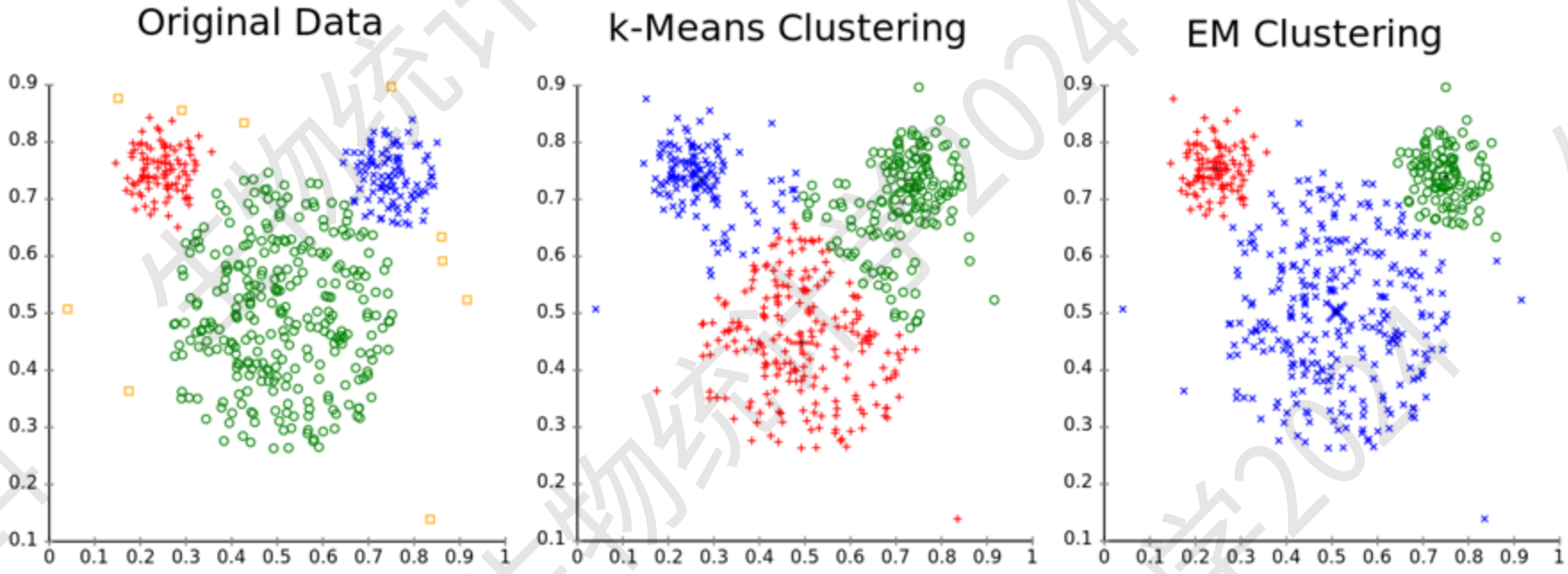
K-means VS. EM





K-means VS. EM

Different cluster analysis results on "mouse" data set:

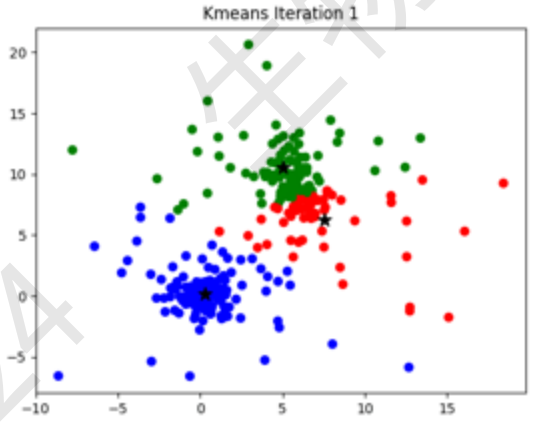


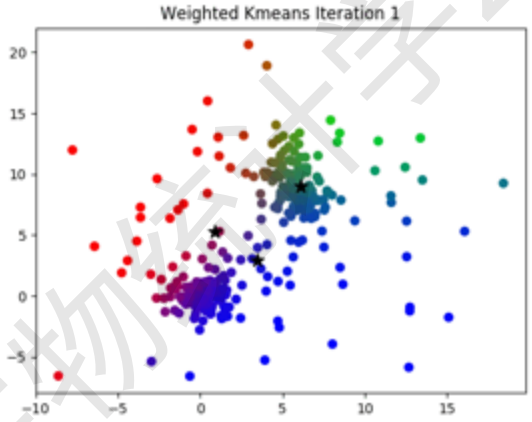
EM聚类与K-Means不同之处：并不计算距离，而是计算概率（并且明显要比K-Means复杂的多），用一个给定的多元高斯概率分布模型来估计出一个数据点属于一个聚类的概率，即将每一个聚类看作是一个高斯模型。

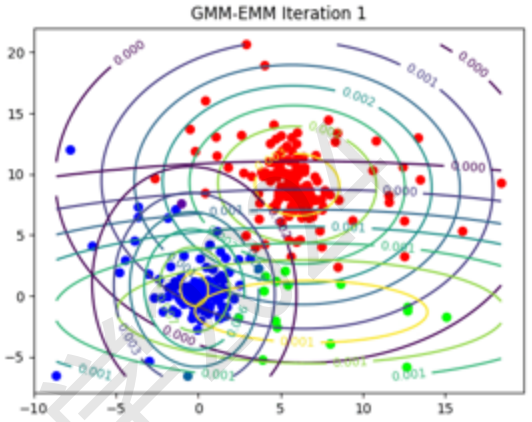
Reference:

K-means VS. EM

GMM-EM: Gaussian Mixture Model (GMM) using Expectation Maximization (EM)



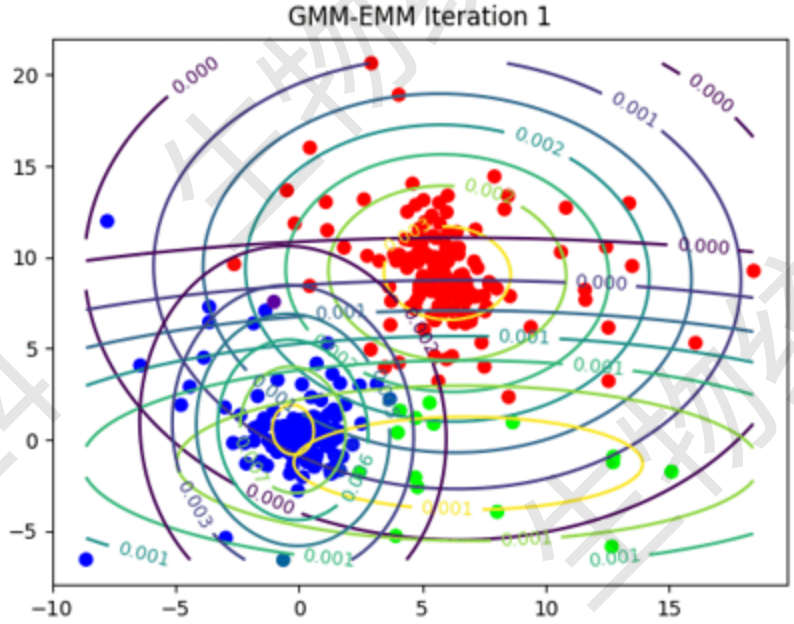


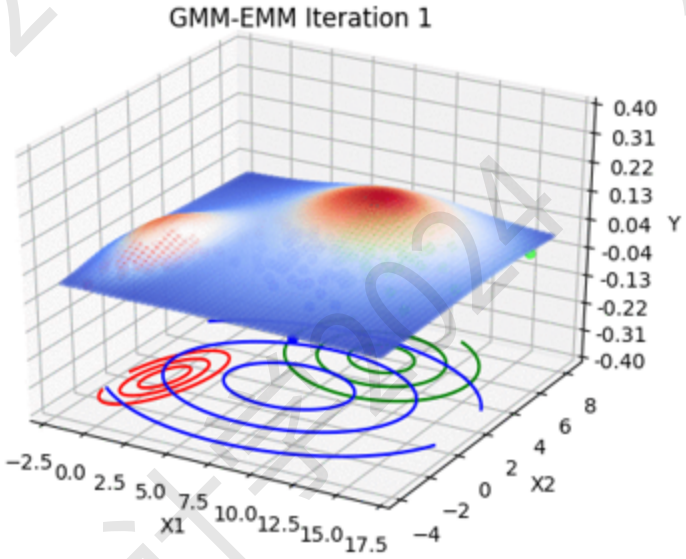


Reference: https://sandipanweb.wordpress.com/2017/03/19/hard-soft-clustering-with-k-means-weighted-k-means-and-gmm-em/

K-means VS. EM

GMM-EM: Gaussian Mixture Model (GMM) using Expectation Maximization (EM)





Reference: https://sandipanweb.wordpress.com/2017/03/19/hard-soft-clustering-with-k-means-weighted-k-means-and-gmm-em/

Choose the Number of Clusters

- Akaike Information Criterion (AIC)

$$AIC = -2l(\hat{\Theta}_g) + 2v_g$$

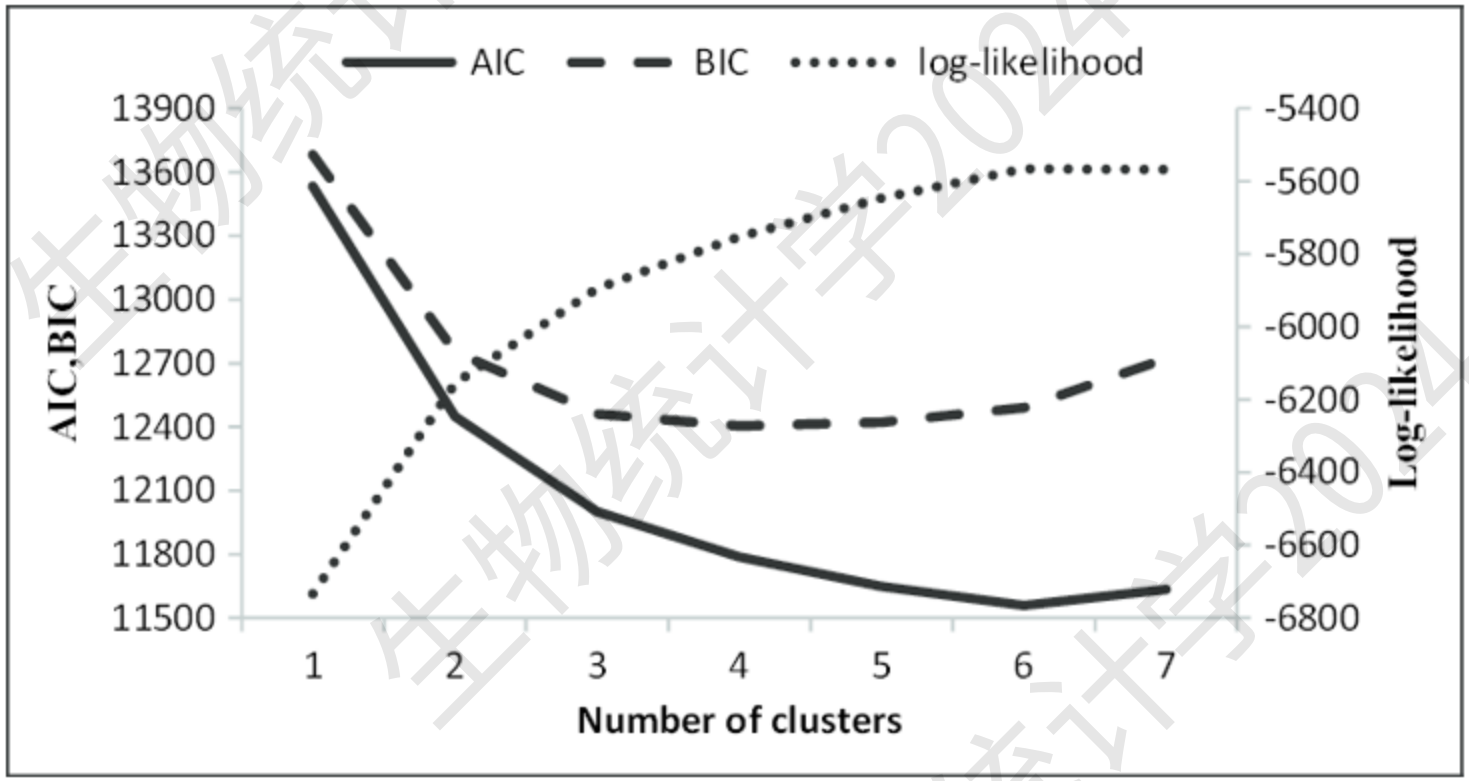
- Bayesian Information Criterion (BIC)

$$BIC = -2l(\hat{\Theta}_g) + v_g \log(n)$$

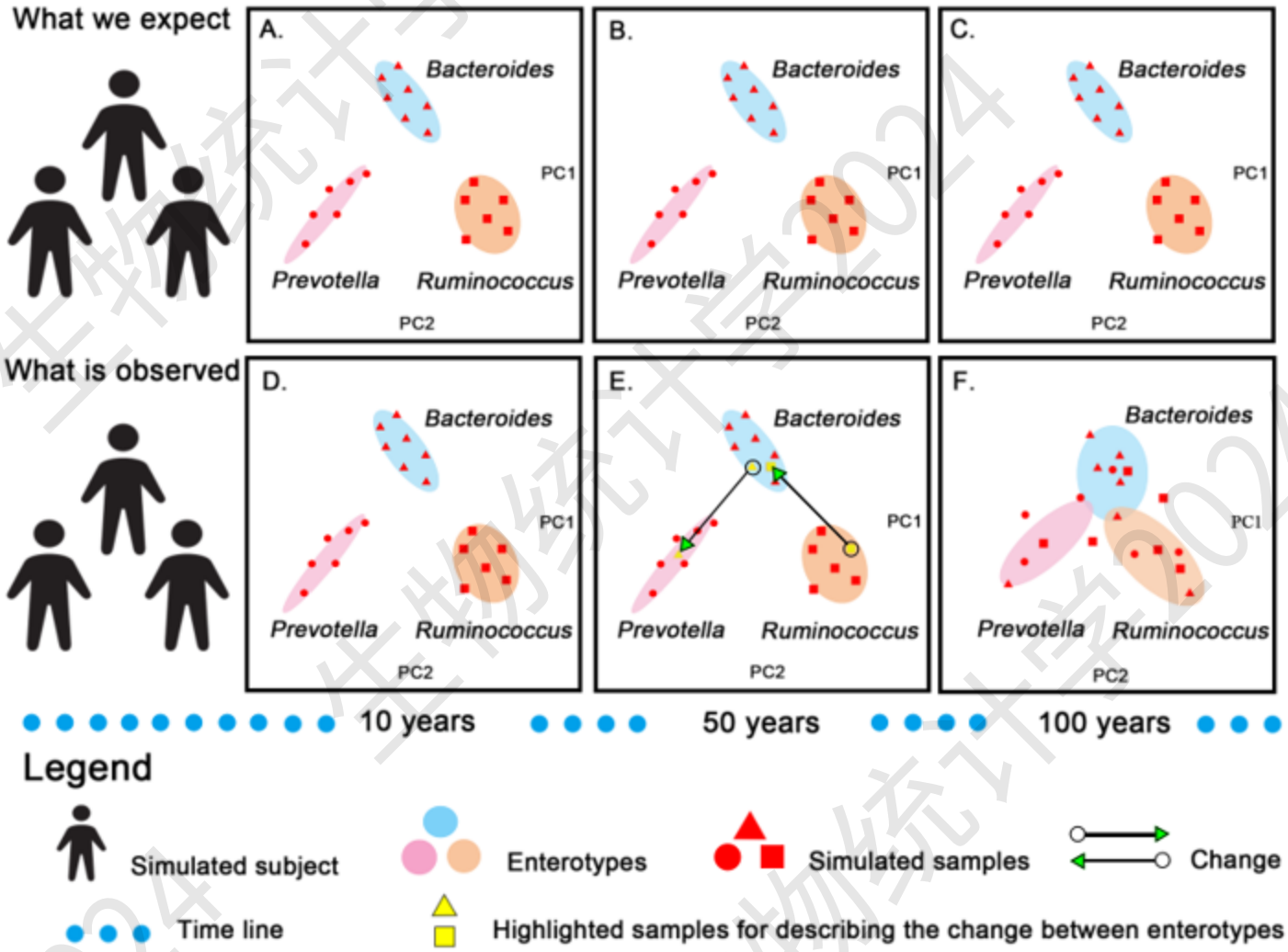
where v_g is the number of independent parameters

1. Akaike H: Information theory and an extension of the maximum likelihood principle. In 2nd Int Symp Information Theory. Edited by Petrov BN, Csaki F. Budapest: Akademiai Kiado,1973, 267-281.
2. Schwartz G: Estimating the dimensions of a model. Annls Statistics 1978, 6:461-464.

Choose the Number of Clusters



Enterotype calculation



Reference: Cheng, et al., GPB, 2018

第6-2章: Classification and Prediction

- Bayesian decision rule
- Fisher linear discriminant analysis
- SVM
- Aggregating classifiers

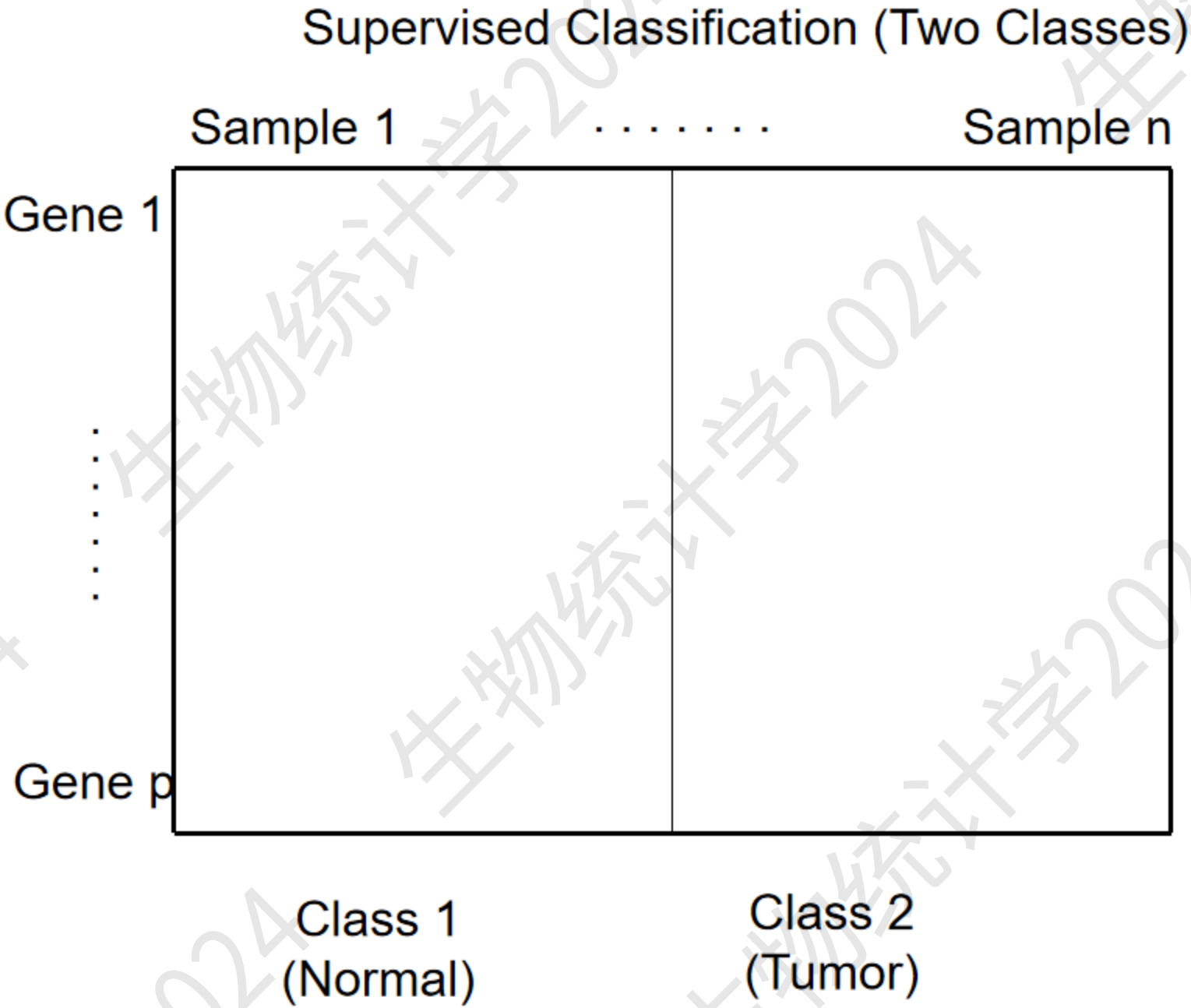
- Reference

- T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, E. S. Lander. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science 15 October 1999: Vol. 286 no. 5439 pp. 531-537.
- Ji Zhu and Trevor Hastie. Classification of gene microarrays by penalized logistic regression. Biostatistics (2004), 5, 3, pp. 427–443.

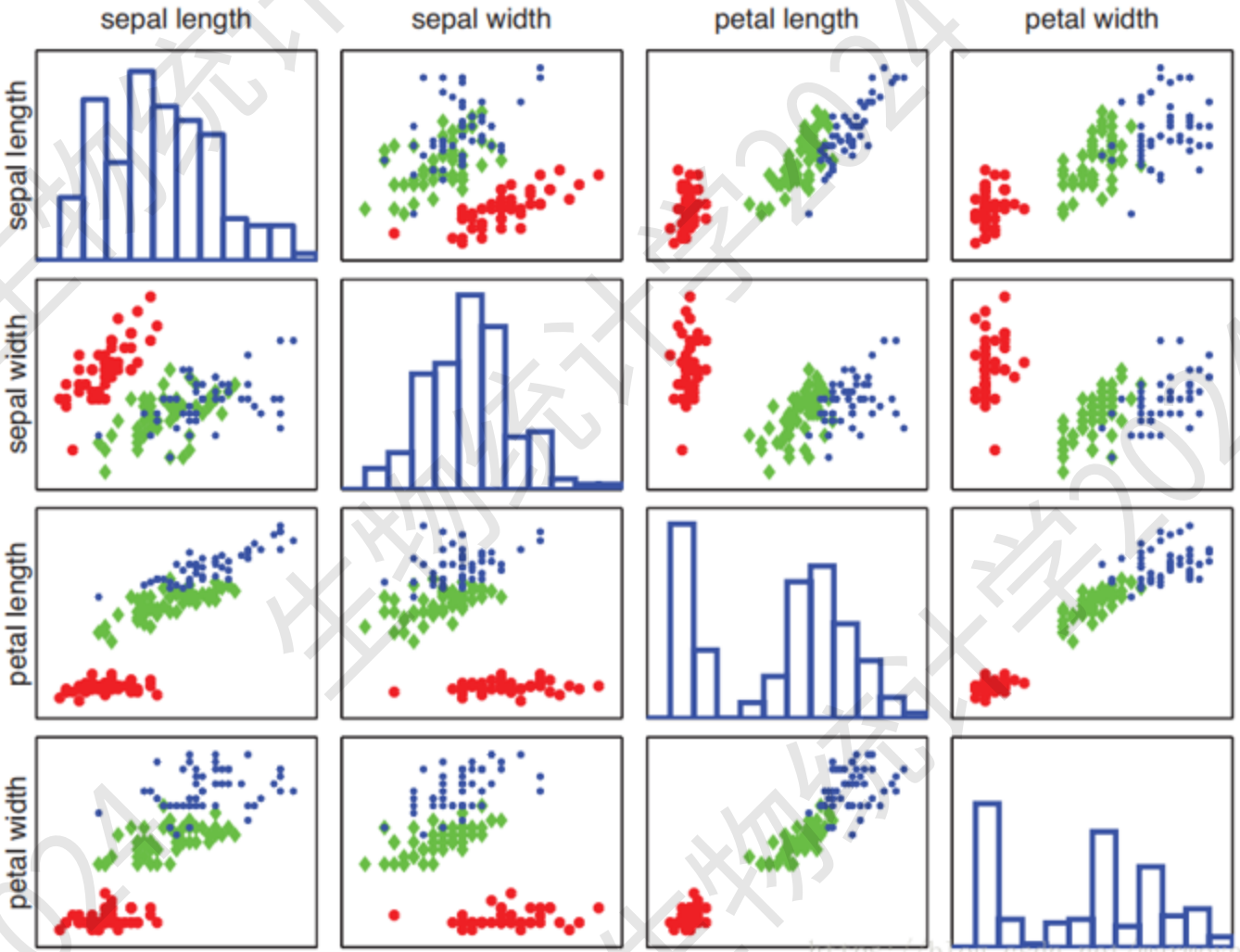
部分Slides来自于http://www.epibiostat.ucsf.edu/biostat/cbmb/courses/CBMBdiscrimination.ppt

Classification (分类)

- Task: assign objects to classes (groups) on the basis of measurements made on the objects
- Unsupervised: classes unknown, want to discover them from the data (cluster analysis)
- Supervised: classes are predefined, want to use a (training or learning) set of labeled objects to form a classifier for classification of future observations


Supervised Classification (Two Classes)
Sample 1
Sample n
Gene 1
Gene p
Class 1
(Normal)
Class 2
(Tumor)

Classification (分类)



Example: Tumor Classification

- Reliable and precise classification essential for successful cancer treatment
- Current methods for classifying human malignancies rely on a variety of morphological, clinical and molecular variables
- Characterize molecular variations among tumors by monitoring gene expression (microarray)
- Hope: that microarrays will lead to more reliable tumor classification (and therefore more appropriate treatments and better outcomes)

Tumor Classification Using Array Data

Three main types of statistical problems associated with tumor classification:

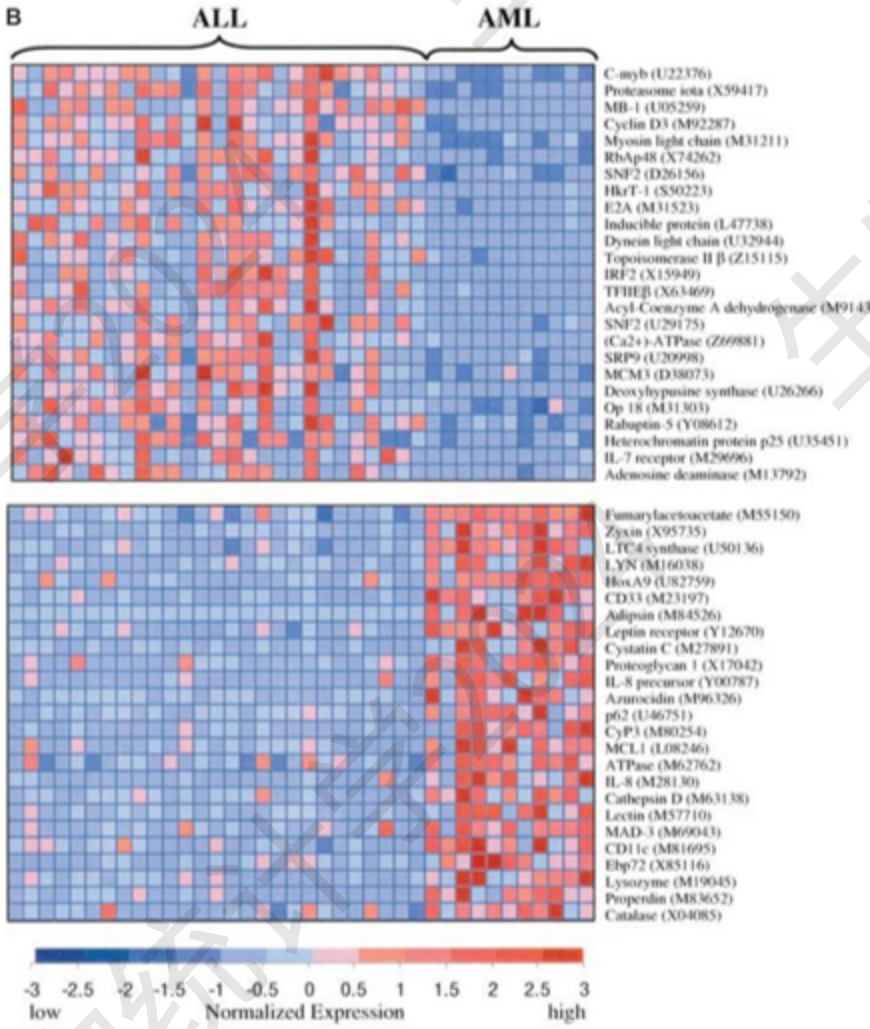
- Identification of new/unknown tumor classes using gene expression profiles (unsupervised learning – clustering)
- Classification of malignancies into known classes (supervised learning – discrimination)
- Identification of "marker" genes that characterize the different tumor classes (feature or variable selection).

Molecular classification of cancer

Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring

T. R. Golub,[1,2*†] D. K. Slonim,[1†] P. Tamayo,[1] C. Huard,[1] M. Gaasenbeek,[1] J. P. Mesirov,[1] H. Coller,[1] M. L. Loh,[2] J. R. Downing,[3] M. A. Caligiuri,[4] C. D. Bloomfield,[4] E. S. Lander[1,5*]

Although cancer classification has improved over the past 30 years, there has been no general approach for identifying new cancer classes (class discovery) or for assigning tumors to known classes (class prediction). Here, a generic approach to cancer classification based on gene expression monitoring by DNA microarrays is described and applied to human acute leukemias as a test case. A class discovery procedure automatically discovered the distinction between acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) without previous knowledge of these classes. An automatically derived class predictor was able to determine the class of new leukemia cases. The results demonstrate the feasibility of cancer classification based solely on gene expression monitoring and suggest a general strategy for discovering and predicting cancer classes for other types of cancer, independent of previous biological knowledge.



Molecular classification of cancer: class discovery and ... - NCBI

https://www.ncbi.nlm.nih.gov › pubmed

by TR Golub - 1999 - Cited by 13212 - Related articles

Classifiers

- A predictor or classifier partitions the space of gene expression profiles into K disjoint subsets, A_1, ..., A_K, such that for a sample with expression profile $\mathbf{X}=(X_1, ...,X_G) \in A_k$ the predicted class is k
- Classifiers are built from a learning set (LS)

$$L = (\mathbf{X}_1, Y_1), ..., (\mathbf{X}_n, Y_n)$$

- Classifier C built from a learning set L:

$$C(. ,L): \mathbf{X} \rightarrow \{1,2, ... ,K\}$$

- Predicted class for observation **X**:

$$C(\mathbf{X},L) = k \text{ if } \mathbf{X} \text{ is in } A_k$$

Bayes公式

- 后验概率: $p(\omega_k|\overrightarrow{x})$
- 先验概率: π_k, 其中$\Sigma\pi_k$=1.
- Bayes公式

$$p(\omega_k|\overrightarrow{x}) = \frac{p(\overrightarrow{x}|\omega_i)\pi_i}{\sum_{l=1}^{K} p(\overrightarrow{x}|\omega_l)\pi_l}$$

Bayes决策

- 将观测向量$\overrightarrow{x}$归入后验概率最大的类，即

$$\hat{\omega} = \underset{k}{\text{Argmax}}\, p(\omega_k|\overrightarrow{x})$$

- 利用Bayes公式，上式等价于

$$\hat{\omega} = \underset{k}{\text{Argmax}}\, p(\overrightarrow{x}|\omega_k)p(\omega_k)$$

最小错误Bayes决策

- 错误函数

$$\Pr(\text{error}) = \sum_{i=1}^{K} \Pr(\text{error}|\omega_i)p(\omega_i)$$

- 单个类别的错误函数

$$\Pr(\text{error}|\omega_i) = \int_{\overline{\Omega_i}} \Pr(\vec{x}|\omega_i)d\vec{x}$$

其中Ω_i为归入类别ω_i的区域。

最小错误Bayes决策

$$
\begin{aligned}
\Pr(\text{error}) &= \sum_{i=1}^{K} \int_{\overline{\Omega_i}} p(\overrightarrow{x}|\omega_i) p(\omega_i) d\overrightarrow{x} \\
&= \sum_{i=1}^{K} p(\omega_i)(1 - \int_{\Omega_i} p(\overrightarrow{x}|\omega_i) d\overrightarrow{x}) \\
&= 1 - \sum_{i=1}^{K} p(\omega_i) \int_{\Omega_i} p(\overrightarrow{x}|\omega_i) d\overrightarrow{x}
\end{aligned}
$$

最小错误Bayes决策

- 因此,为了使p(error)最小,只要选择区域Ω_i,使得正确分类概率最大

$$\sum_{i=1}^{K} p(\omega_i) \int_{\Omega_i} p(\vec{x}|\omega_i) d\vec{x}$$

- Bayes决策使得正确分类概率最大

$$c = \int \max_i p(\omega_i) p(\vec{x}|\omega_i) d\vec{x}$$

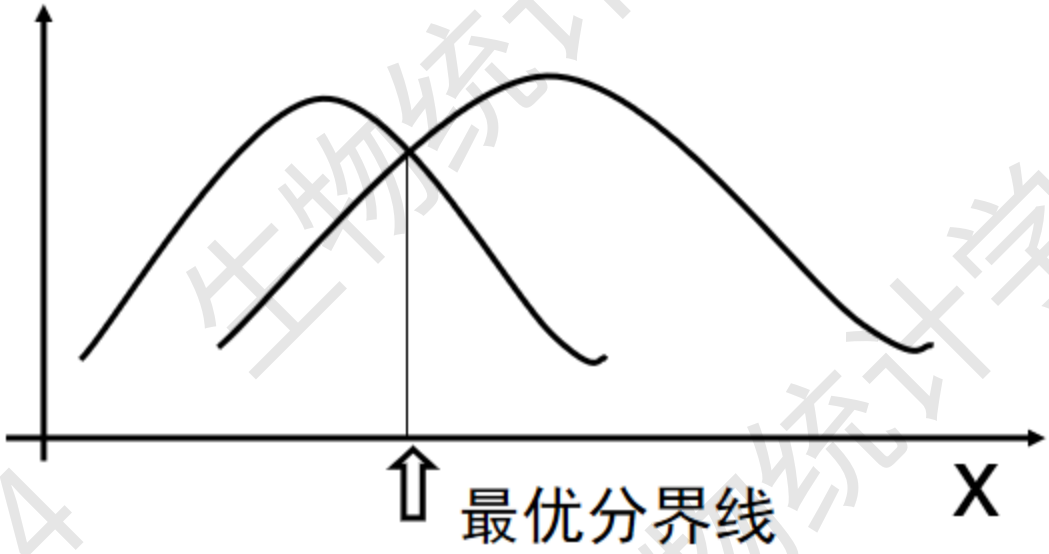
- Bayes错分概率为

$$e_B = 1 - c = 1 - \int \max_i p(\omega_i) p(\vec{x}|\omega_i) d\vec{x}$$

两类问题Bayes决策

- 对于一维观测量x, 最小错误率对应于如下的最优分界线下的错误率,

$$e_B = p(\omega_2)\int_{\Omega_1} p(x|\omega_2)dx + p(\omega_1)\int_{\Omega_2} p(x|\omega_1)dx$$



判别分析

- 最初由 R.A. Fisher提出.

Fisher判别分析

寻找一个投影方向，使得在该方向上

- 极小化组内距离。
- 极大化组间距离。

➔ 基本的目标在于提取最有效的分类特征

Fisher判别分析：分散矩阵

- 组间距

$$S_B = \sum_{i=1}^{K} \frac{n_i}{n} (\mu_i - \mu)(\mu_i - \mu)^T$$

- 组内距

$$S_W = \sum_{i=1}^{K} \frac{n_i}{n} \hat{\Sigma}_i$$

投影方向上的分散矩阵

- 投影变换$u=\alpha^T x$
- 投影后组间距$S_B(u)=\alpha^T S_B \alpha$.
- 投影后组内距$S_W(u)=\alpha^T S_W \alpha$.

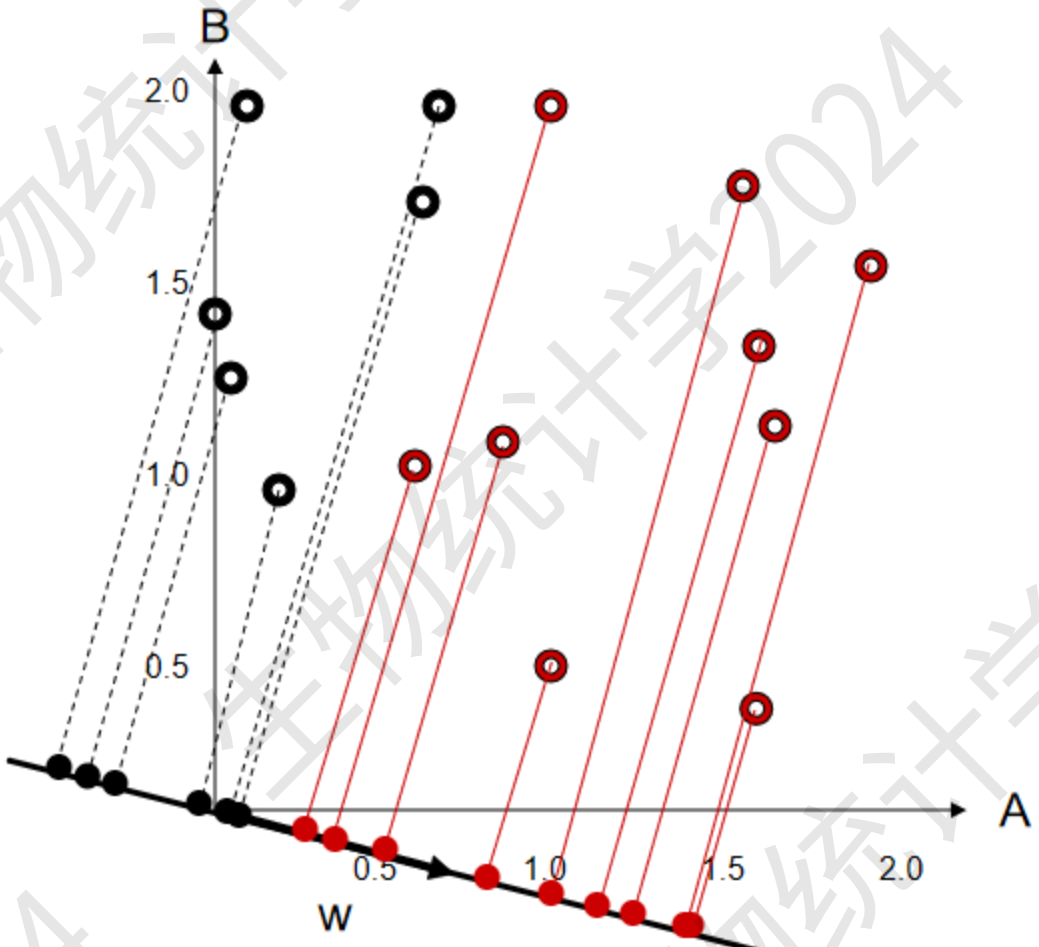
Fisher判别分析

- Fisher准则：投影后的组间距和组内距比值最大，

$$J_F(\alpha) = \frac{\alpha^T S_B \alpha}{\alpha^T S_W \alpha}$$

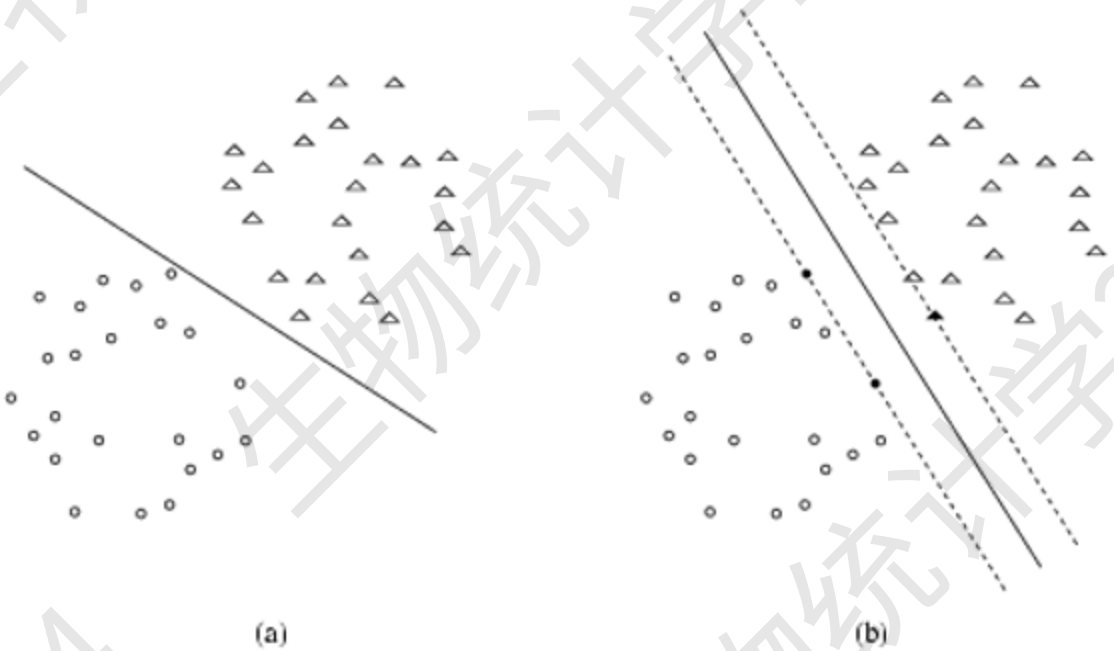
$$\hat{\alpha} = \underset{\alpha}{\text{Argmax}}\, J_F(\alpha)$$

两类问题的Fisher判别分析



Find Optimal Hyperplane

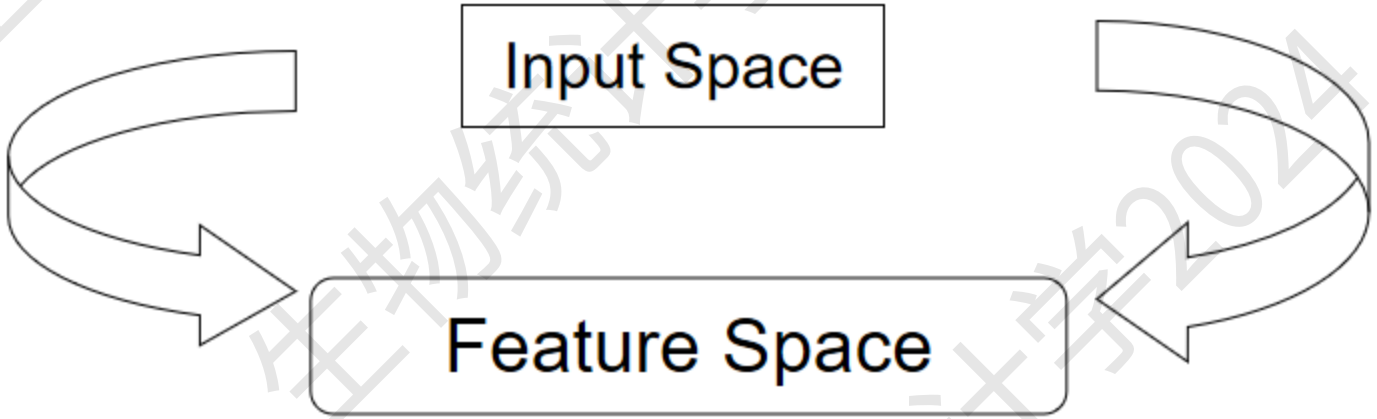
- Assumption: samples are linearly separable.
- A better generalization is expected from (b).



SVM

Extending the Hyperplane

- Mapping the input samples to feature space in higher dimension.

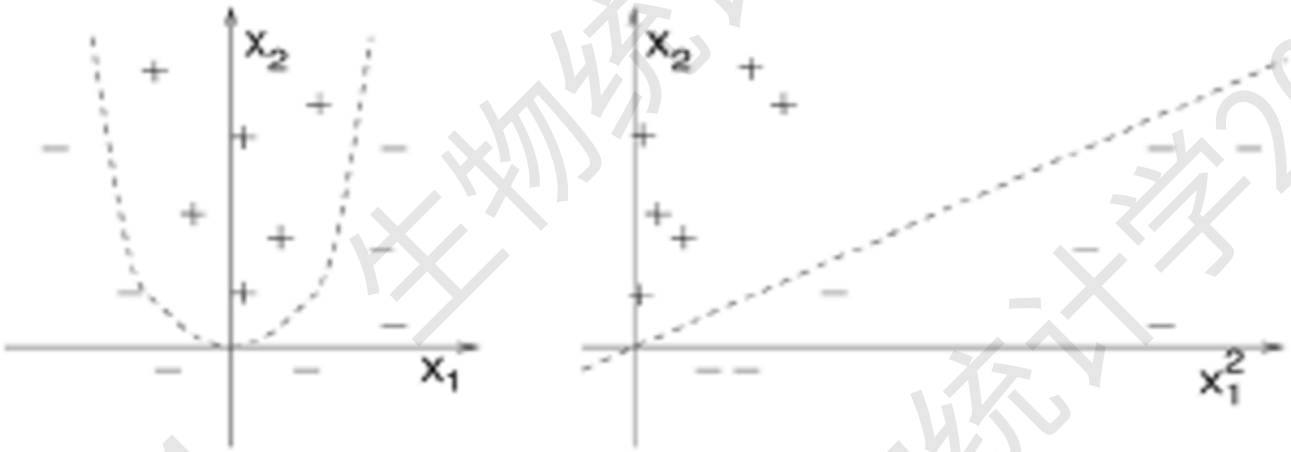


Finding the hyperplance in feature space.

SVM

Example

- Input space with two attributes: (x_1,x_2).
- Feature space with 6 attributes: $(x_1^2, x_2^2, x_1, x_2, x_1x_2, 1)$



SVM

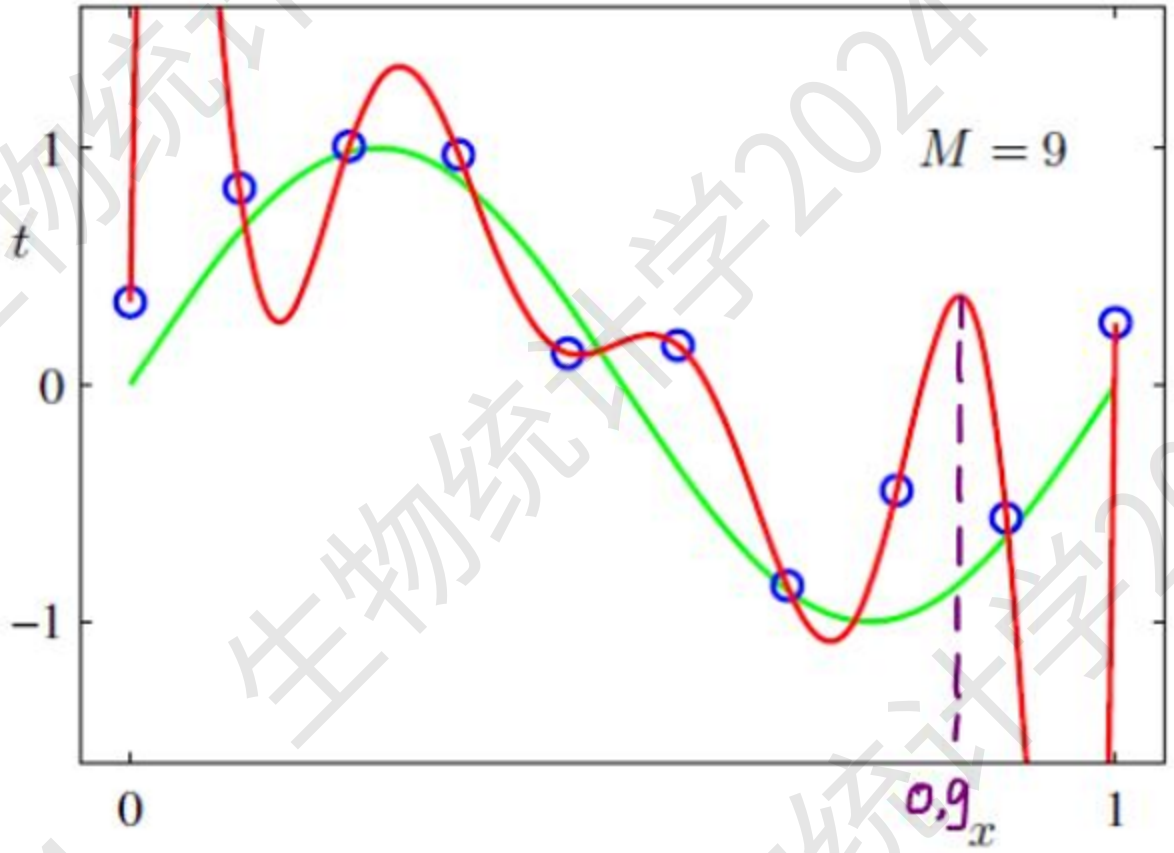
Why does SVM work well

- Measurement: the risk of misclassifying of training samples and test samples (generalization problem).
- Principle: structural risk minimization.
- Key concept: VC dimension.

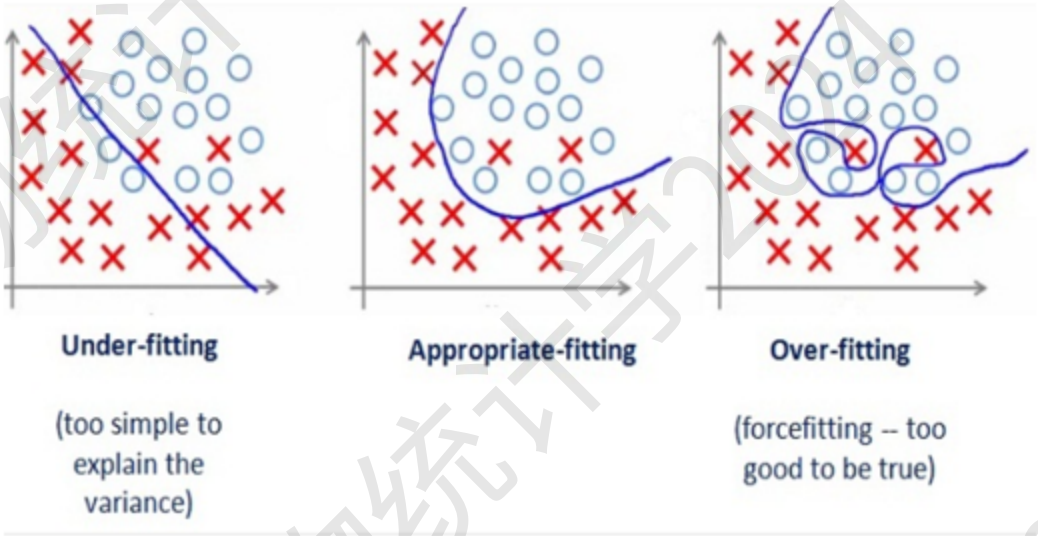
Other Classifiers Include…

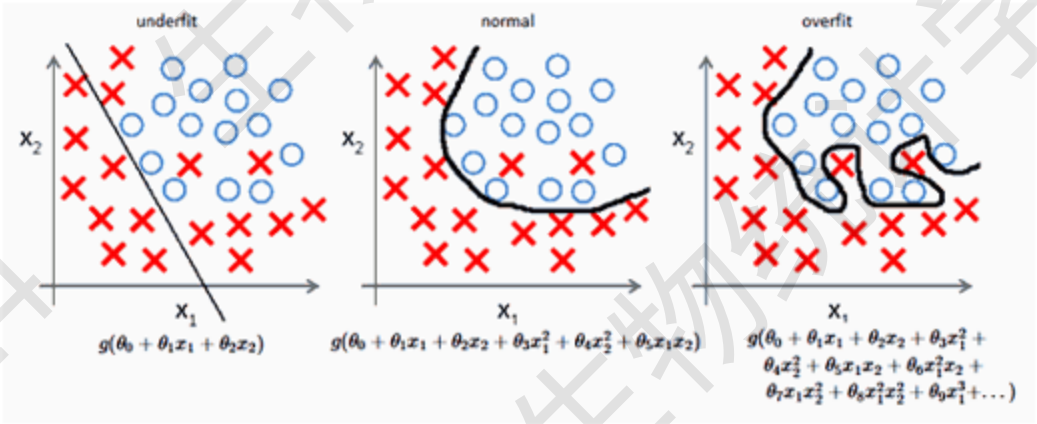
- Decision tree (CART, C4.5)
- Neural networks
- Nearest neighbour (KNN)
- Logistic regression
- Projection pursuit
- Bayesian belief networks

Overfitting problem

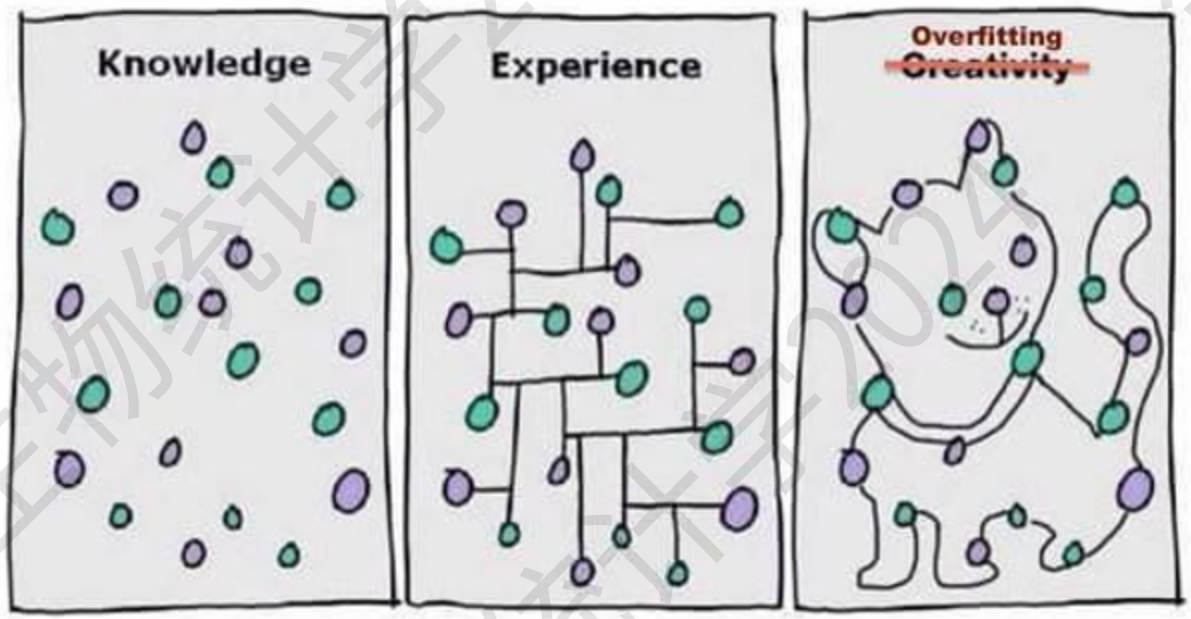


Overfitting problem





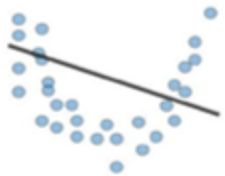
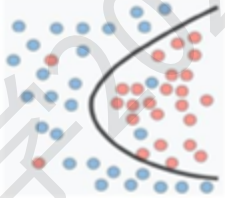
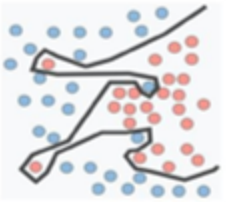
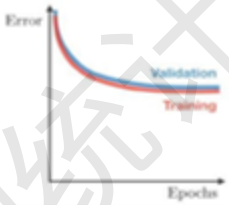
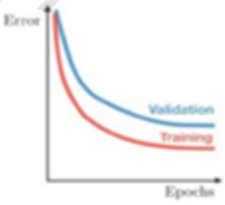
过拟合



- 太简单—欠拟合，没有充分的挖掘数据中的规律
- 太复杂—过拟合，过分挖掘数据中的规律了

https://www.hrwhisper.me/machine-learning-regularization/

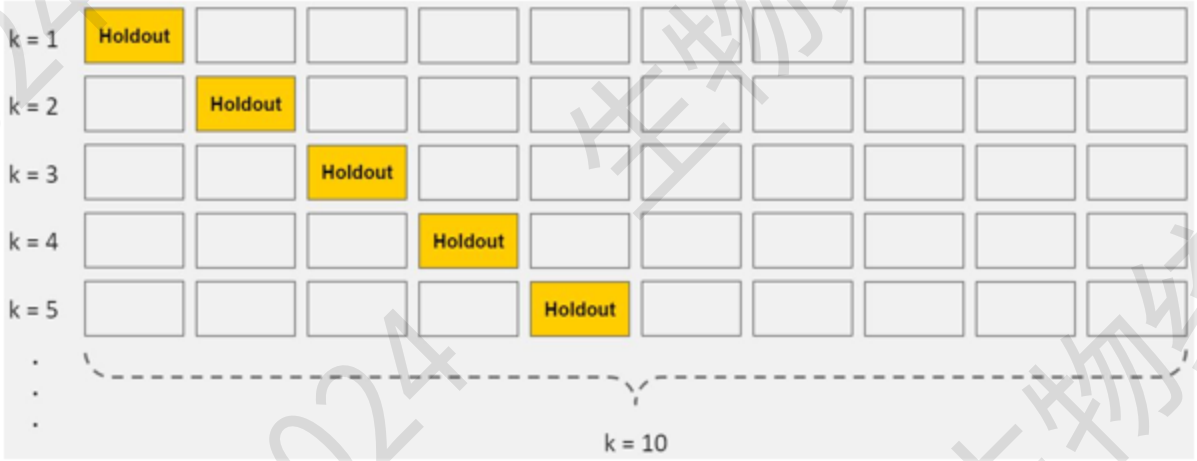
Bias/variance tradeoff

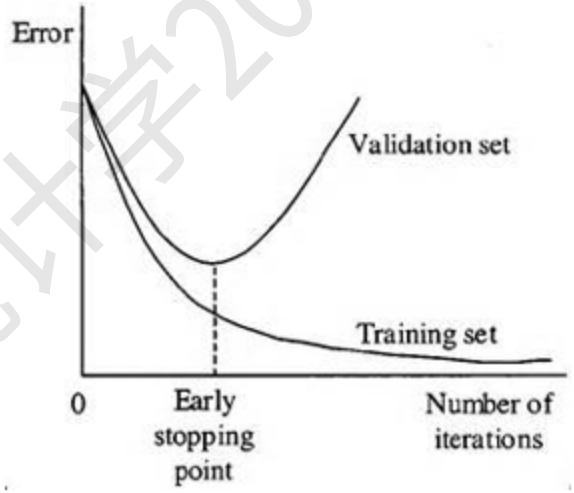
	Underfitting	Just right	Overfitting
Symptoms	• High training error • Training error close to test error • High bias	• Training error slightly lower than test error	• Very low training error • Training error much lower than test error • High variance
Regression illustration			
Classification illustration			
Deep learning illustration			
Possible remedies	• Complexify model • Add more features • Train longer		• Perform regularization • Get more data

Overfitting problem

How to avoid overfitting

- Cross-validation
- Train with more data
- Remove features
- Early stopping





评价统计模型对某个数据集的预测精度

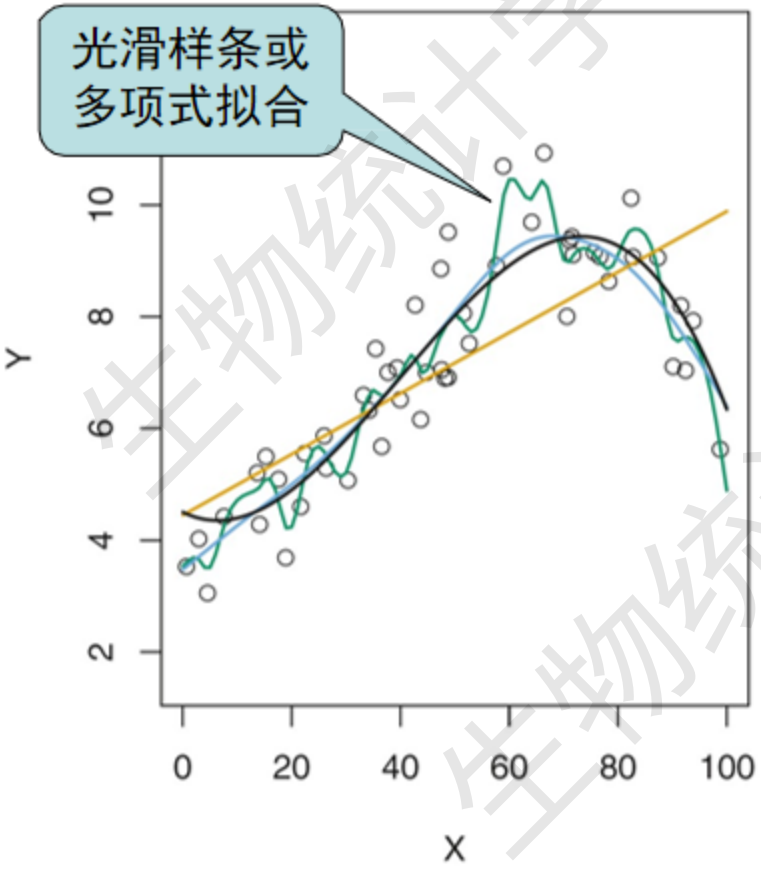
- 对一个给定的观测，测量预测的响应值与真实响应值间的接近程度。均方误差（mean square error):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$$

- **存在问题**：MSE使用训练数据计算（training MSE），而训练数据也用来估计f，可能会低估误差。

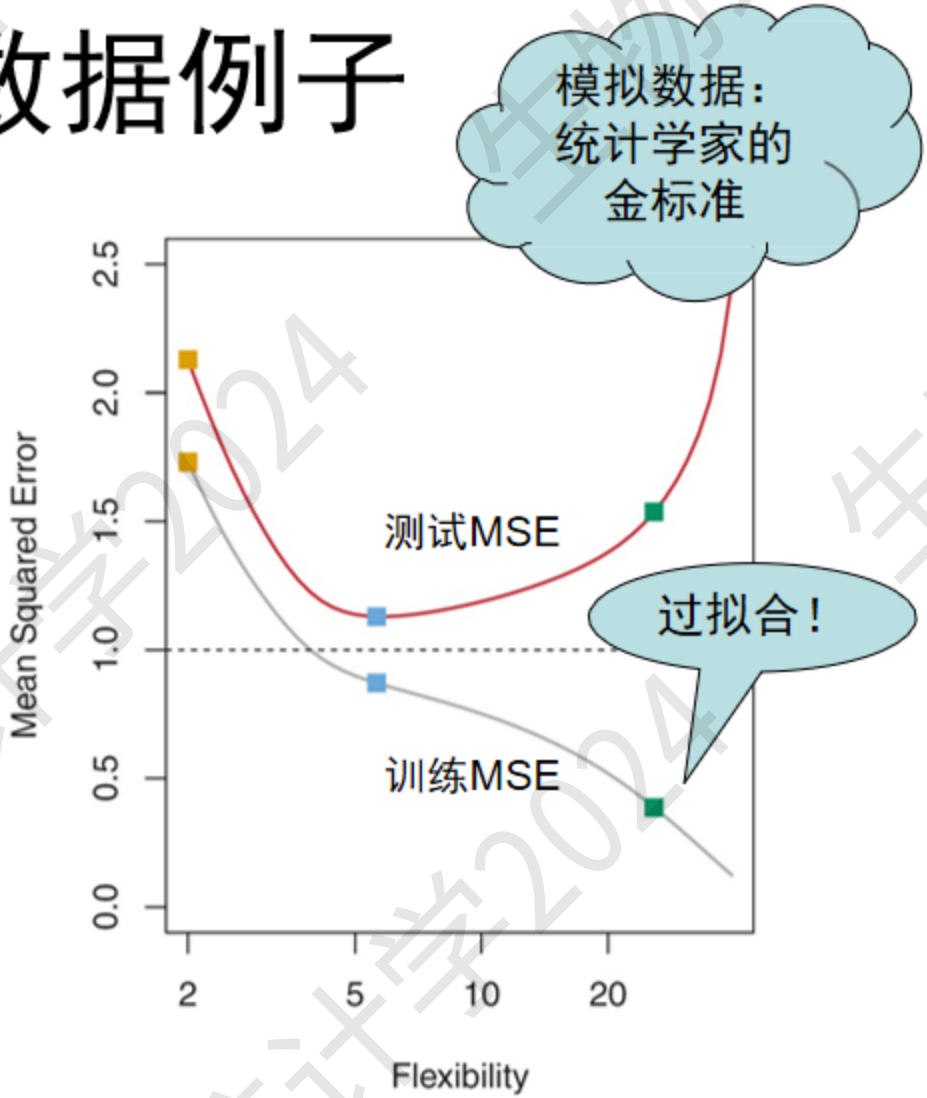
- 更关心将模型用于测试数据时的预测精度（test MSE）
  - 预测未来的股票价格、新病人的疾病风险
  - 测试数据没有用作拟合模型

模拟数据例子



黑色曲线是真实函数 ***f***，
用以产生模拟数据点

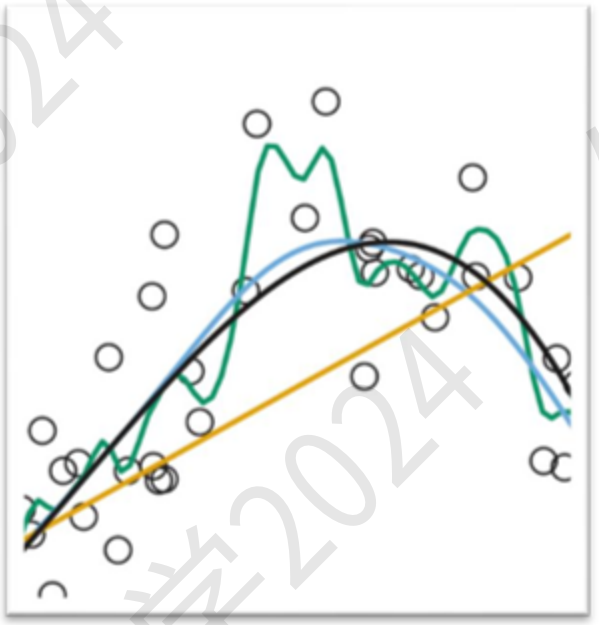
$$Y = f(X) + \epsilon$$



自由度：描述曲线灵活程度，
简单线性回归是2（参数的个数少，限定性强）

过拟合

- **原因**：The statistical learning procedure is working too hard to find patterns in the training data, and may be picking up some patterns that are just caused by random chance rather than by true properties of the unknown function *f*.



- Overfitting refers specifically to the case in which a less flexible model would have yielded a smaller test MSE.

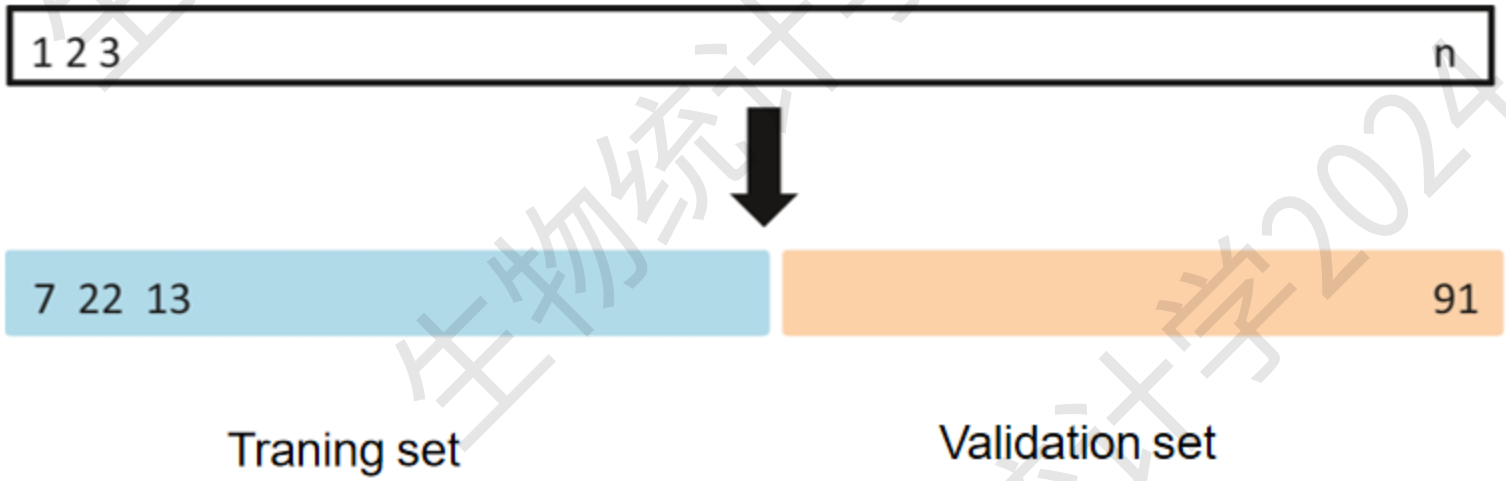
- **实际数据问题**：不易得到测试数据来估计测试MSE。

解决办法：**交叉验证**

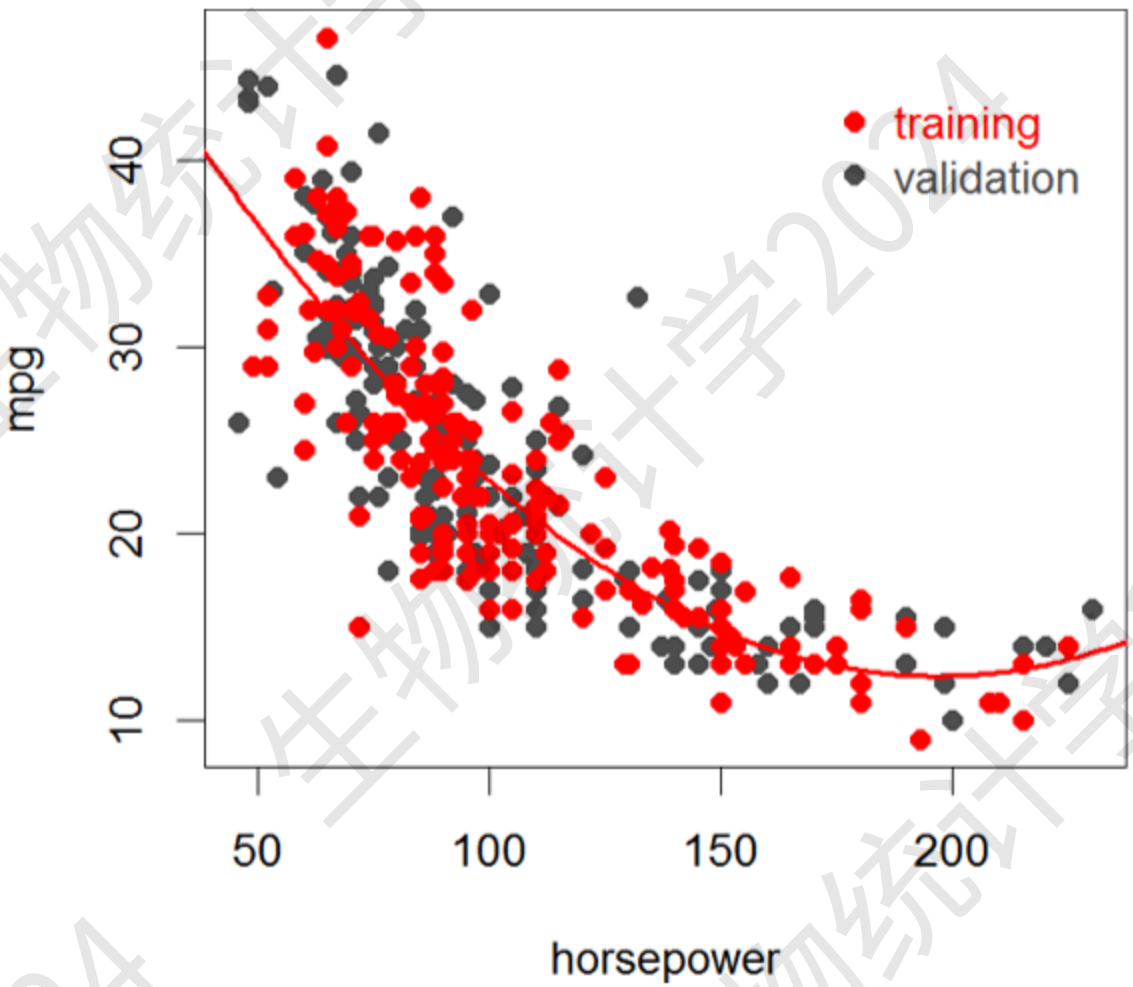
交叉验证法

在拟合过程中，保留出训练数据的一个子集不用作训练模型，而用来估计test MSE。

1、验证集（validation set）方法

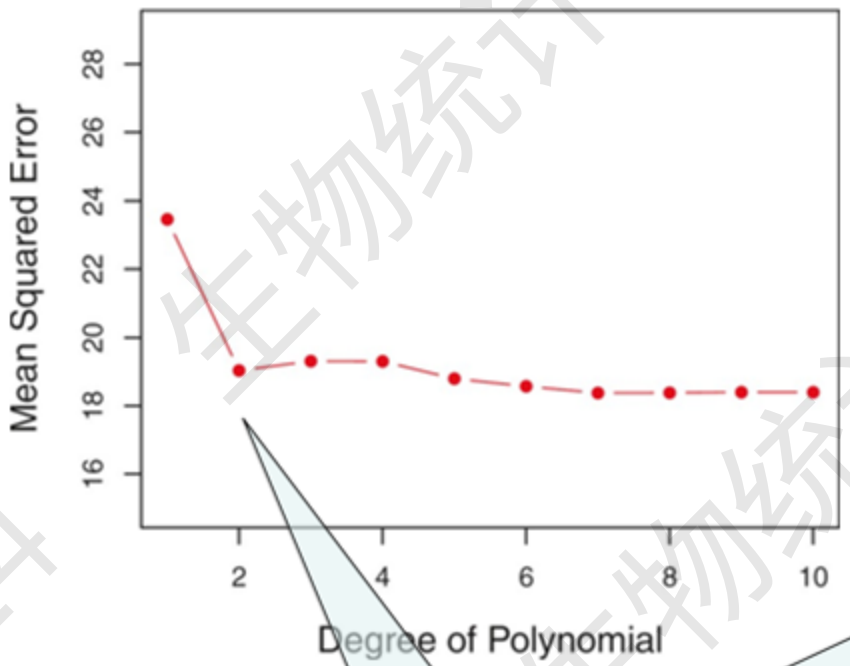


Auto 数据集

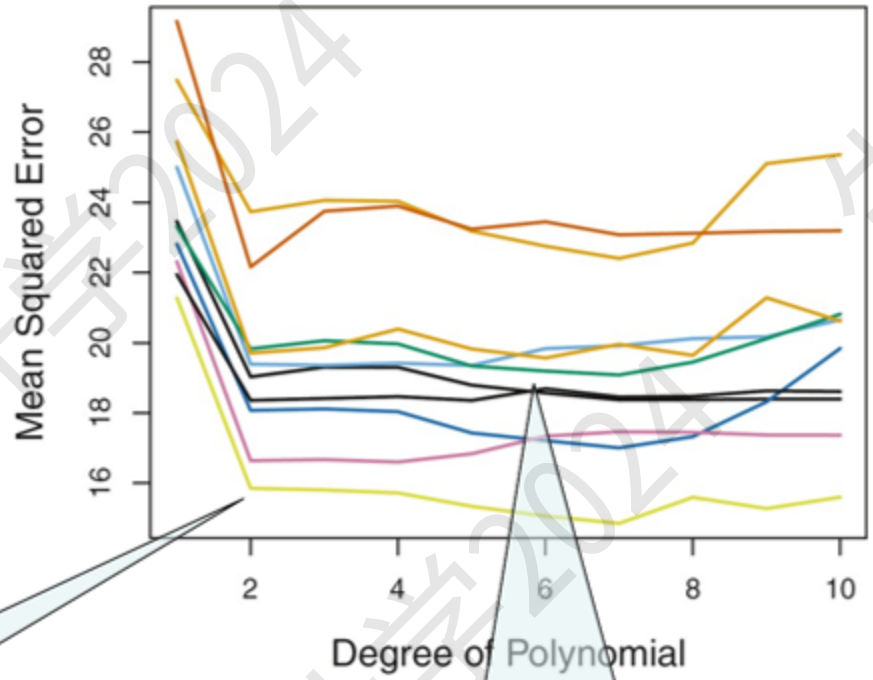


比较不同模型的验证集MSE

一次分割



10次不同的分割



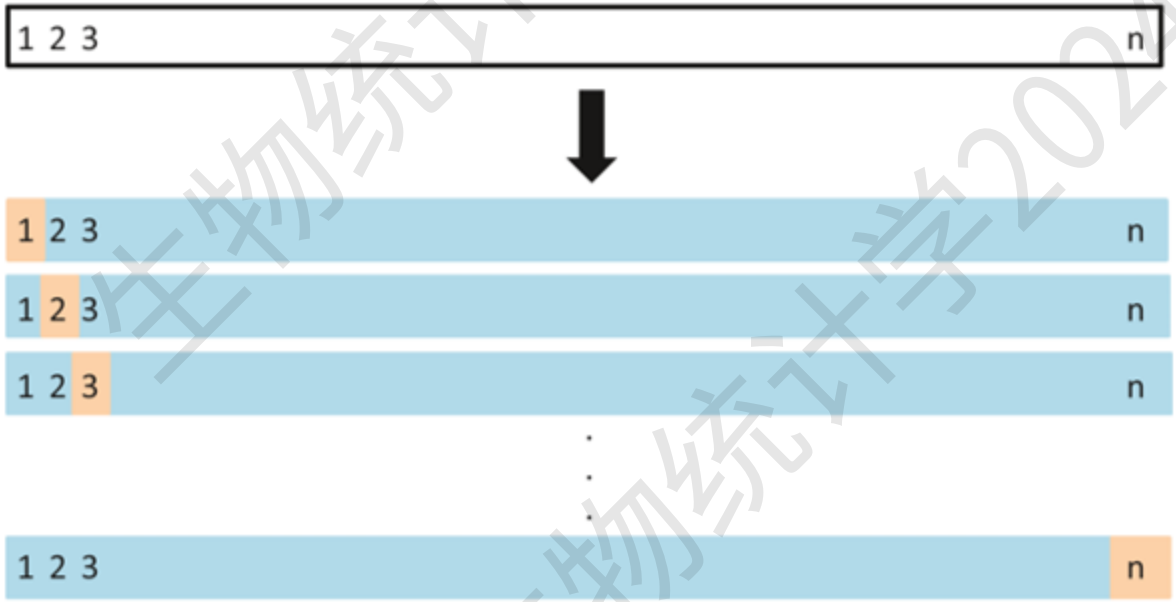
用2次多项式拟合，预测精度提升较大

只使用部分数据训练，模型可能高估MSE，不是最优

不同分割对MSE的估计差异较大，最小MSE的模型也不一样

留一交叉验证法

- Leave-one-out cross-validation (**LOOCV**)



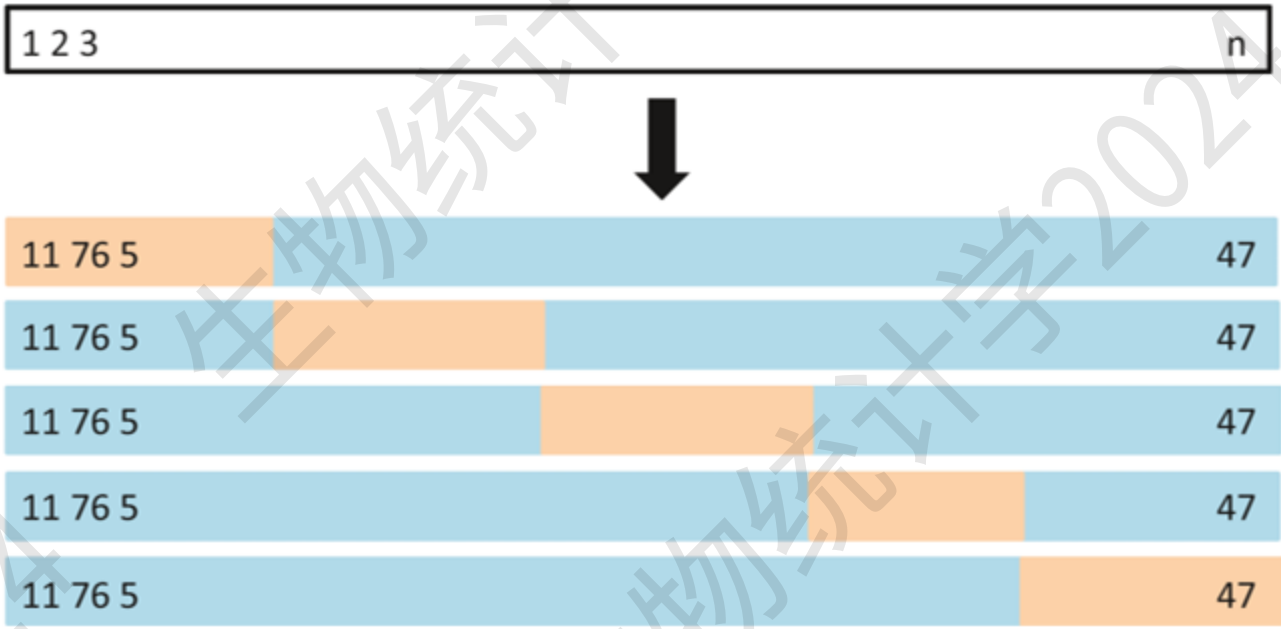
$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{MSE}_i$$

优点

1、使用n-1个样本点训练，和使用所有数据点接近，模型偏差较小，不容易高估测试MSE。

2、多次使用LOOCV总会得到相同的结果

k 折交叉验证法 (k-fold CV)



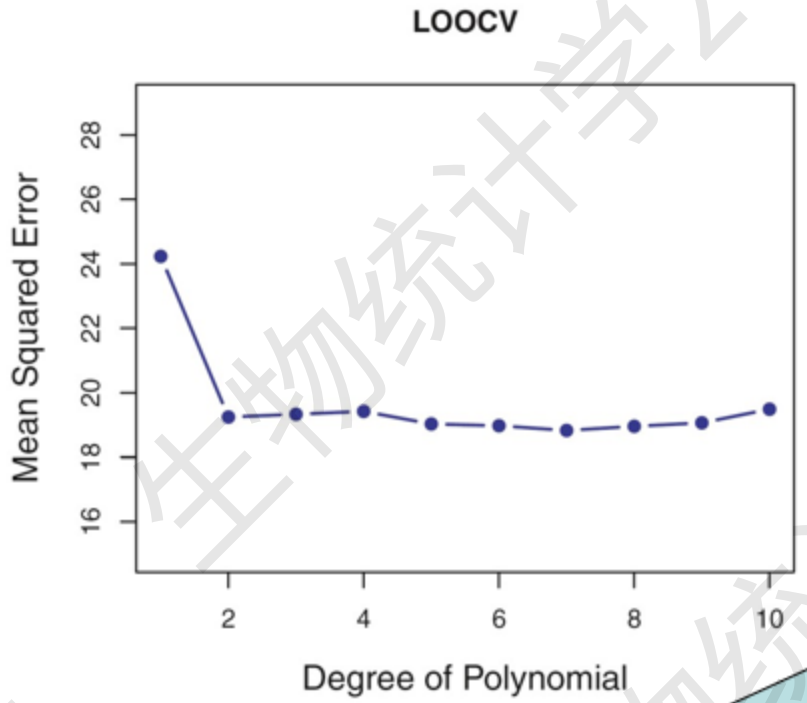
$$\mathrm{CV}_{(k)} = \frac{1}{k}\sum_{i=1}^{k}\mathrm{MSE}_i$$

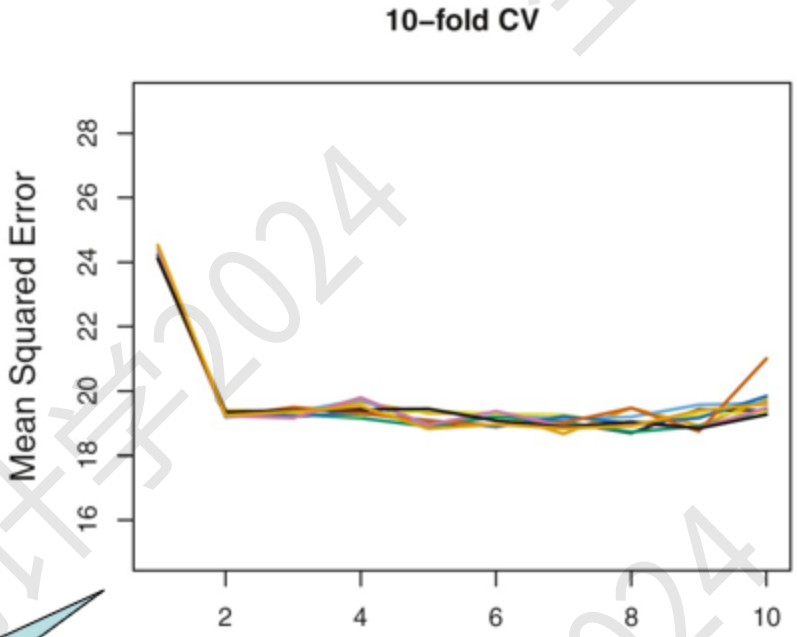
优点

1、LOOCV拟合 n 次模型，计算量较大。k-fold CV 拟合 k （5或10）次模型。

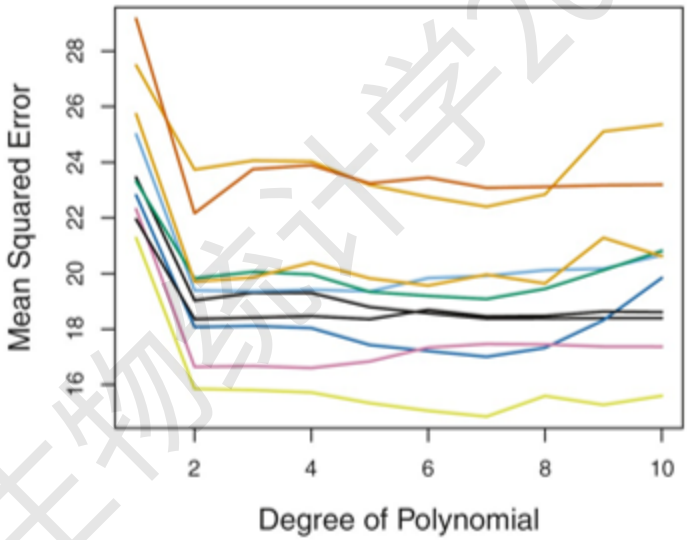
2、每个模型使用 (k-1) / k 样本点，相比验证集方法，更接近使用全部样本点的模型。

LOOCV of the Auto data



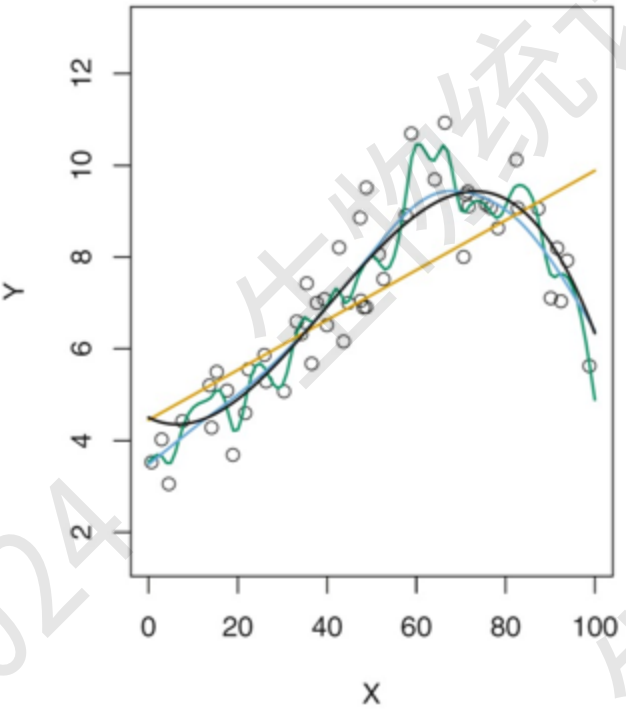


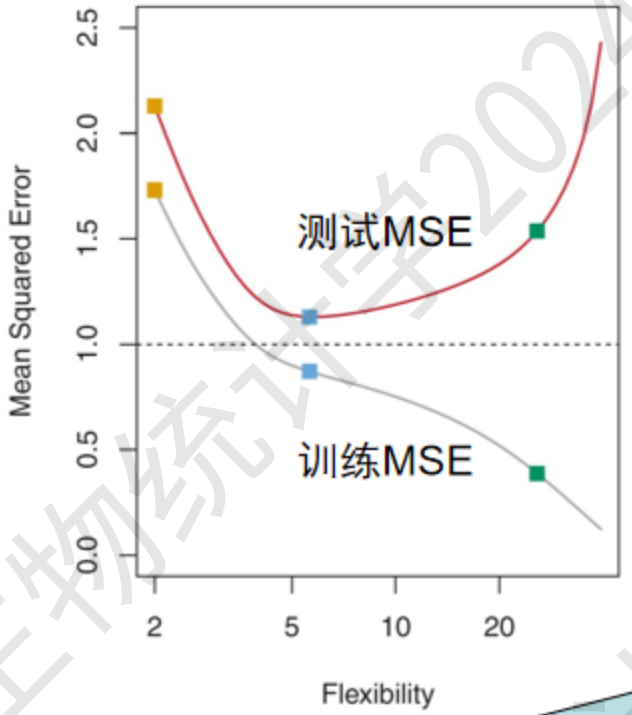
多次不同分割的10-fold CV:
MSE的估计有波动性，但比
验证集方法的波动性小

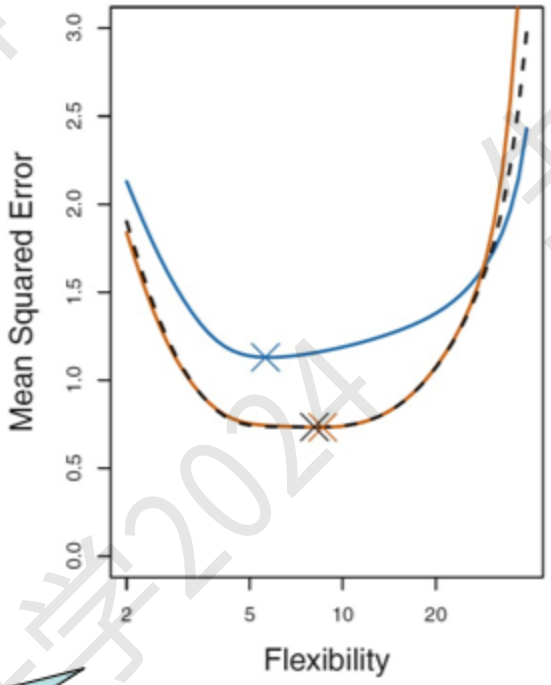


模拟数据例子

统计学家的正对照







- CV估计的MSE曲线形状正确，但低估了测试MSE。
- 可以帮助确定正确模型的参数数量（MSE最小时）

测试MSE（利用真实曲线模拟测试数据）

LOOCV估计的MSE

10-fold CV估计的MSE

Measuring the Accuracy of the Classifier

- Hold-out test
  - Hold a certain fraction of samples for test.
- Leave one out cross-validation
- K-fold cross-validation
  - 将数据集分为k个子集；
  - 用*k-1*个子集作训练集，1个子集作测试集，然后k次交叉验证；

Measuring the Accuracy of the Classifier

	Real Negative	Real Positive
Claimed Positive	False Positve (FP)	True Positive (TP)
Claimed Negative	True Negative (TN)	False Negative (FN)

Measuring the Accuracy of the Classifier

- Sensitivity (Sn)
- False Positive Rate (FPR)
- Correlation Coefficient (CC)
- Approximate Correlation (AC)

$$Sn = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{TN + FP}$$

$$CC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN)(TN + FP)(TP + FP)(TN + FN)}}$$

$$AC = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TP}{TP + FP} + \frac{TN}{TN + FP} + \frac{TN}{TN + FN}\right) - 1$$

Measuring the Accuracy of the Classifier

$$\text{TPR /Recall / Sensitivity} = \frac{TP}{TP + FN}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

$$\text{FPR} = 1 - \text{Specificity}$$

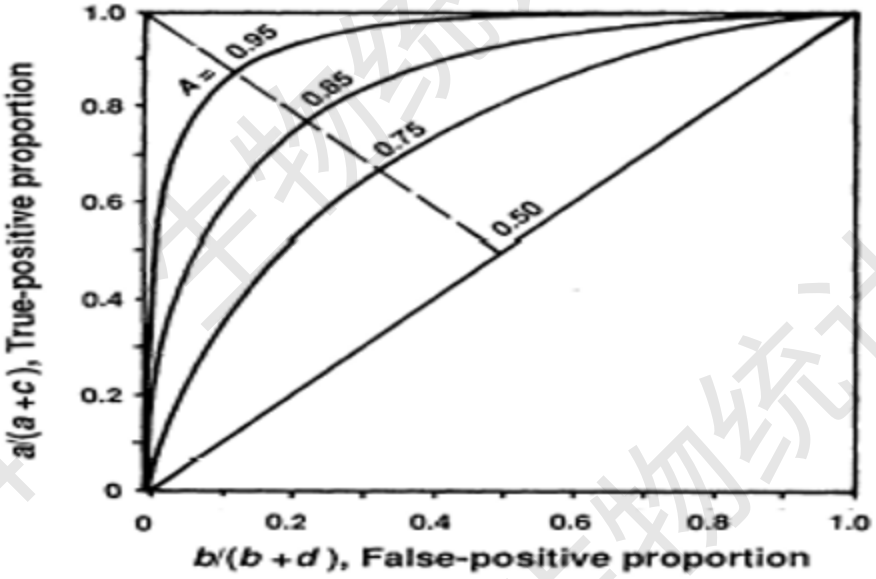
$$= \frac{FP}{TN + FP}$$

Sensitivity ⬆ TPR ⬆
Specificity ⬇ FPR ⬆

Sensitivity ⬇ TPR ⬇
Specificity ⬆ FPR ⬇

Measuring the Accuracy of the Classifier

- ROC: Receiver Operating Characteristic or Relative Operating Characteristic. True positive proportion v.s. False-positive proportion.

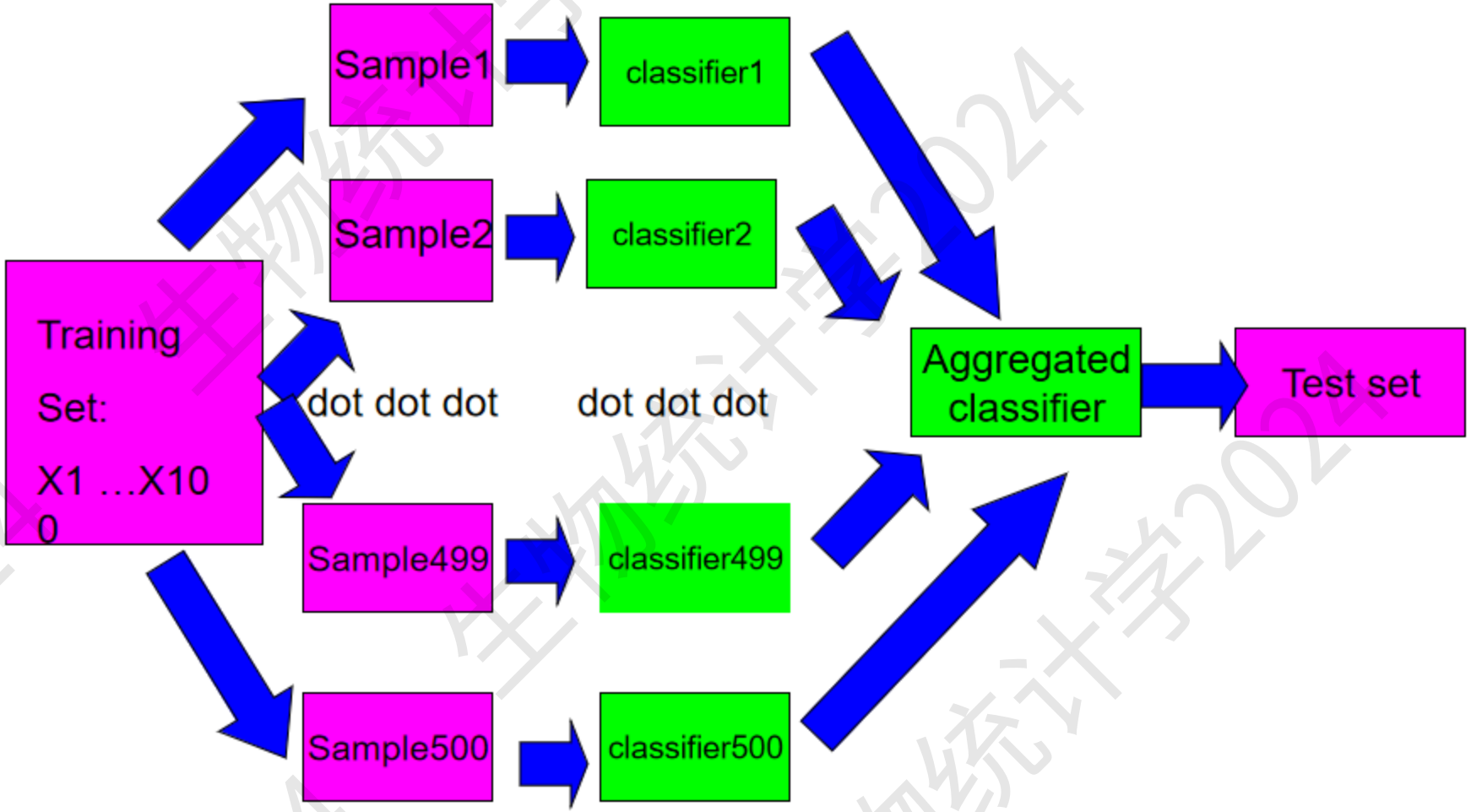


Aggregating classifiers

- Breiman (1996, 1998) found that gains in accuracy could be obtained by aggregating predictors built from perturbed versions of the learning set; the multiple versions of the predictor are aggregated by voting.
- Let $C(., L_b)$ denote the classifier built from the bth perturbed learning set L_b, and let w_b denote the weight given to predictions made by this classifier. The predicted class for an observation x is given by

$$\text{argmax}_k \sum_b w_b I(C(x,L_b) = k)$$

Diagram of aggregating classifiers



Bagging

- Bagging = Bootstrap aggregating
- Non-parametric Bootstrap (standard bagging): perturbed learning sets drawn at random with replacement from the learning sets; predictors built for each perturbed dataset and aggregated by plurality voting ($w_b = 1$)
- Parametric Bootstrap: perturbed learning sets are multivariate Gaussian
- Convex pseudo-data (Breiman 1996)

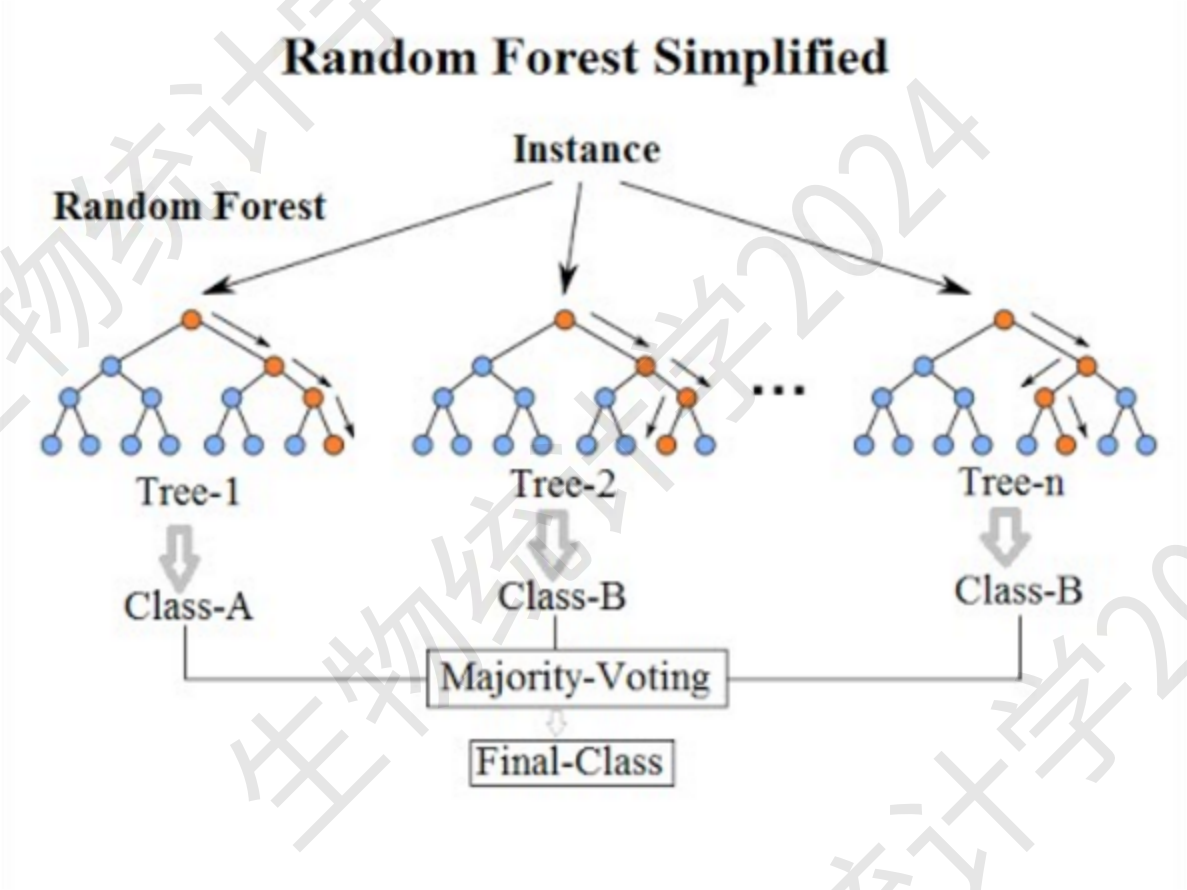
Boosting

- Freund and Schapire (1997), Breiman (1998)
- Data resampled adaptively so that the weights in the resampling are increased for those cases most often misclassified
- Predictor aggregation done by weighted voting

Random Forests

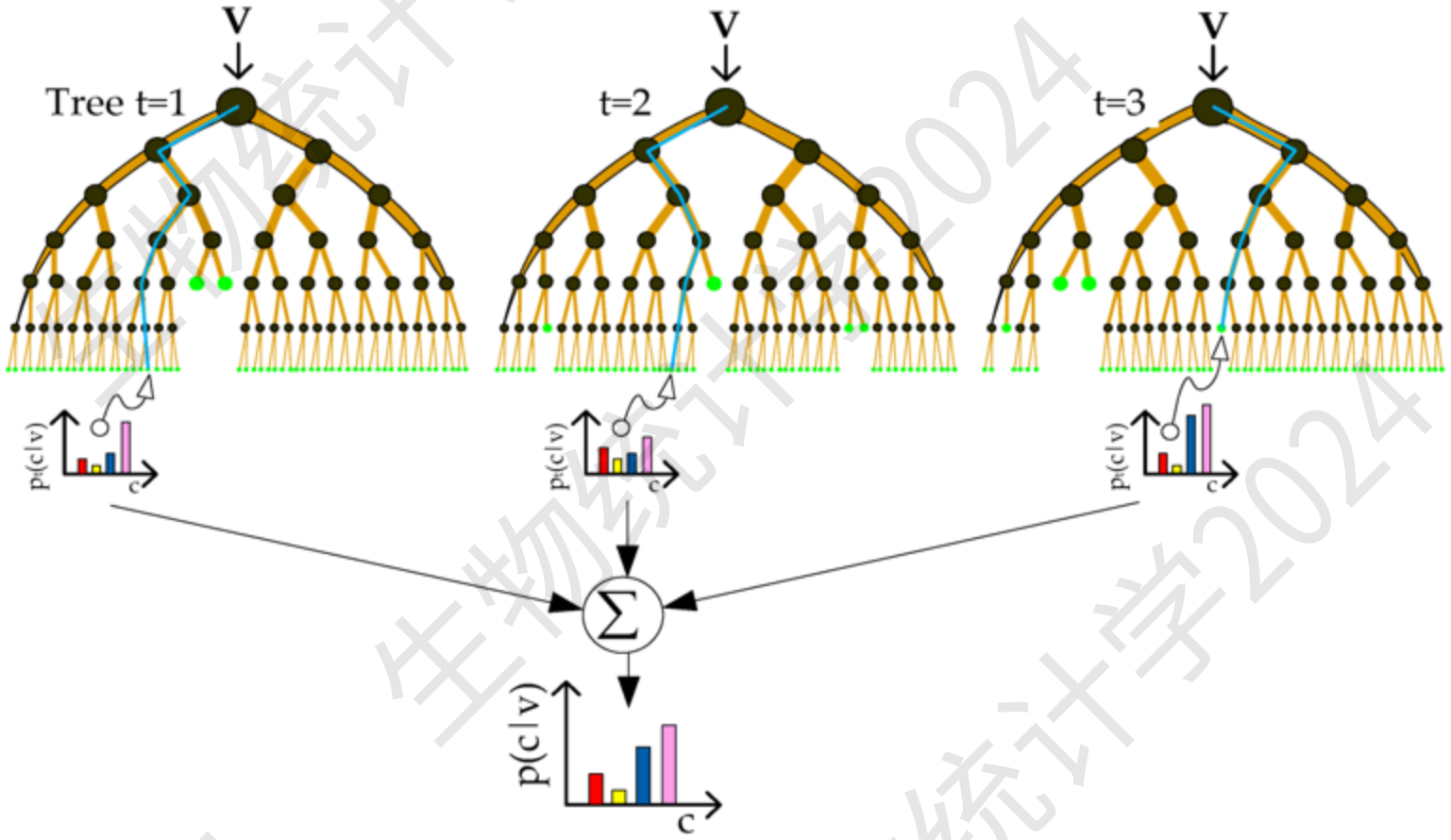
- Perturbed learning sets are drawn at random with replacement from the learning sets
- The exploratory tree is built for each perturbed dataset in a way that at each node, the pre-specified number of features is randomly sub-sampled without replacement and only these variables are used to decide the split at that node.
- The resulting trees are aggregated by plurality voting ($w_b = 1$)

Random Forests

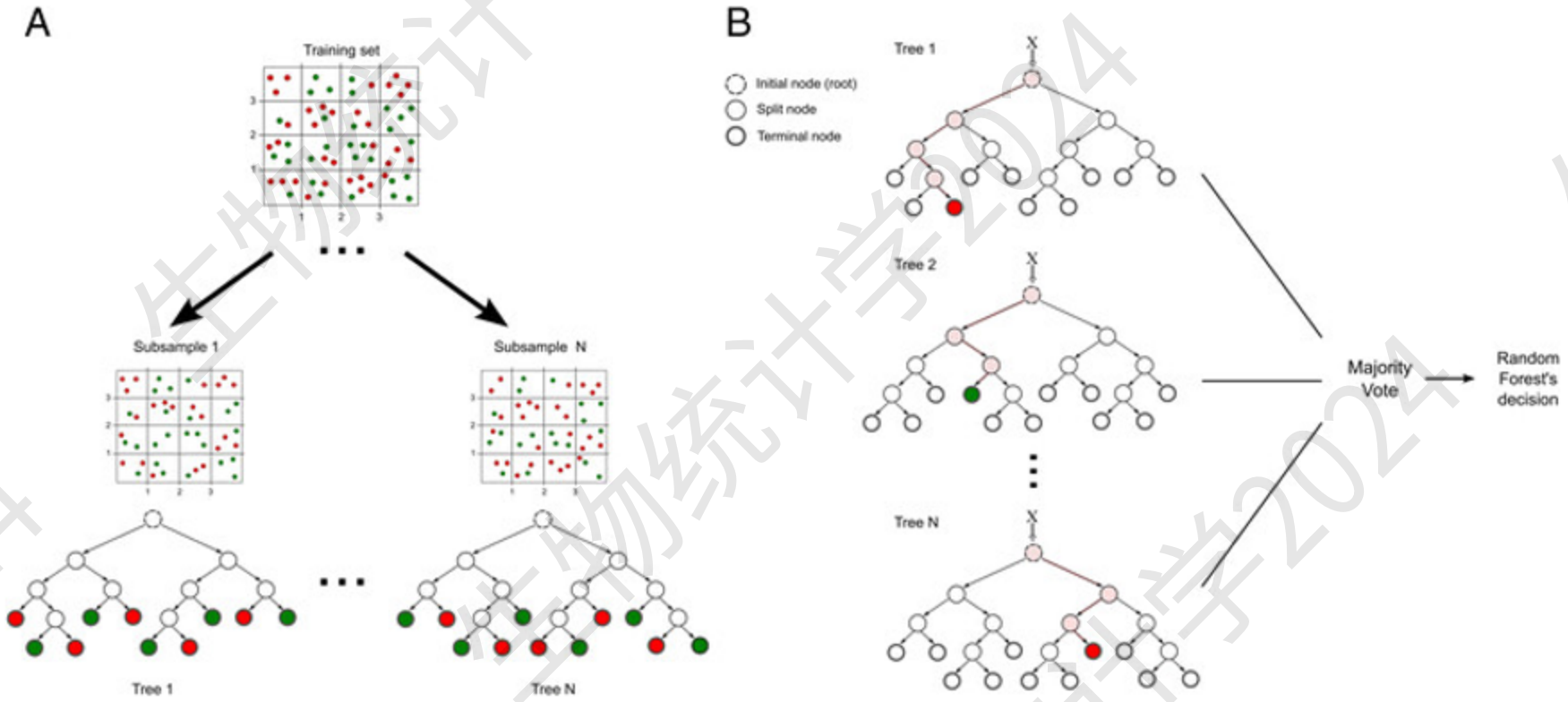


Reference:
https://medium.com/@williamkoehrsen/random-forest-simple-explanation-377895a60d2d

Random Forests



Random Forests

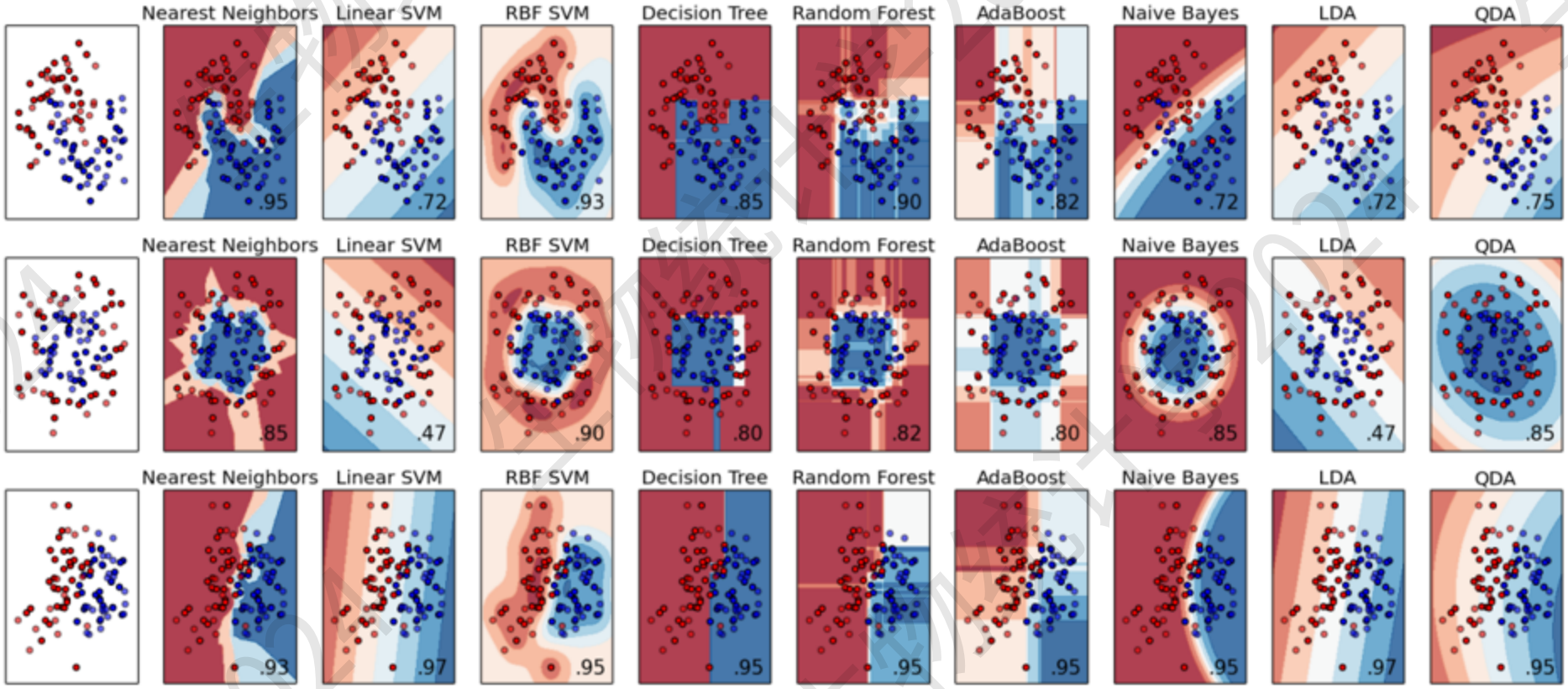


Reference:
https://www.researchgate.net/publication/280533599_What_variables_are_important_in_predicting_bovine_viral_diarrhea_virus_A_random_forest_approach/figures?lo=1&utm_source=google&utm_medium=organic

Random Forests

- It is unexcelled in accuracy among current algorithms (在当前所有算法中，具有极好的准确率)
- It runs efficiently on large dataset (能够有效地运行在大数据集上)
- It can handle thousands of input variables without variable deletion (能够处理具有高维特征的输入样本，而且不需要降维)
- It gives estimates of what variables are important in the classification (能够评估各个特征在分类问题上的重要性)
- It generates an internal unbiased estimate of the generalization error as the forest building progresses (在生成过程中，能够获取到内部生成误差的一种无偏估计)
- It has an effective method for estimating missing data and maintains accuracy when a large proportion of the data are missing (对于缺省值问题也能够获得很好得结果)

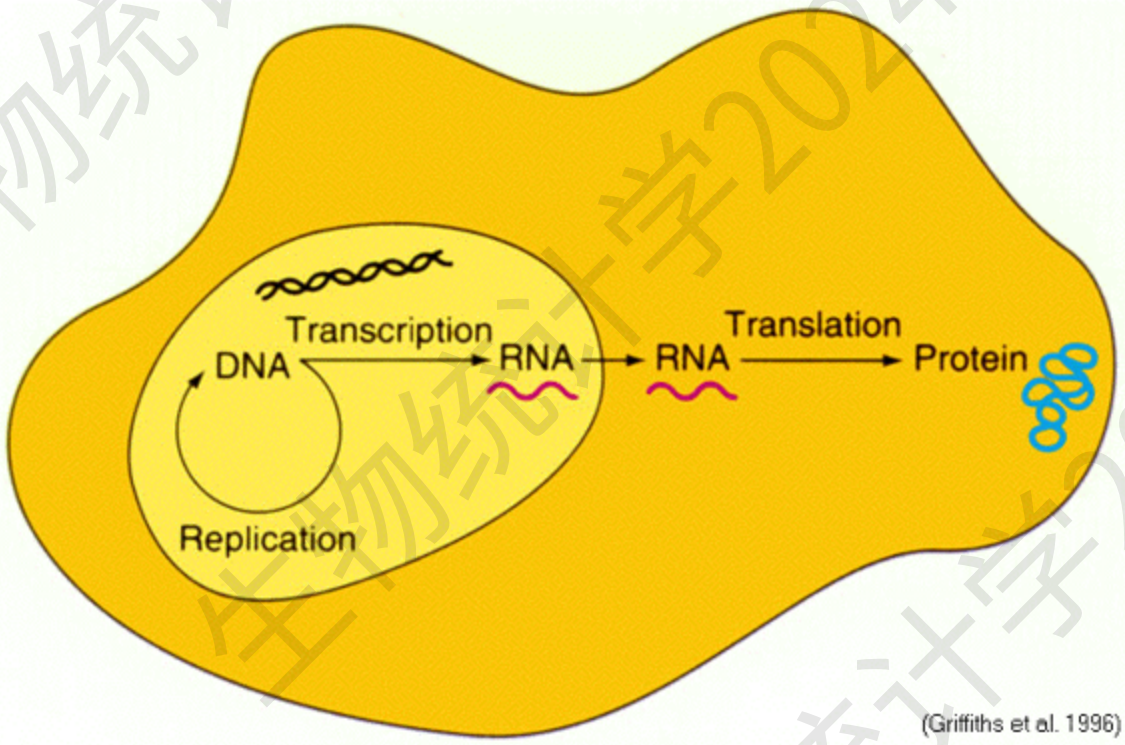
Random Forests



第6-3章: Gene Expression and Expression Measurements

- Introduction to gene expression
- Expression measurements

Transcriptome



(Griffiths et al. 1996)

Gene Expression

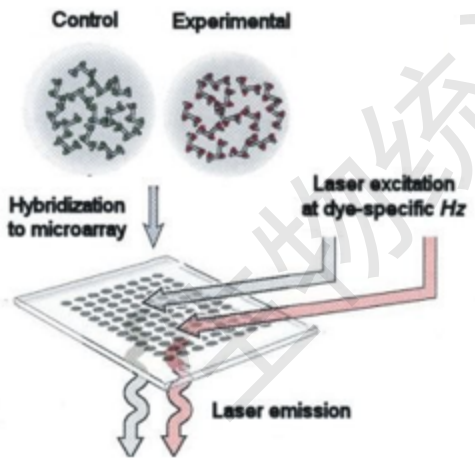
- Each cell contains a complete set of DNA.
- Only a fraction of these are used (or “expressed”) in any particular cell at any given time. For example, genes specific for erythroid cells, such as the hemoglobin genes, are not expressed in brain cells.

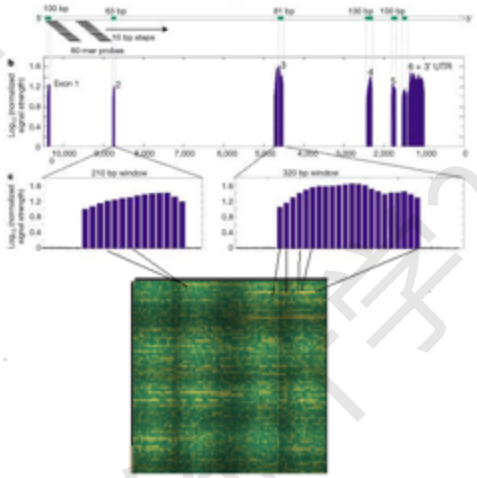
What is a DNA Microarray?

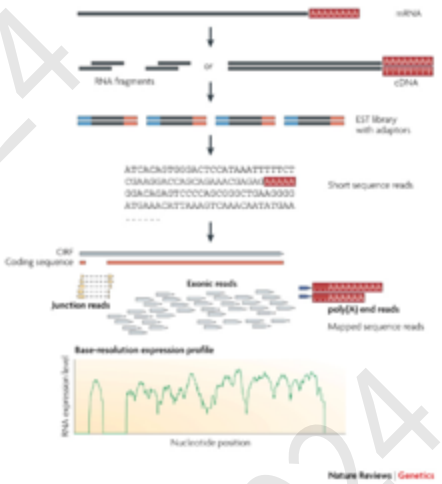
- Also known as DNA Chip
- Allows simultaneous measurement of the level of transcription for every gene in a genome (gene expression)
- Transcription?
  - Process of copying of DNA into messenger RNA (mRNA)
  - Environment dependent!
- Microarray detects mRNA, or rather the more stable cDNA

The Evolution of Transcriptomics

Hybridization-based







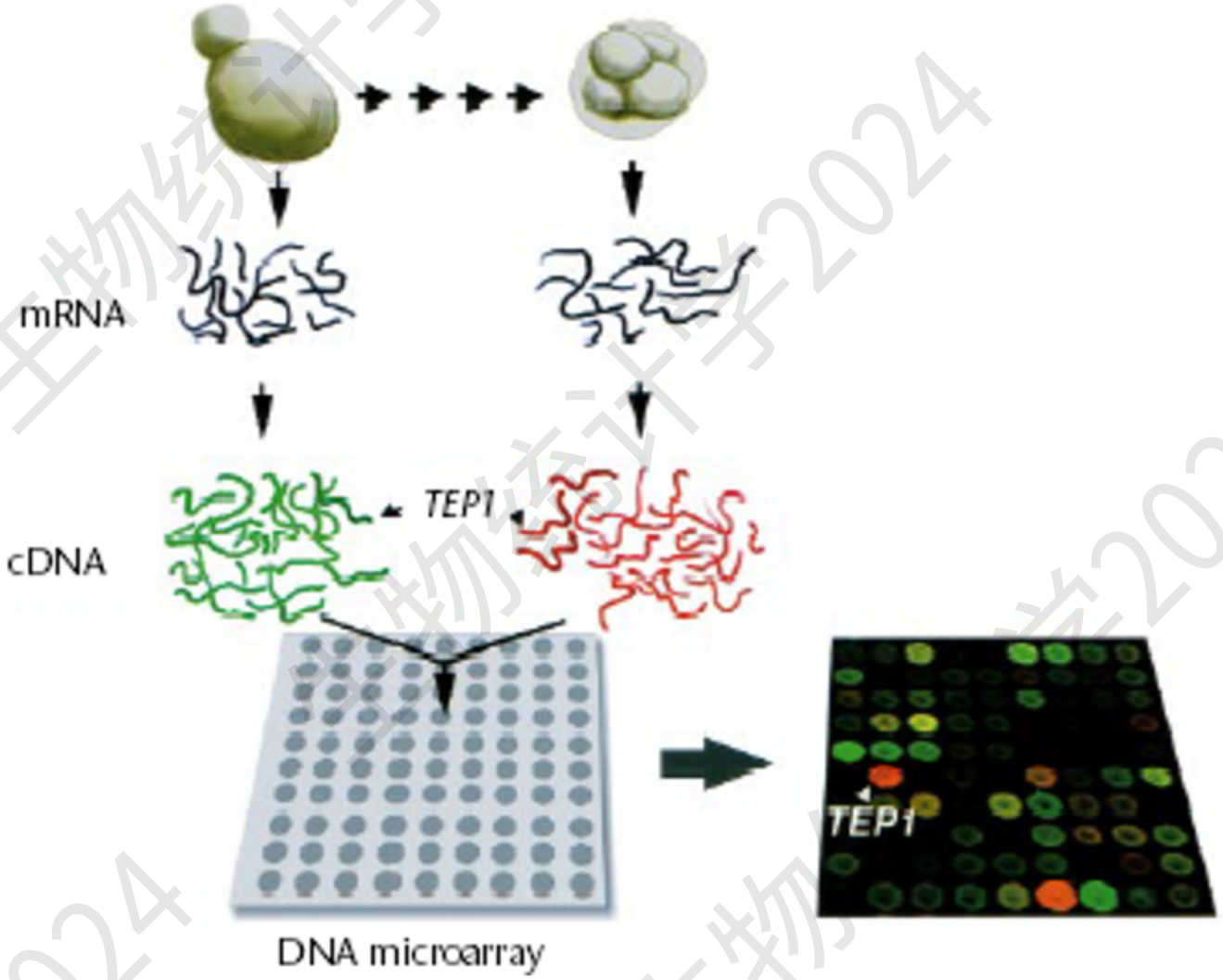
RNA-seq is still a technology under active development

1995 P. Brown, et. al. Gene expression profiling using spotted cDNA microarray: expression levels of known genes

2002 Affymetrix, whole genome expression profiling using tiling array: identifying and profiling novel genes and splicing variants

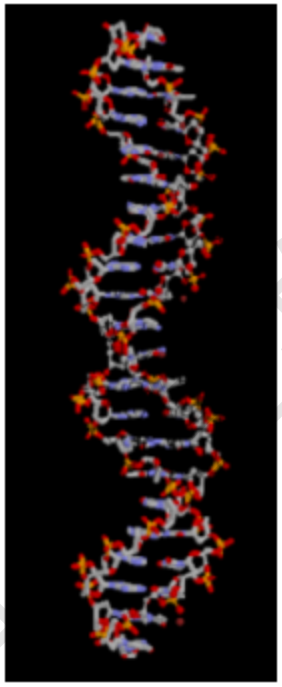
2008 many groups, mRNA-seq: direct sequencing of mRNAs using next generation sequencing techniques (NGS)

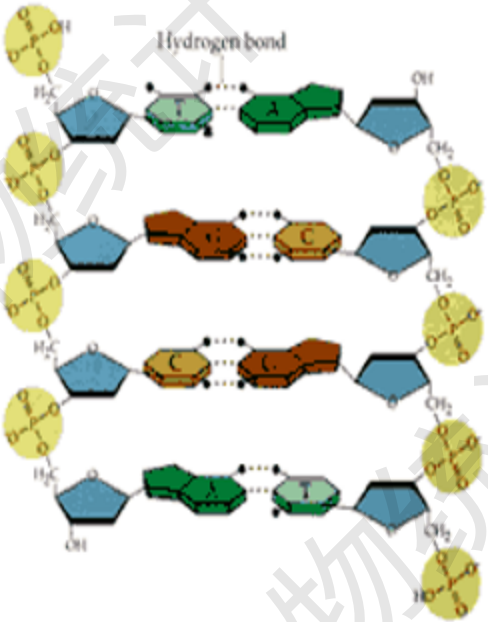
cDNA Microarray

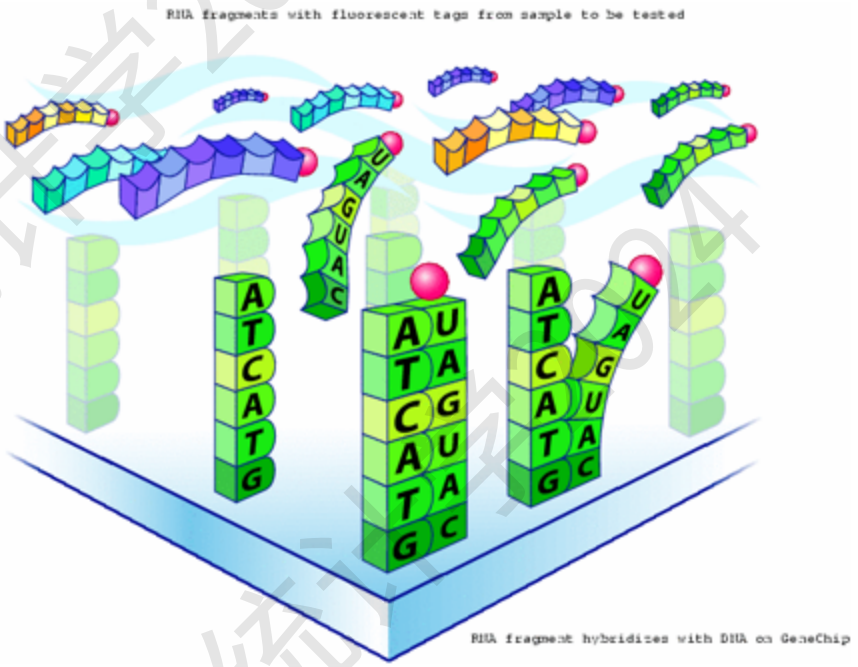


Brown & Botstein, 1999

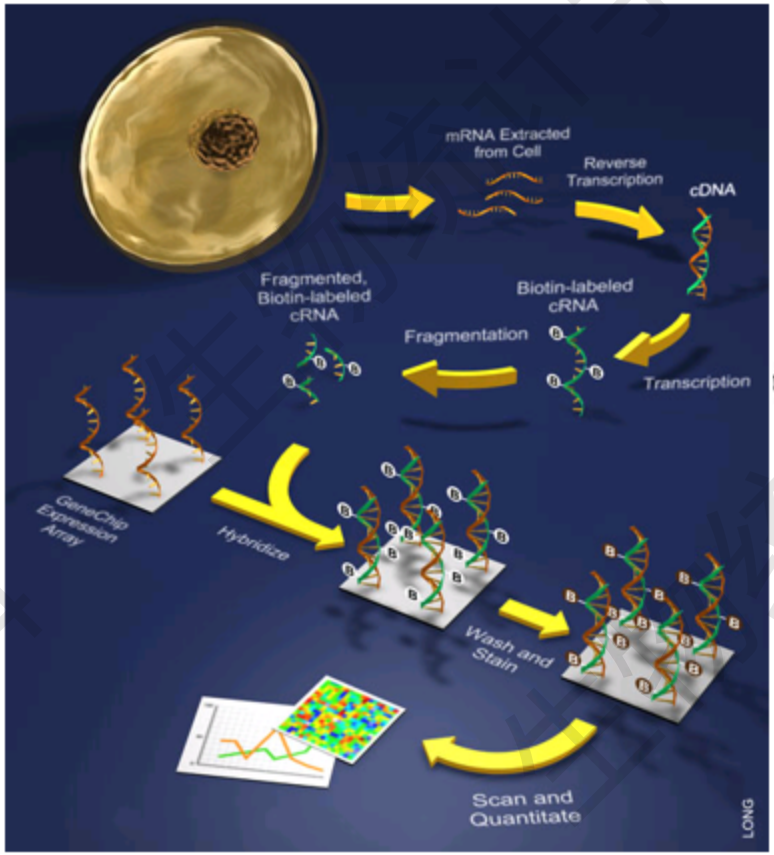
杂交机制(hybridization)->生物芯片

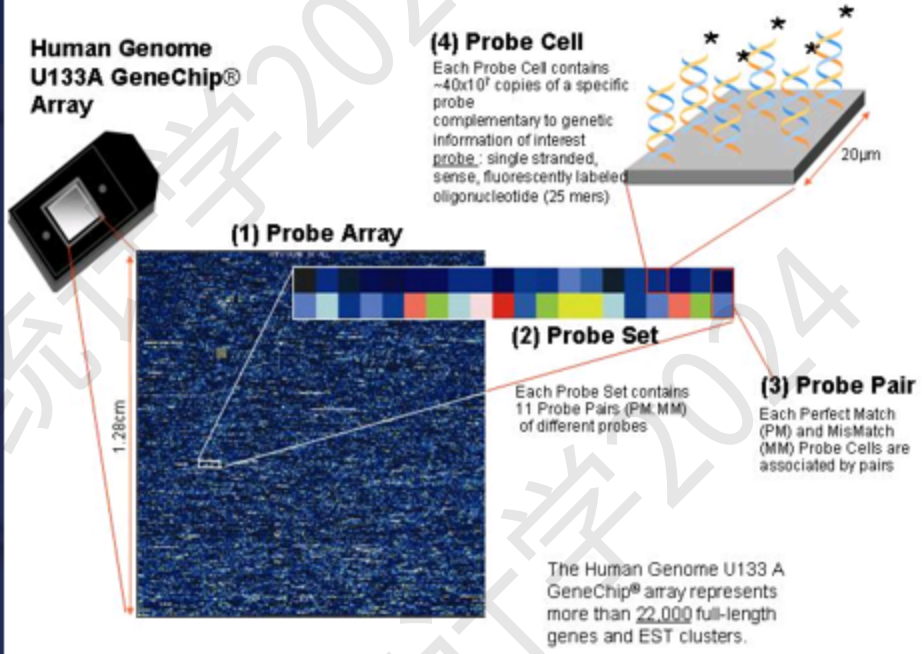




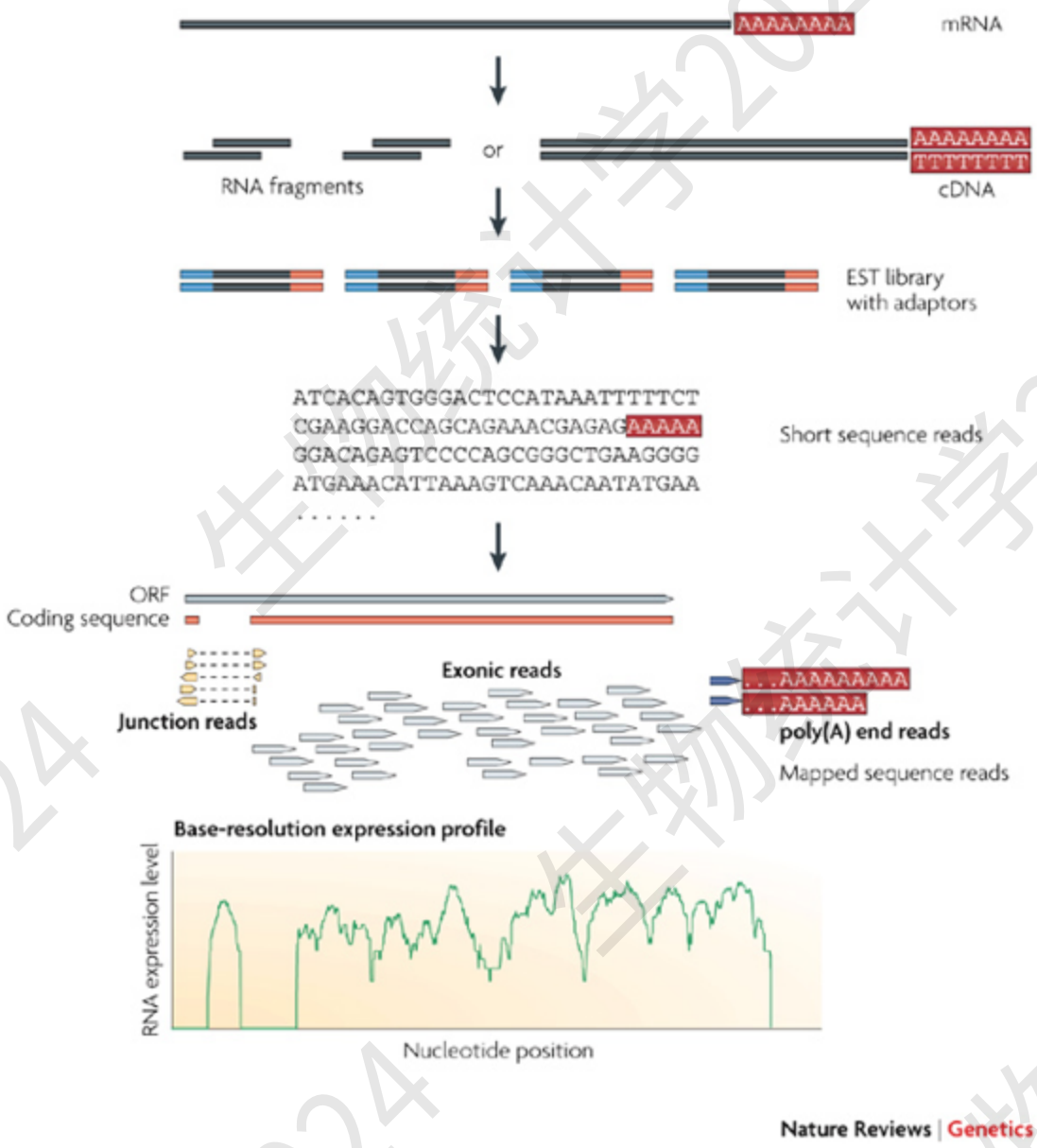


Affymetrix 表达芯片





How RNA-seq works



Sample preparation

Next generation sequencing (NGS)

Data analysis:
- ✓Mapping reads
- ✓Visualization (Gbrowser)
- ✓De novo assembly
- ✓Quantification

Figure from Wang et. al, **RNA-Seq: a revolutionary tool for transcriptomics,** Nat. Rev. Genetics 10, 57-63, 2009).

FPKM (RPKM): Expression Values

- **F**ragments ~~Reads~~ **P**er **K**ilobase of exon model per **M**illion mapped fragments
- Nat Methods. 2008, Mapping and quantifying mammalian transcriptomes by RNA-Seq. Mortazavi A et al.

$$FPKM = 10^9 \times \frac{C}{NL}$$

C= the number of reads mapped onto the gene's exons
N= total number of reads in the experiment
L= the sum of the exons in base pairs.

第6-4章: Class Comparison

1. Statistical test
2. Gene set enrichment analysis (GSEA)

References

- Subramanian et al. PNAS 102:15546, 2005.
- Tian et al. PNAS 102:13544, 2005.
- Mootha et al. Nature Genetics 2003

Class comparison

- What genes are up regulated between control and test or multiple test conditions
  - Normal v tumor
  - Treated v untreated
- Fold change
  - Not sufficient, need statistics
- Statistics
  - t test, non-parametric, fdr
- Depends on underlying assumptions about data

Class Comparison

- What genes are up regulated between control and test or multiple test conditions
- Many analysis methods
  - May produce different results
  - Different underlying statistics and methods
    - t test
    - SAM
    - Non parametric (relative entropy)
    - Emperical bayesian
- Depends on underlying assumptions about data

Measuring the Accuracy of the Classifier

	Real Negative	Real Positive
Claimed Positive	False Positve (FP)	True Positive (TP)
Claimed Negative	True Negative (TN)	False Negative (FN)

Measuring the Accuracy of the Classifier

- Sensitivity (Sn)
- False Positive Rate (FPR)
- Correlation Coefficient (CC)
- Approximate Correlation (AC)

$$Sn = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{TN + FP}$$

$$CC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN)(TN + FP)(TP + FP)(TN + FN)}}$$

$$AC = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TP}{TP + FP} + \frac{TN}{TN + FP} + \frac{TN}{TN + FN}\right) - 1$$

Measuring the Accuracy of the Classifier

$$\text{TPR /Recall / Sensitivity} = \frac{TP}{TP + FN}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

$$\text{FPR} = 1 - \text{Specificity}$$

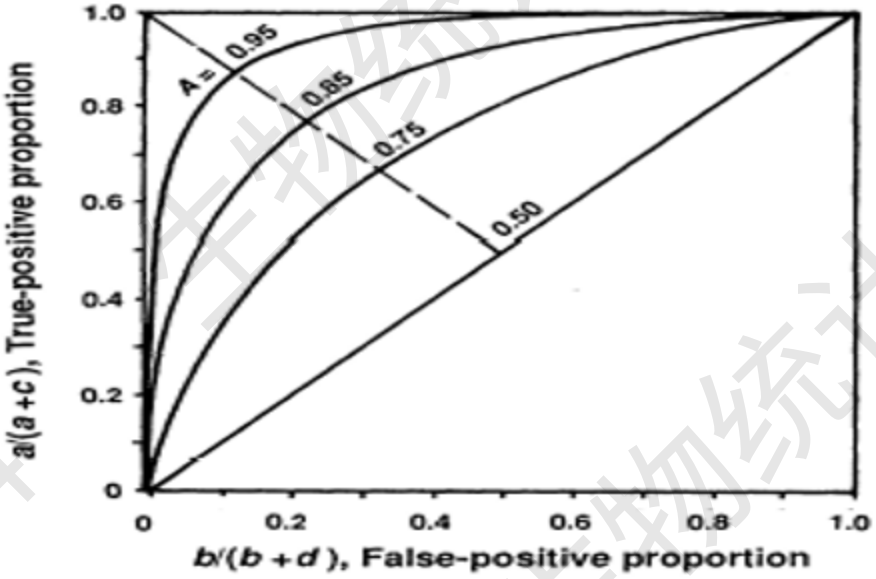
$$= \frac{FP}{TN + FP}$$

Sensitivity ⬆ TPR ⬆
Specificity ⬇ FPR ⬆

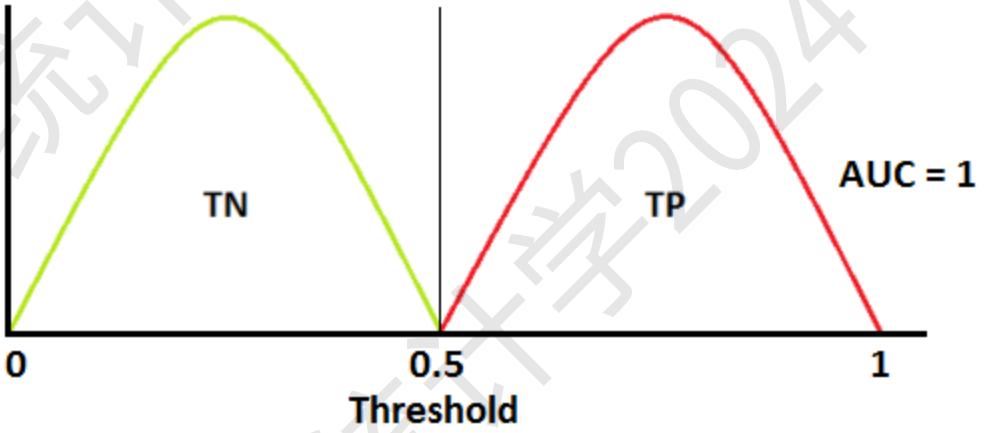
Sensitivity ⬇ TPR ⬇
Specificity ⬆ FPR ⬇

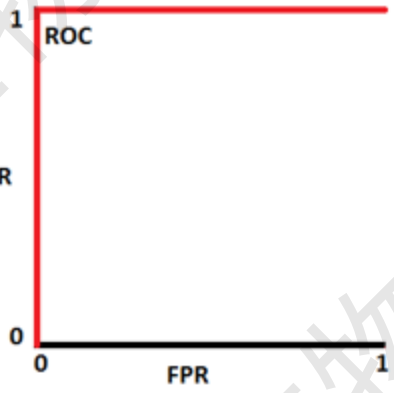
Measuring the Accuracy of the Classifier

- ROC: Receiver Operating Characteristic or Relative Operating Characteristic. True positive proportion v.s. False-positive proportion.

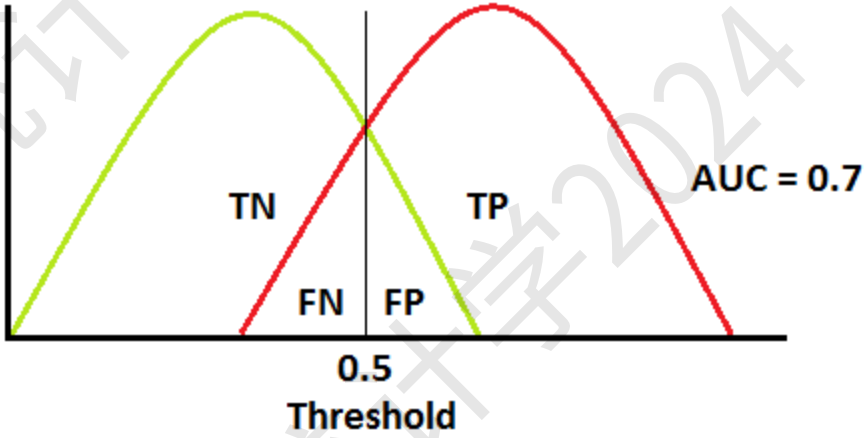


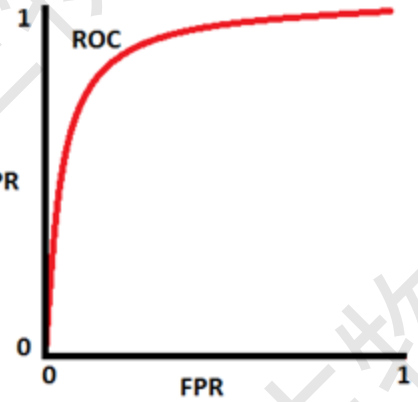
Measuring the Accuracy of the Classifier



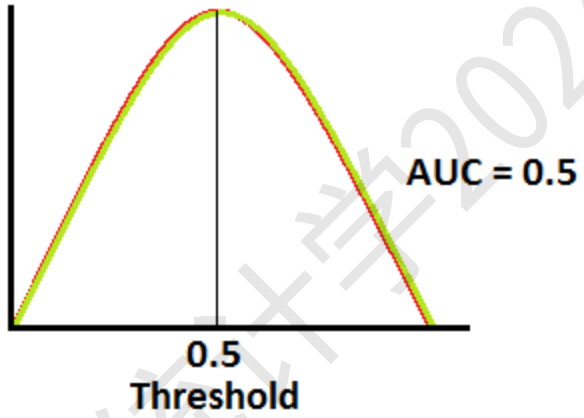


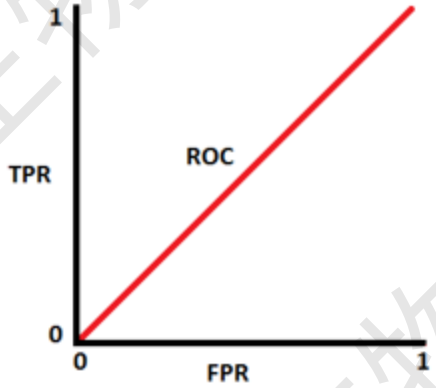
Measuring the Accuracy of the Classifier



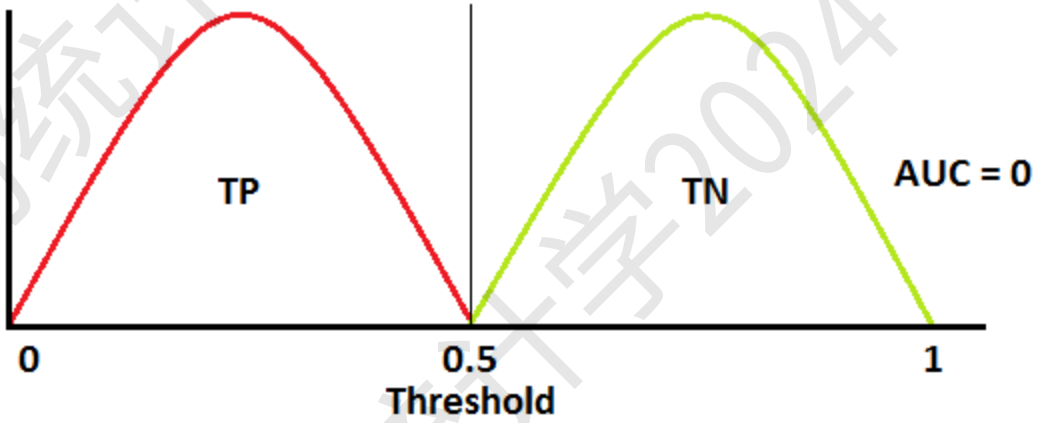


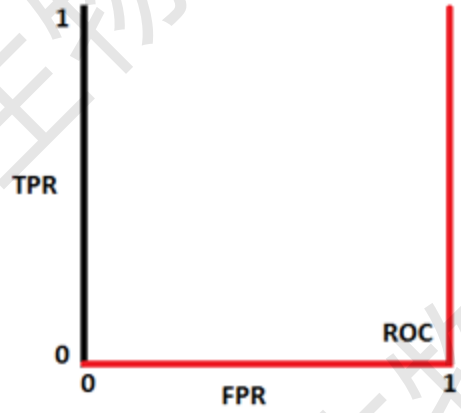
Measuring the Accuracy of the Classifier





Measuring the Accuracy of the Classifier





Measuring the Accuracy of the Classifier

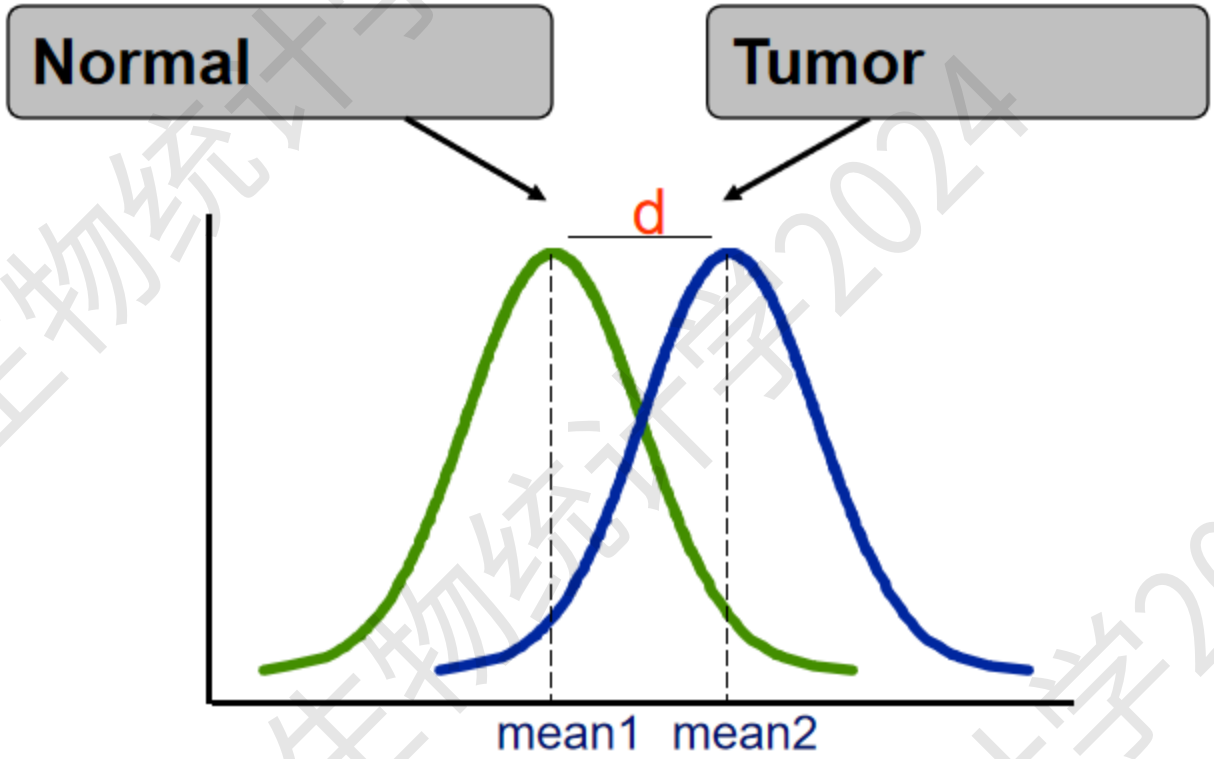
Sources: [15][16][17][18][19][20][21][22] view • talk • edit

		Predicted condition			
	Total population = P + N	**Positive (PP)**	**Negative (PN)**	Informedness, bookmaker informedness (BM) = TPR + TNR − 1	Prevalence threshold (PT) = $\frac{\sqrt{TPR \times FPR} - FPR}{TPR - FPR}$
Actual condition	**Positive (P)**	**True positive (TP), hit**	**False negative (FN), type II error, miss, underestimation**	True positive rate (TPR), recall, sensitivity (SEN), probability of detection, hit rate, power = $\frac{TP}{P}$ = 1 − FNR	False negative rate (FNR), miss rate = $\frac{FN}{P}$ = 1 − TPR
	Negative (N)	**False positive (FP), type I error, false alarm, overestimation**	**True negative (TN), correct rejection**	False positive rate (FPR), probability of false alarm, fall-out = $\frac{FP}{N}$ = 1 − TNR	True negative rate (TNR), specificity (SPC), selectivity = $\frac{TN}{N}$ = 1 − FPR
	Prevalence = $\frac{P}{P+N}$	Positive predictive value (PPV), precision = $\frac{TP}{PP}$ = 1 − FDR	False omission rate (FOR) = $\frac{FN}{PN}$ = 1 − NPV	Positive likelihood ratio (LR+) = $\frac{TPR}{FPR}$	Negative likelihood ratio (LR−) = $\frac{FNR}{TNR}$
	Accuracy (ACC) = $\frac{TP+TN}{P+N}$	False discovery rate (FDR) = $\frac{FP}{PP}$ = 1 − PPV	Negative predictive value (NPV) = $\frac{TN}{PN}$ = 1 − FOR	Markedness (MK), deltaP (Δp) = PPV + NPV − 1	Diagnostic odds ratio (DOR) = $\frac{LR+}{LR-}$
	Balanced accuracy (BA) = $\frac{TPR+TNR}{2}$	F_1 score = $\frac{2PPV \times TPR}{PPV + TPR}$ = $\frac{2TP}{2TP + FP + FN}$	Fowlkes–Mallows index (FM) = $\sqrt{PPV \times TPR}$	Matthews correlation coefficient (MCC) = $\sqrt{TPR \times TNR \times PPV \times NPV} - \sqrt{FNR \times FPR \times FOR \times FDR}$	Threat score (TS), critical success index (CSI), Jaccard index = $\frac{TP}{TP + FN + FP}$

Measuring the Accuracy of the Classifier

A	B	C	C′
TP=63 FN=37 100 FP=28 TN=72 100 91 109 200	TP=77 FN=23 100 FP=77 TN=23 100 154 46 200	TP=24 FN=76 100 FP=88 TN=12 100 112 88 200	TP=76 FN=24 100 FP=12 TN=88 100 88 112 200
TPR = 0.63	TPR = 0.77	TPR = 0.24	TPR = 0.76
FPR = 0.28	FPR = 0.77	FPR = 0.88	FPR = 0.12
PPV = 0.69	PPV = 0.50	PPV = 0.21	PPV = 0.86
F1 = 0.66	F1 = 0.61	F1 = 0.23	F1 = 0.81
ACC = 0.68	ACC = 0.50	ACC = 0.18	ACC = 0.82

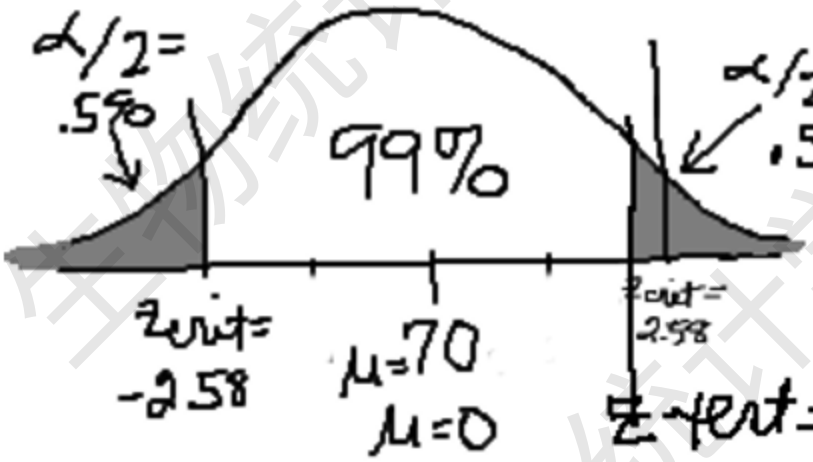
Hypothesis Testing



Null hypothesis $H_0 : \mu_1 = \mu_2$

Alternative hypotheses $H_1 : \mu_1 \neq \mu_2$

Hypothesis Testing

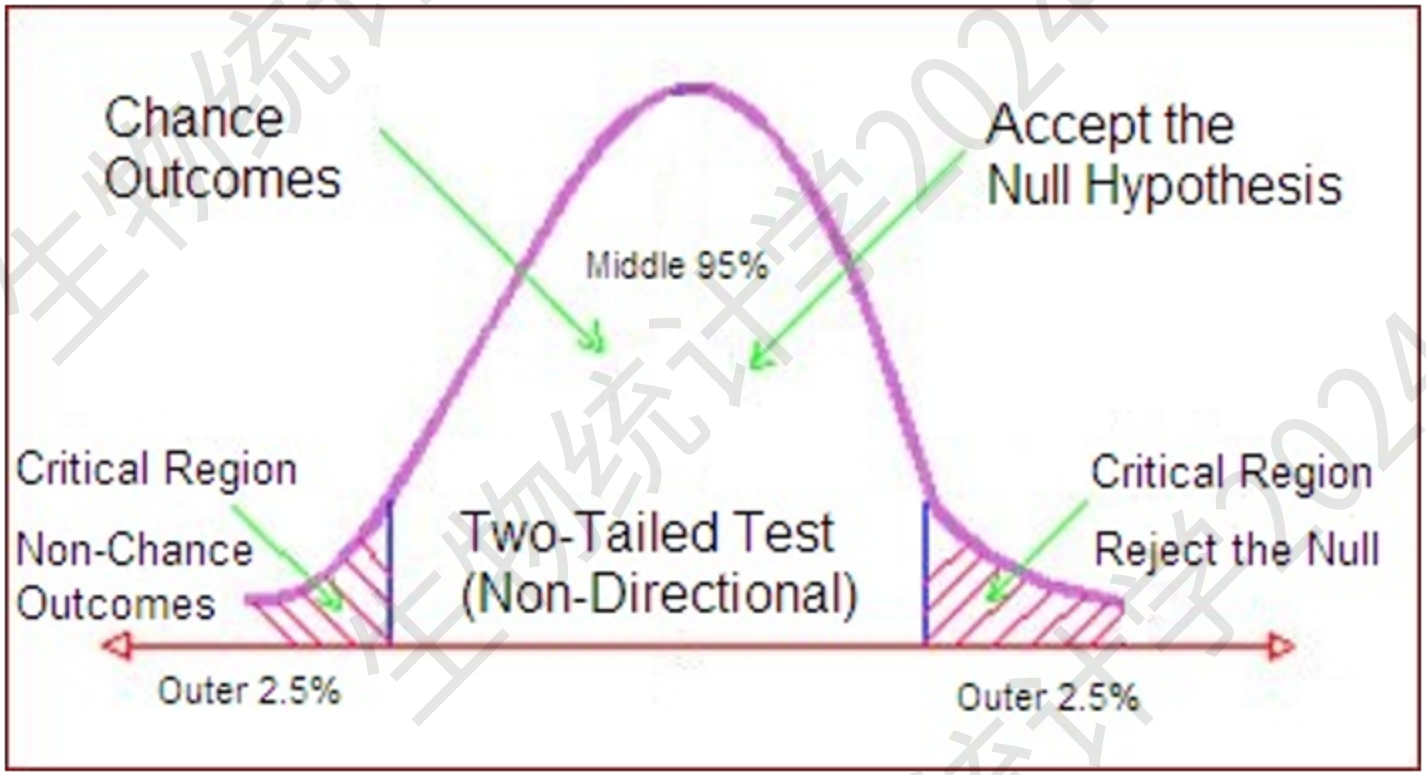


$H_0: \mu = 70$

$H_1: \mu \neq 70$

$$Z_{test} = \frac{\bar{x} - M}{\frac{\sigma}{\sqrt{n}}} = \frac{75 - 70}{\frac{12}{\sqrt{36}}} = \frac{5}{2} = 2.5$$

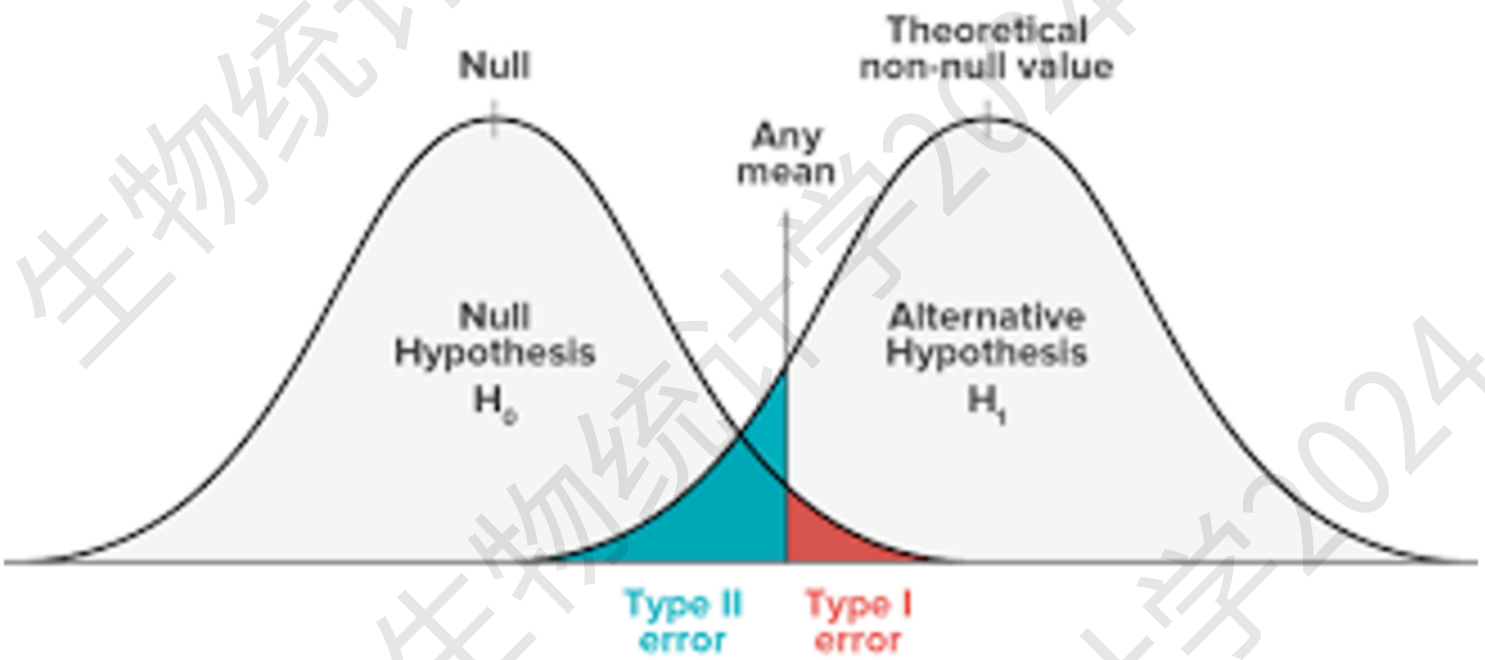
Hypothesis Testing



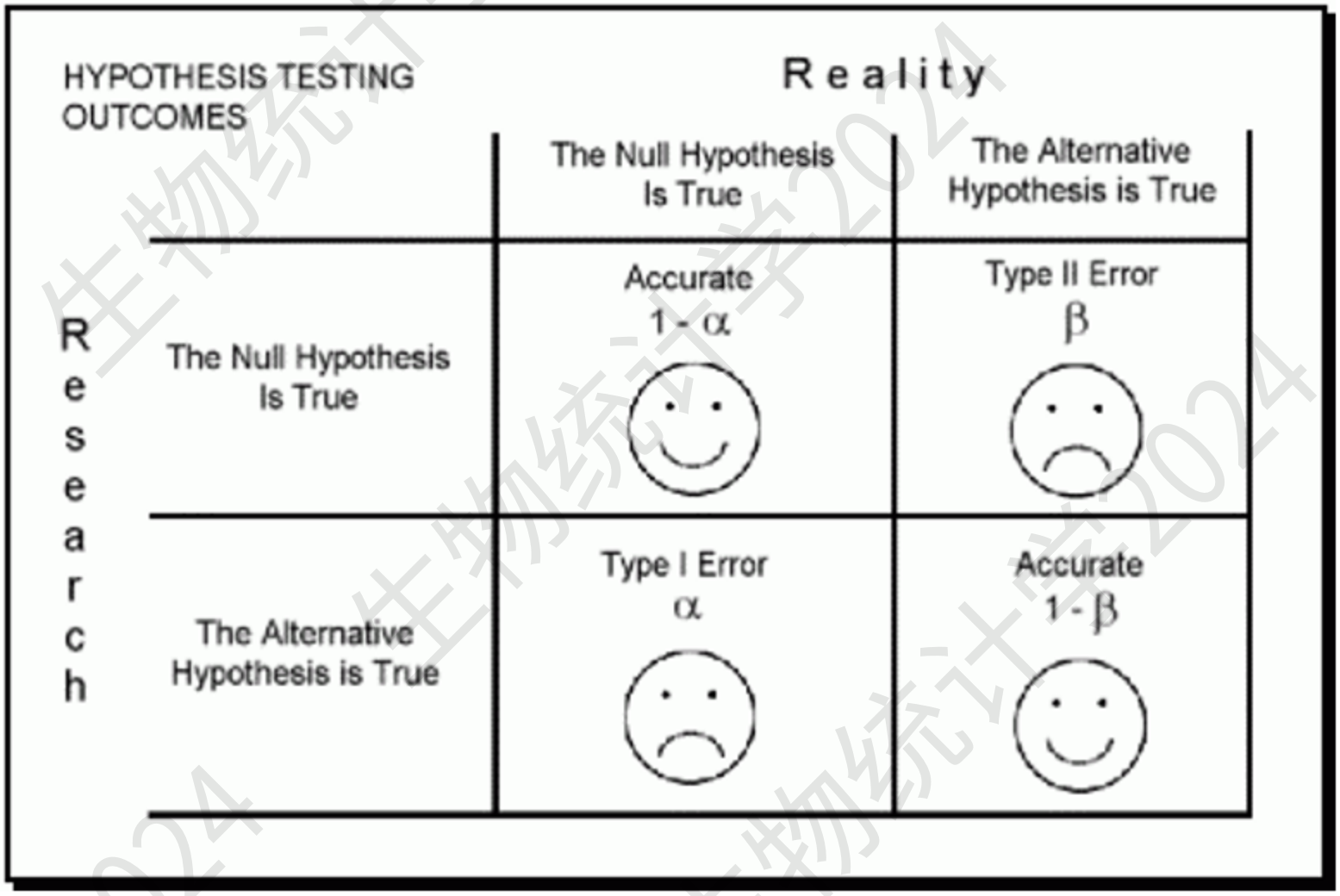
Type I and Type II Error

	Retain Null	Reject Null
H_0	√	type I error
H_1	type II error	√

Type I and Type II Error



Type I and Type II Error



Two-Sample *t*-Statistic

- Student's *t*-statistic

$$T_j = \frac{\bar{\mathbf{y}}_{1j} - \bar{\mathbf{y}}_{2j}}{\sqrt{s_{1j}^2 / n_1 + s_{2j}^2 / n_2}}$$

- Normal assumption

Multiple Test Problem

- Perform a test for each gene to determine the statistical significance of differential expression for that gene.

- Problem: When many hypotheses are tested, the probability of a type I error (false positive) increases sharply with the number of hypotheses.

Example

- Suppose we measure the expression of 10,000 genes in a microarray experiment.
- p <= 0.05 says that 95% confidence means are different; therefore 5% due to chance
- 500 genes are picked up by chance
- Suppose t tests selects 1000 genes at a p of 0.05
- 500/1000 ;Approximately 50% of the genes will be false, very high false discovery rate; need more confidence

Corrections for Multiple Comparisons

- Involve corrections to the p-value so that the actual p-value is higher
- Bonferroni correction
- Benjamin-Hochberg procedure
- Significance Analysis of Microarrays
  - Tusher et al. at Stanford

The Bonferroni Method

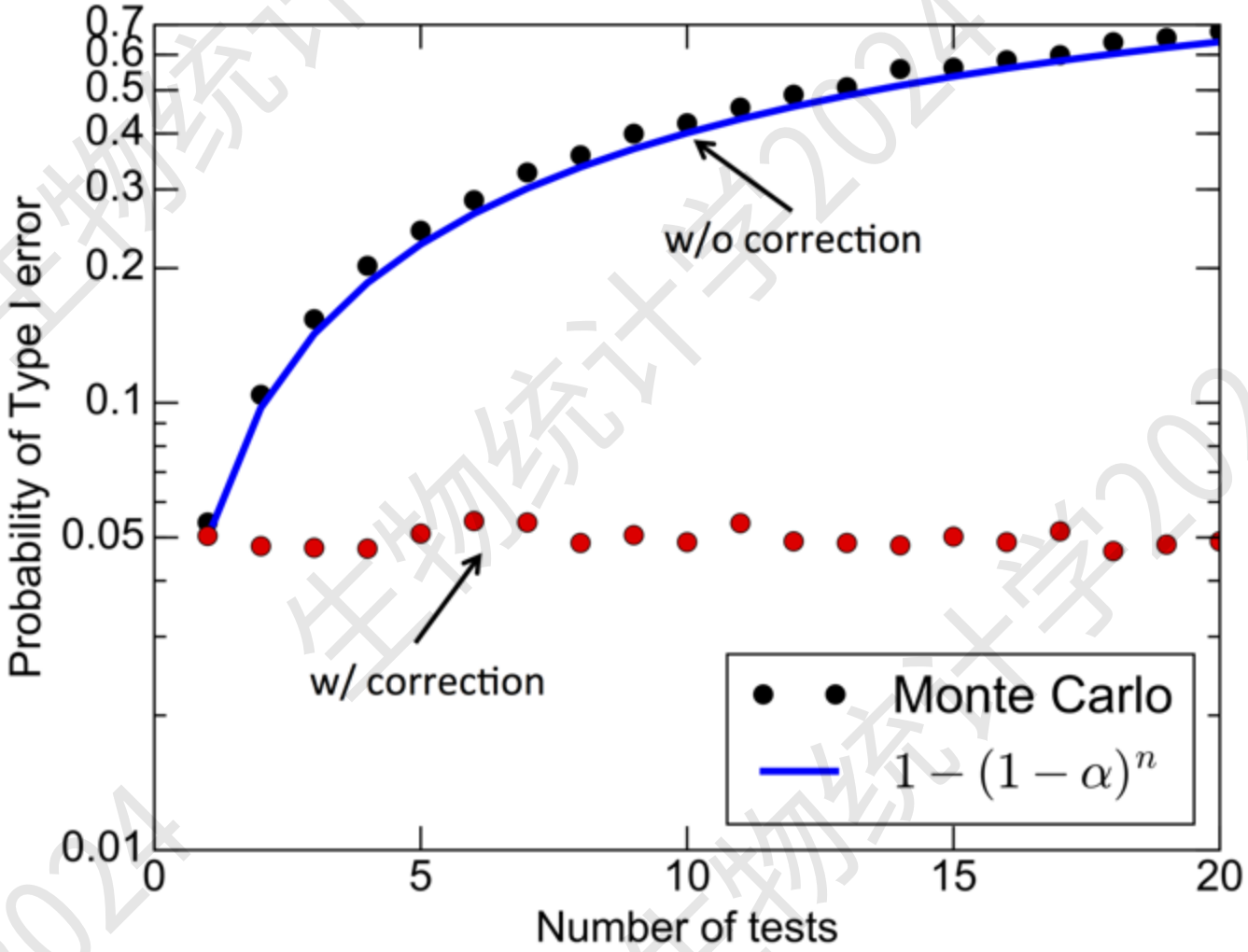
- Controls the family wise error rate (FWER) FWER is the probability that at least one false positive error will be made.

- But this method is very conservative, as it tries to make it unlikely that even one false rejection is made.

The Bonferroni Method

- 如要在同一数据集上检验两个独立的假设，显著水平设为常见的0.05。
- 用于检验该两个假设应使用更严格的0.025。即0.05* (1/2)。
- 该方法是由Carlo Emilio Bonferroni发展的，因此称Bonferroni校正。

The Bonferroni Method



False Discovery Rate (FDR)

- The FDR is essentially the expectation of the proportion of false positives among the identified differentially expressed genes.

$$\text{FDR} \approx \frac{\#(\text{False Positives})}{\#(\text{Rejected Hypotheses})}$$

Measuring the Accuracy

	Real Negative	Real Positive
Claimed Positive	False Positve (FP)	True Positive (TP)
Claimed Negative	True Negative (TN)	False Negative (FN)

False Discovery Rate

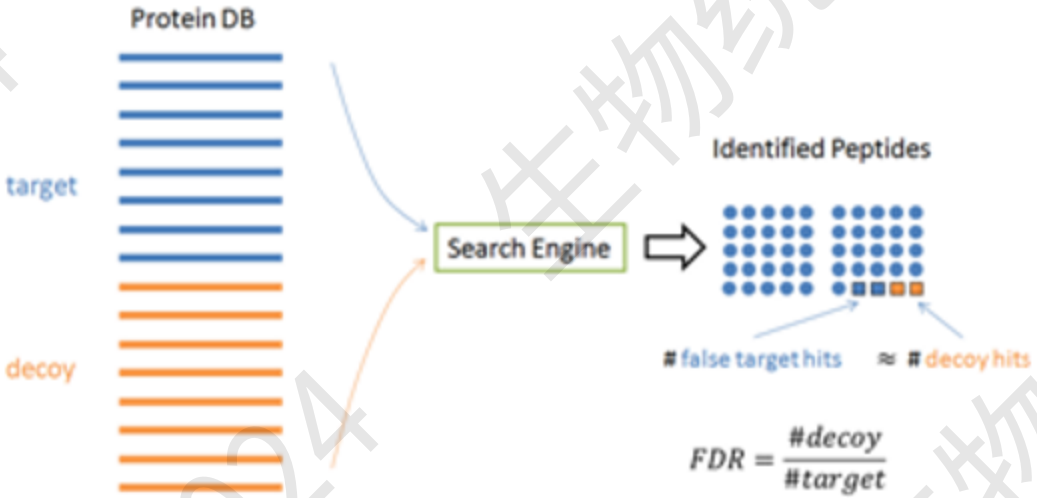
	Accept Null	Reject Null	Total
Null True	N_{00}	N_{01}	N_0
Non-True	N_{10}	N_{11}	N_1
Total	$N - N_r$	N_r	N

$$FDR \approx \frac{N_{01}}{N_r}$$

FDR（false discovery rate）

- FDR（错误发现率）错误控制法是Benjamini于1995年提出的一种方法，基本原理是通过控制FDR值来决定P值的阈值。相对Bonferroni来说，FDR用比较温和的方法对p值进行了校正。

- 试图在假阳性和假阴性间达到平衡，将假/真阳性比例控制到一定范围之内。例如，如果检验1000次，我们设定的阈值为0.05（5%），那么无论我们得到多少个差异蛋白，这些差异蛋白中出现假阳性的概率保持在5%之内，即FDR＜5%。



较适合于基因组学问题：我们可以容忍一定数量的假阳性基因，但希望它在阳性基因中的比例较低

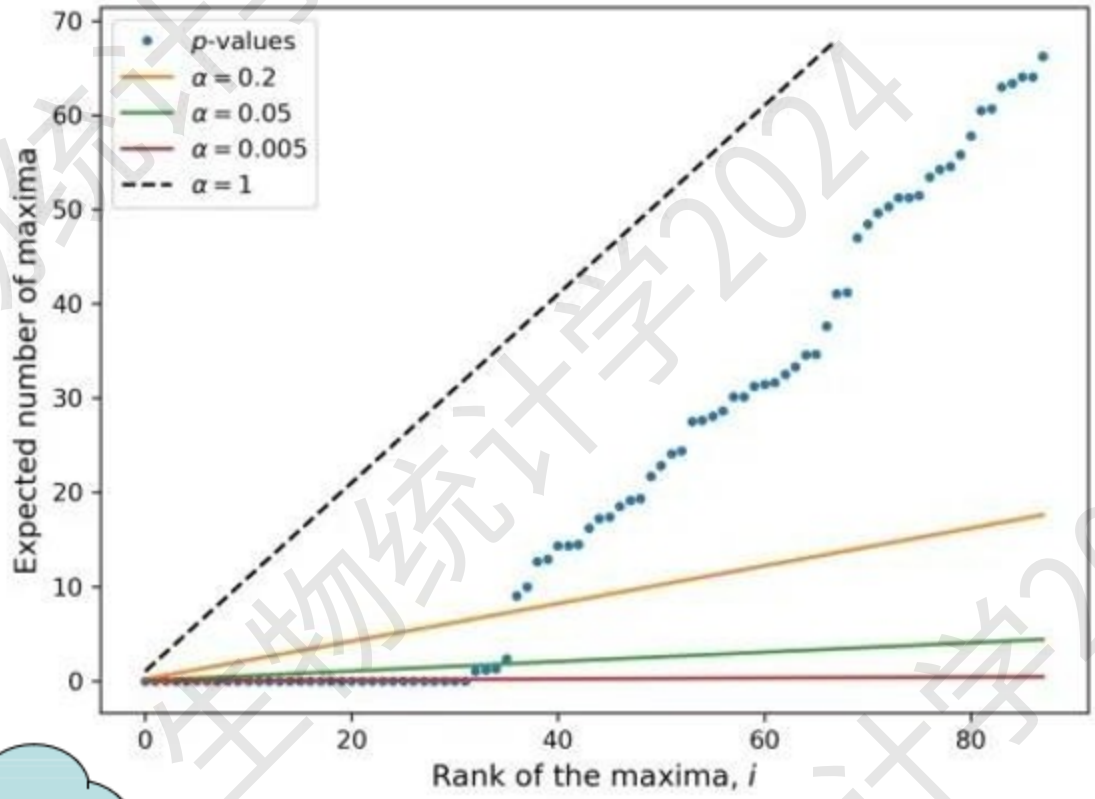
FDR计算

1. 将一系列p值、校正方法（BH）以及所有p值的个数（length(p)）输入到R语言p.adjust函数中。
2. 将一系列的p值按照从大到小排序，然后利用下述公式计算每个p值所对应的FDR值。
   公式：p * (n/i)， p是每一个检验的p-value，n是检验的个数，i是排序后的位置（最大的p值的i值为n，第二大则是n-1，依次至最小为1）。
3. 将计算出来的FDR值赋予给排序后的p值，如果某一个p值所对应的FDR值大于前一位p值（排序的前一位）所对应的FDR值，则放弃公式计算出来的FDR值，选用与它前一位相同的值。因此会产生连续相同FDR值的现象；反之则保留计算的FDR值。
4. 将FDR值按照最初始的p值的顺序进行重新排序，返回结果。

```
p.adjust(p, method = p.adjust.methods, n = length(p))
p.adjust.methods
# c("holm", "hochberg", "hommel", "bonferroni", "BH", "BY",  # "fdr", "none")
```

https://stat.ethz.ch/R-manual/R-devel/library/stats/html/p.adjust.html

FDR阈值



p * n / i < 0.05?
即
p * n < 0.05 * i?

OPEN ACCESS Freely available online

PLOS | COMPUTATIONAL BIOLOGY

Editorial

Ten Simple Rules for Reproducible Computational Research

Geir Kjetil Sandve[1,2*], Anton Nekrutenko[3], James Taylor[4], Eivind Hovig[1,5,6]

1. For Every Result, Keep Track of How It Was Produced
2. Avoid Manual Data Manipulation Steps
3. Archive the Exact Versions of All External Programs Used
4. Version Control All Custom Scripts
5. Record All Intermediate Results, When Possible in Standardized Formats
6. For Analyses That Include Randomness, Note Underlying Random Seeds
7. Always Store Raw Data behind Plots
8. Generate Hierarchical Analysis Output, Allowing Layers of Increasing Detail to Be Inspected
9. Connect Textual Statements to Underlying Results
10. Provide Public Access to Scripts, Runs, and Results

Gene Set Enrichment Analysis Motivation

- **Gene Set Enrichment Analysis (GSEA)** is a computational method that determines whether an a priori defined set of genes shows statistically significant, concordant differences between two biological states (e.g. phenotypes)

- Interpreting the results to gain insights into biological mechanisms remains a major challenge

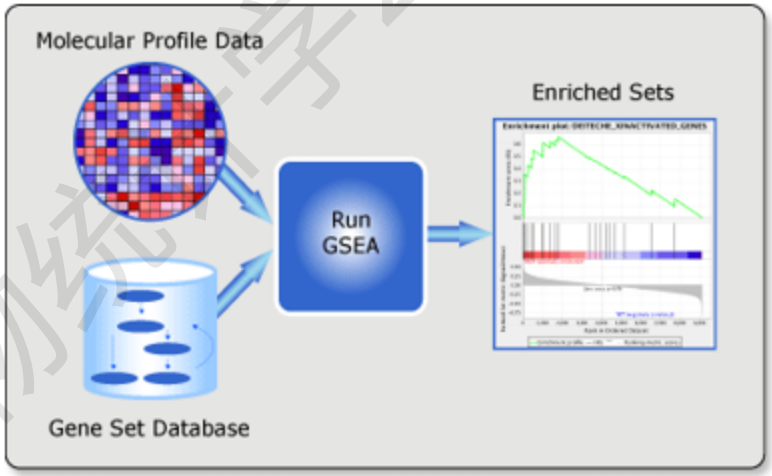
- For a typical study (e.g., experimental condition vs. control, disease state vs. normal, tumor type A vs. tumor type B), a standard approach has been to produce a list of differentially expressed genes (DEGs)

Challenges in Interpreting Gene Microarray Data

- May obtain a long list of statistically significant genes without any obvious unifying biological theme;
- Even with DEG list(s) of up and/or down-regulated genes, still need to accurately extract valid biological inferences. Cutoff for inclusion in DEG lists is somewhat arbitrary. Must address multiple hypothesis testing.

An Existing Way to Study Enrichment of Gene Categories

- Statistical procedures such as ***Fisher's exact test*** based on the hypergeometric distribution are used to test if members of a list of differentially expressed genes are overrepresented in given GO categories or in predefined gene sets compared with the distribution of the whole set of genes represented on the chip.
- Tools developed along this line include:
  - GOMINER;
  - GENMAPP;
  - ONTO-TOOLs;
  - CHIPINFO;
  - GOSTAT.



Limitation of Above Methods

- No further use made of information contained in expression values for the non-DEG list genes
- The level of differential expression of the genes in the significant gene list is not taken into consideration.
- The correlation structure of the expression data is not considered at all.

Introduction of GSEA

- First explored in Mootha's *Nature Genetics (03)* paper, fully formulated in PNAS(05) paper.
- *GSEA:* evaluate microarray data at the level of gene sets, which is defined based on prior knowledge (such as gene sets from GO categories or pathways from KEGG).

Overview of GSEA

- Given a prior defined gene set S , *GSEA* is to determine whether members of S are randomly distributed throughout the list, or primarily found at the top or bottom in the list.
- Step of *GSEA*:
  - Calculation of an enrichment score (*ES*).
  - Estimation of significance level of *ES*.
  - Adjustment for *MHT*.

Calculation of ES

- Notation: D is the expression dataset with N genes and k samples; C is a phenotype or profile of interest; N_H is gene number of S,
- Rank order N genes to form $L=\{ g_1, \cdots, g_N\}$ according their correlation $r(g_j)=r_j$.
- Define:

$$P_{hit}(S,i) = \sum_{g_j \in S, j \le i} \frac{|r_j|^p}{N_R}, \; N_R = \sum_{g_j \in S} |r_j|^p;$$

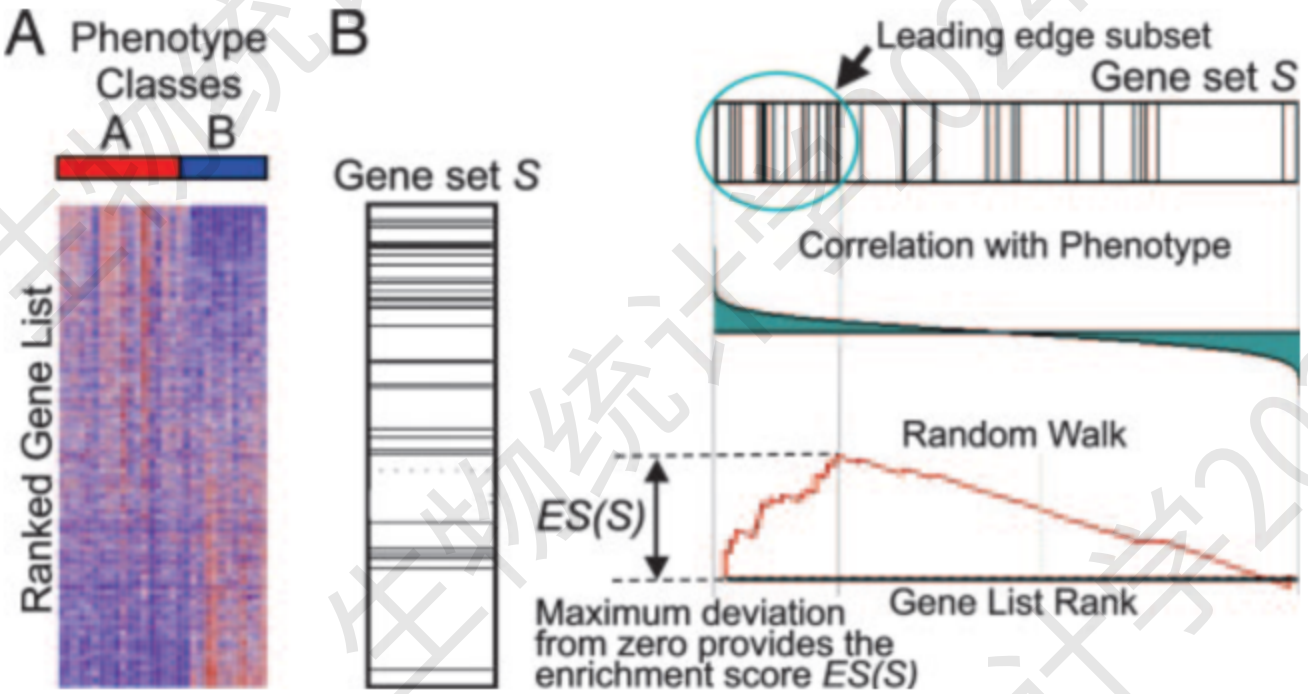
$$P_{miss}(S,i) = \sum_{g_j \notin S, j \le i} \frac{1}{N - N_H}$$

Where p is a constant to control the weight of ranks.

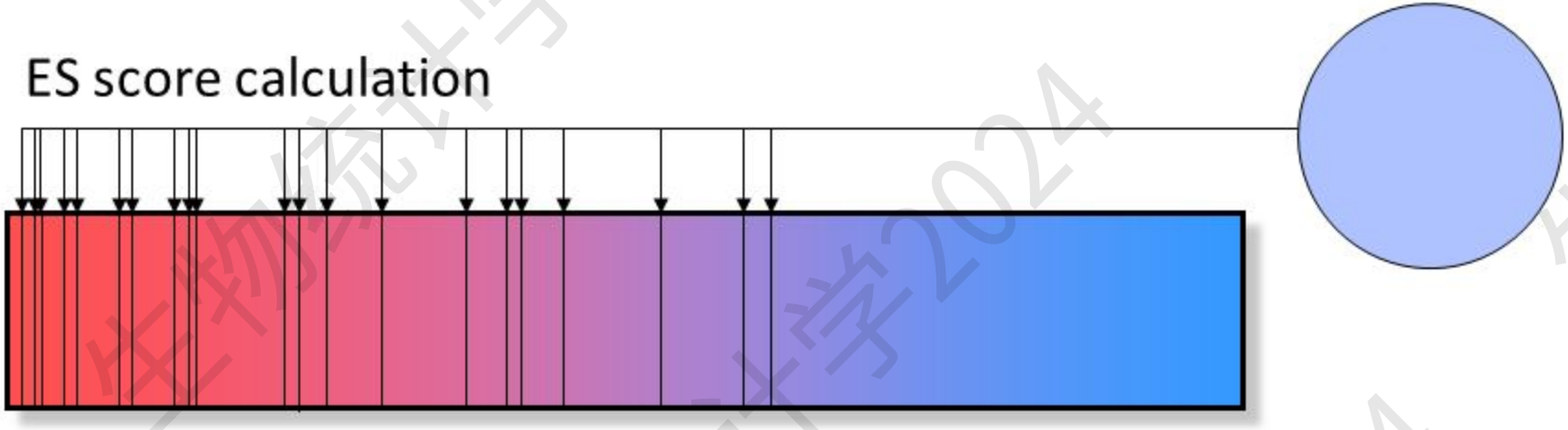
Calculation of ES

- The *ES* is the maximum deviation from zero of P_{hit}-P_{miss}.
- For a randomly distributed S, ES(S) will be small, but if it is concentrated at the top or bottom of the list, the score will be high.
- When p=0, this reduces to the standard *Kolmogorov-Smirnov statistic*.
  - As P_{hit} is the empirical distribution for genes in S, while P_{miss} is the one for genes outside S.

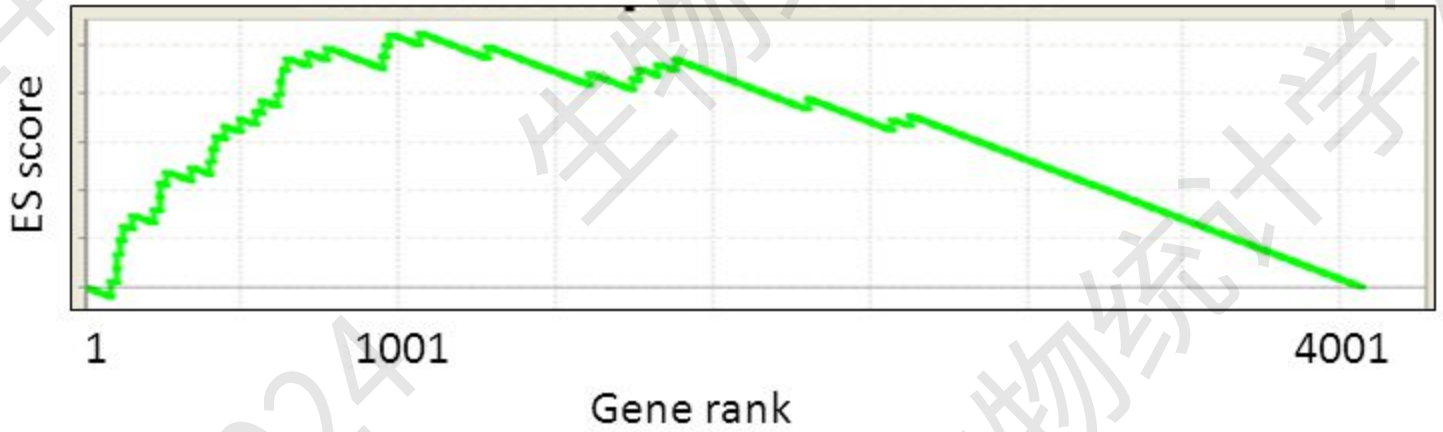
GSEA Overview



GSEA principle



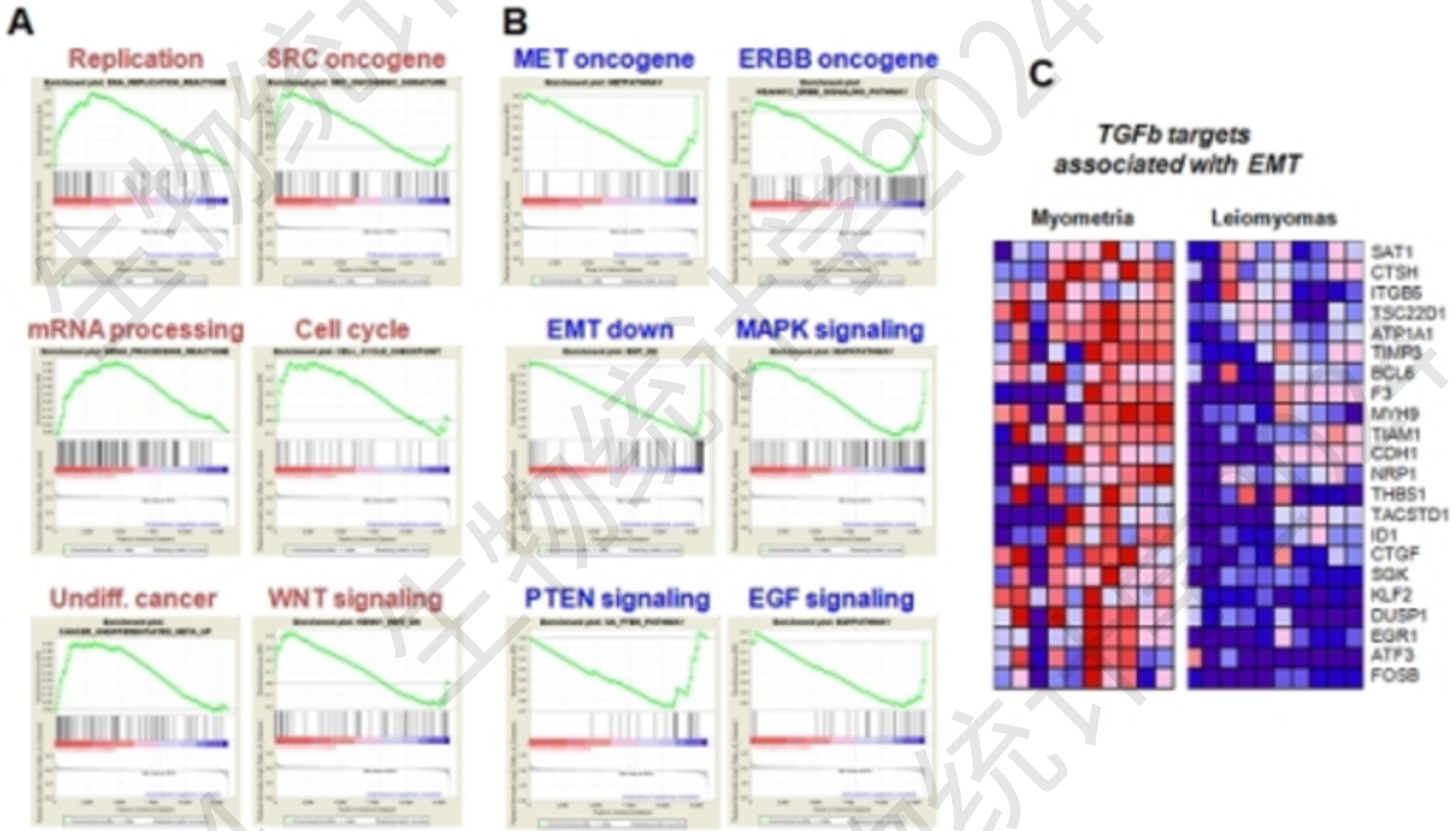
Every present gene (black vertical bar) gives a positive contribution, every absent gene (no vertical bar) gives a negative contribution



Estimating Significance

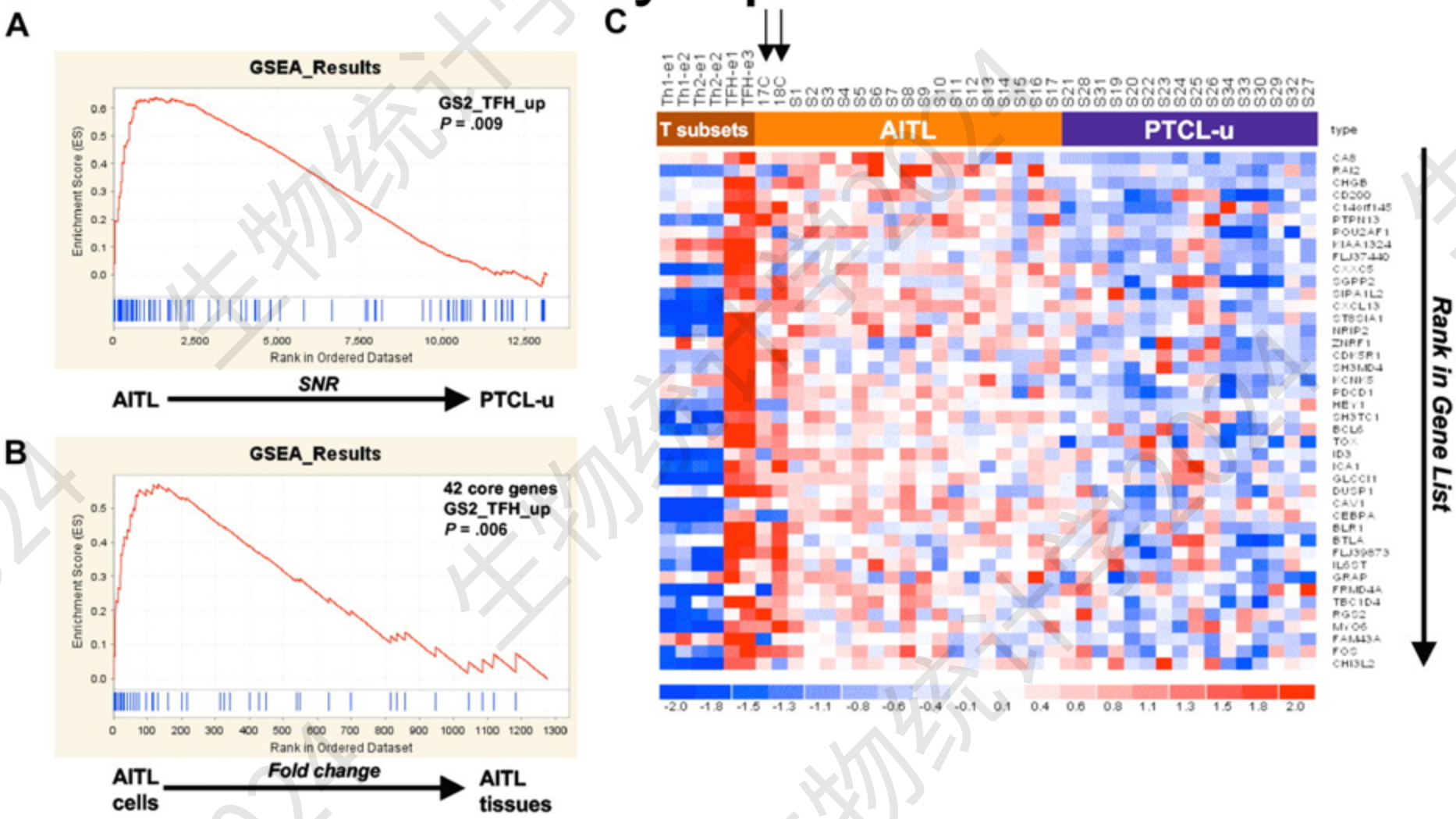
- Randomly assign the original phenotype labels to samples, reorder genes, re-compute ES(S).
- Repeat for 1000 permutations, and create a histogram of the corresponding ES_{NULL};
- Estimate nominal p-value for S from ES_{NULL} and observed ES(S).

GSEA tools: RDAVIDWebService



Reference:
http://bioconductor.org/packages/release/bioc/html/RDAVIDWebService.html

GSEA example: nodal peripheral T-cell lymphoma



Reference: http://www.bloodjournal.org/content/109/11/4952?sso-checked=true

第6-5章 Variable Selection

Variable Selection Problem

- A common problem is that there is a large set of candidate predictor variables.

- Goal is to choose a small subset from the larger set so that the resulting regression model is **simple**, yet have **good predictive ability**.

Two basic Methods of Selecting Predictors

- **Stepwise regression**: **Enter and remove predictors, in a stepwise manner, until there is no justifiable reason to enter or remove more.**
- **Best subsets regression**: Select the subset of predictors that do the best at meeting some well-defined objective criterion.

Stepwise Regression: the Idea

- Start with no predictors in the "**stepwise model**."
- At each step, enter or remove a predictor based on partial *F*-tests (that is, the *t*-tests).
- Stop when no more predictors can be justifiably entered or removed from the stepwise model.

Drawbacks of Stepwise Regression

- The final model is not guaranteed to be optimal in any specified sense.
- The procedure yields a single final model, although in practice there are often several equally good models.
- It doesn't take into account a researcher's knowledge about the predictors.

Stepwise Regression Methods

- Three broad categories:
  - Forward selection
  - Backward elimination
  - Stepwise regression

LASSO

- Lasso是另一种数据降维方法
- 不仅适用于线性情况，也适用于非线性情况
- Lasso是基于惩罚方法对样本数据进行变量选择
  - 通过对原本的系数进行压缩，将原本很小的系数直接压缩至0
  - 将这部分系数所对应的变量视为非显著性变量
  - 将不显著的变量直接舍弃

Lasso Model

(least absolute shrinkage and selection operator)

- Lasso: Least Absolute Shrinkage and Selection Operator
- Minimize

$$\min_{\beta} \sum_{i=1}^{n} \frac{1}{2}(y_i - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

- Equivalent to minimizing sum of squares with constraint (Lagarangian function)

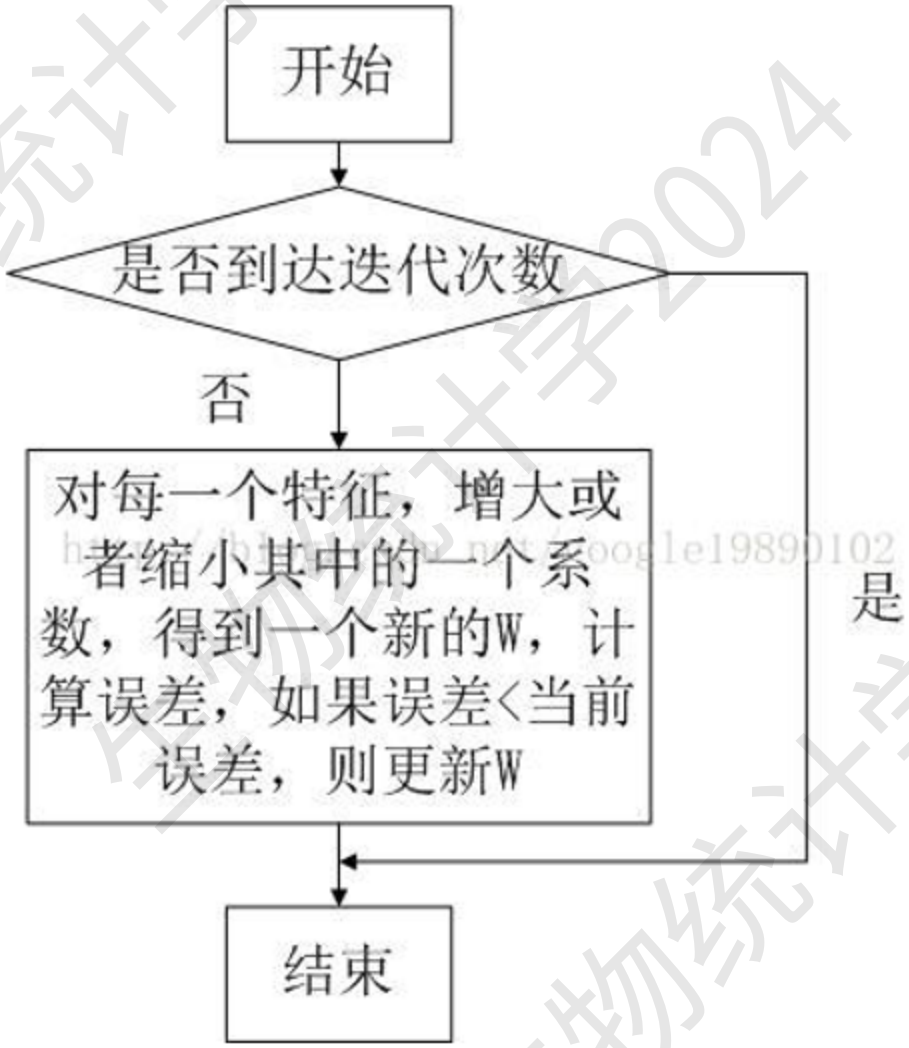
$$\sum_{j=1}^{p} |\beta_j| \leq s$$

Lasso Explanation

- The bound "s" is a tuning parameter. When "s" is large enough, the constraint has no effect and the solution is just the usual multiple linear least squares regression of y on x_1, x_2, ...x_p.
- However when for smaller values of s (s>=0) the solutions are shrunken versions of the least squares estimates. Often,some of the coefficients b_j are zero. Choosing "s" is like choosing the number of predictors to use in a regression model, and cross-validation is a good tool for estimating the best value for "s".

http://statweb.stanford.edu/~tibs/lasso/simple.html

Algorithms for Lasso



Algorithms for Lasso

- Standard convex optimizer
- Least angle regression (LAR) - Efron et al 2004- computes
- Entire path of solutions. State-of-the-Art until 2008
- Pathwise coordinate descent---New

Ridge Regression

- Minimize

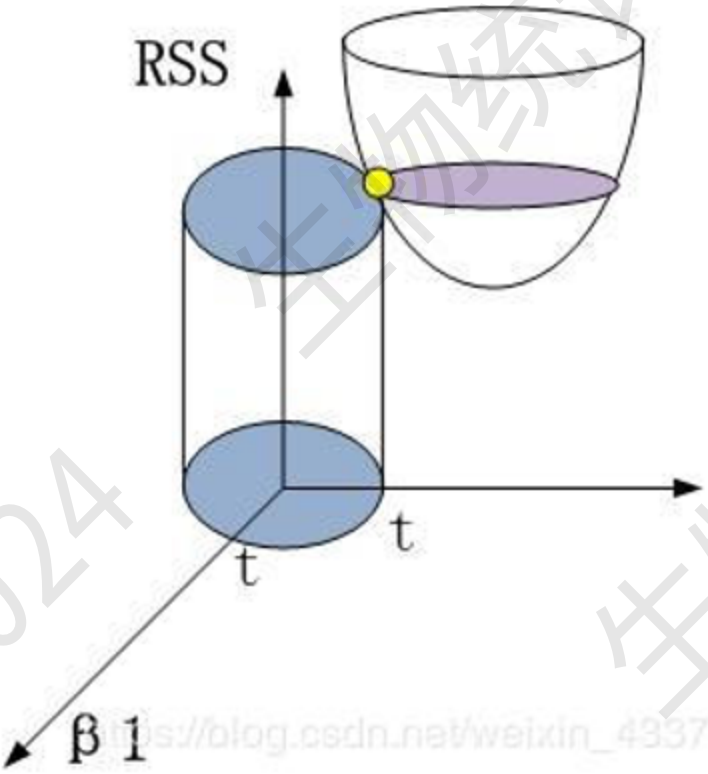
$$\min_{\beta} \frac{1}{2} \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} |\beta_j^2|$$

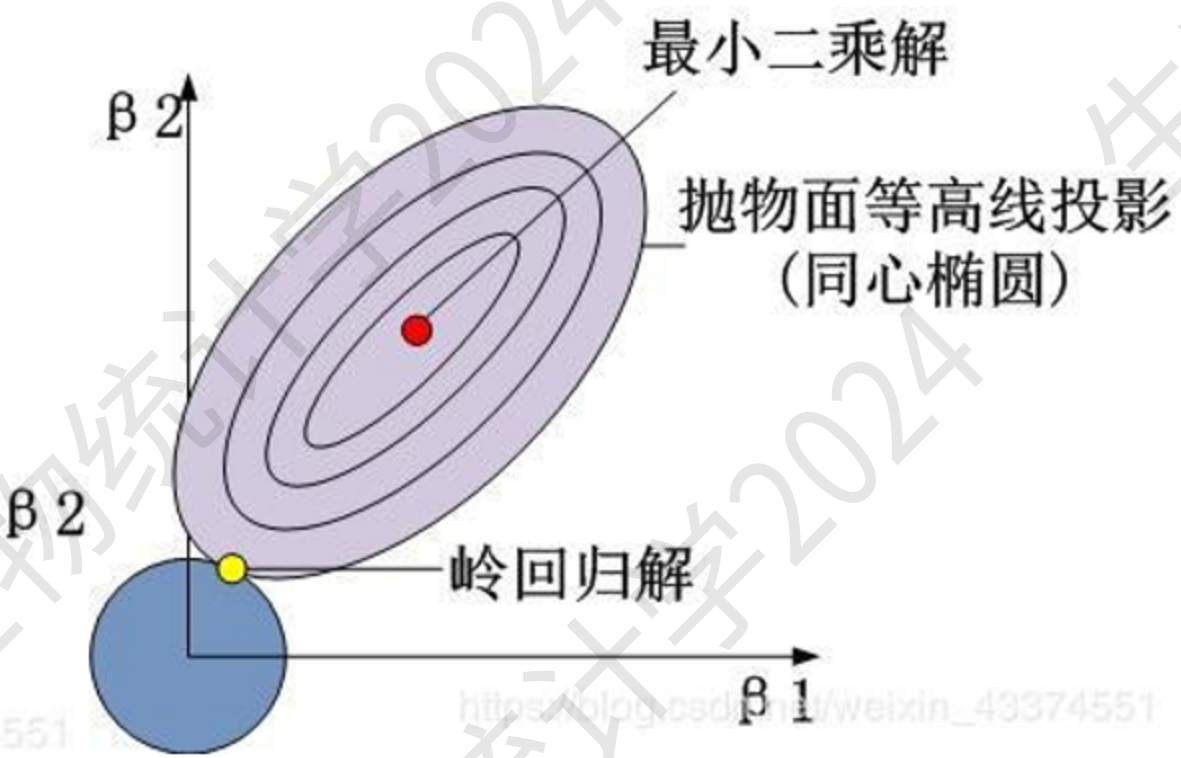
- Equivalent to minimizing sum of squares with constraint $\sum_{j=1}^{p} |\beta_j|^2 \leq s$

- Close-form solution

$$\beta^{\text{ridge}} = (X^T X + \lambda I)^{-1} X^T Y$$

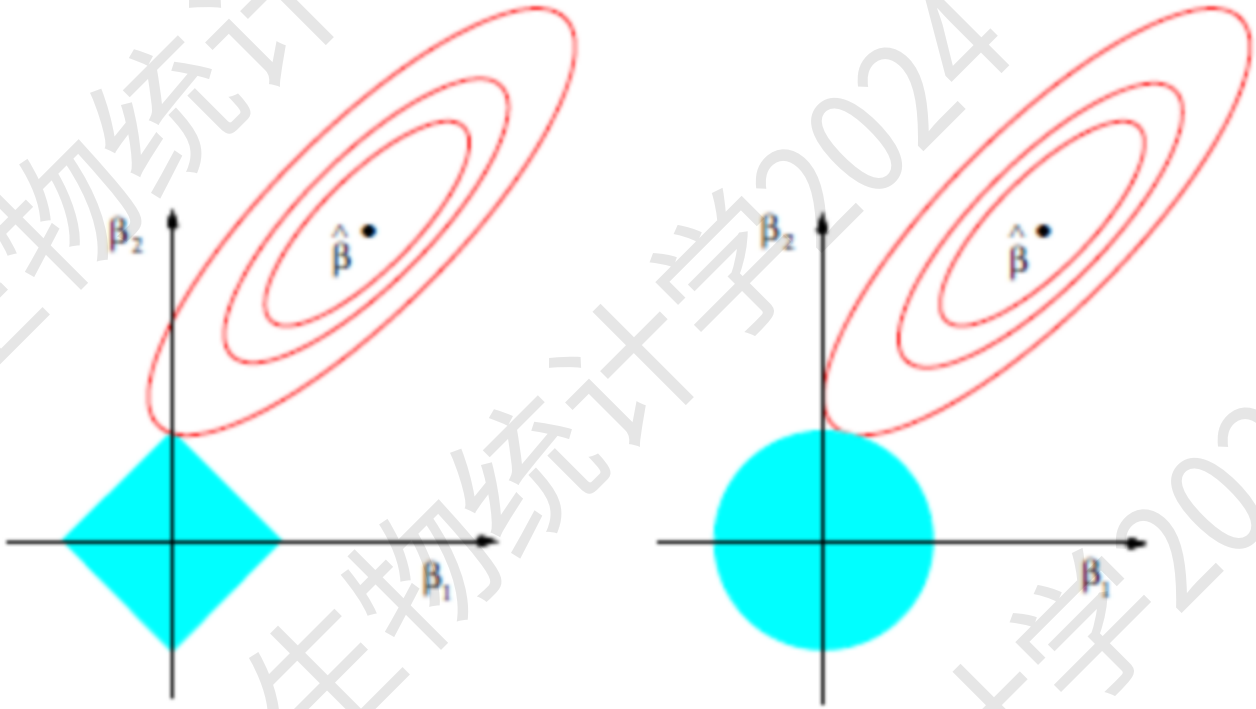
Ridge Regression





Ridge方法

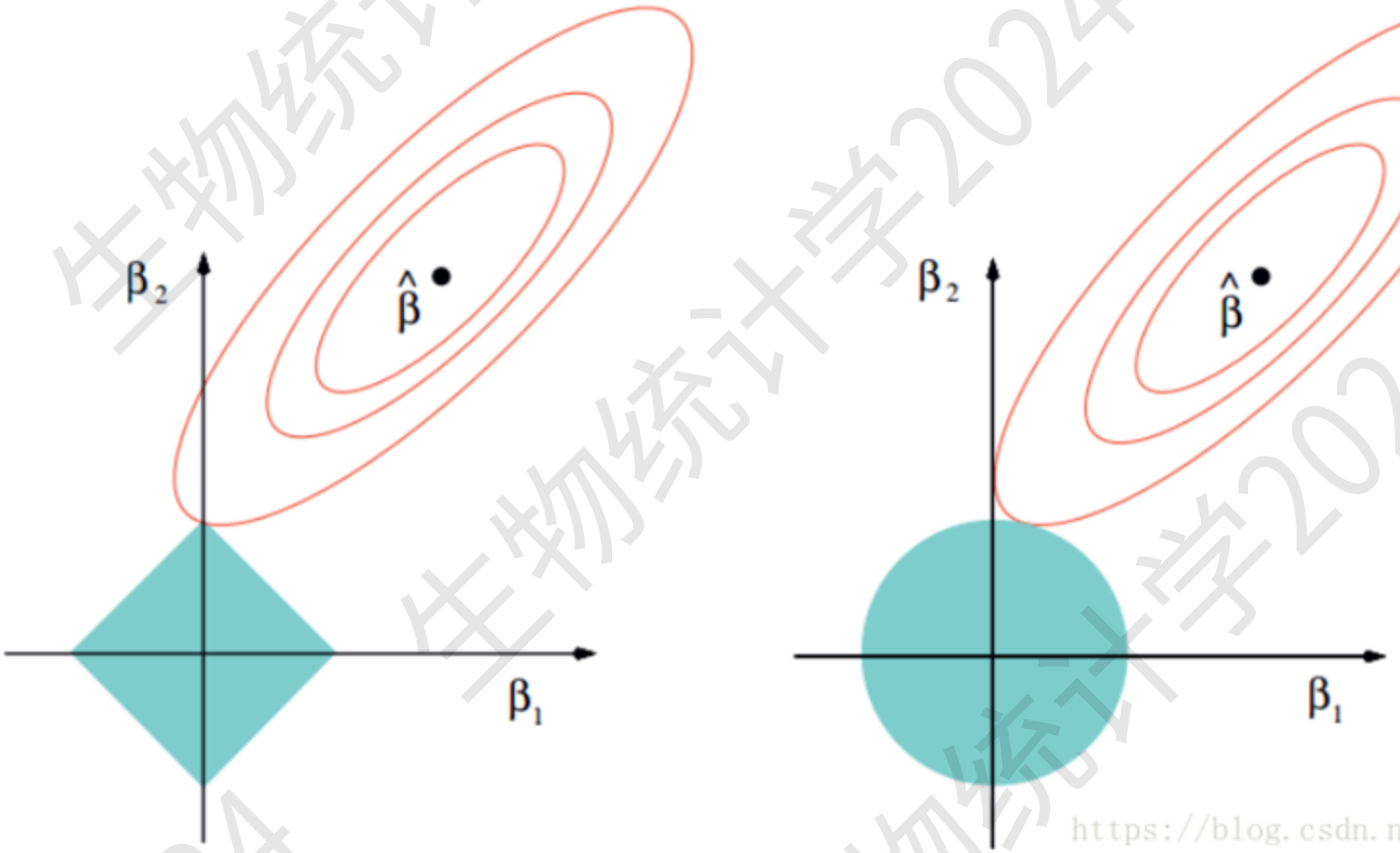
Lasso and Ridge Regression



左图：只要不是特殊情况下与正方形的边相切，一定是与某个顶点优先相交，那必然存在横纵坐标轴中的一个系数为0，起到对变量的筛选的作用。

右图：这个圆的限制下，点可以是圆上的任意一点，所以q＝2的时候也叫做岭回归，岭回归是起不到压缩变量的作用的，在这个图里也是可以看出来的。

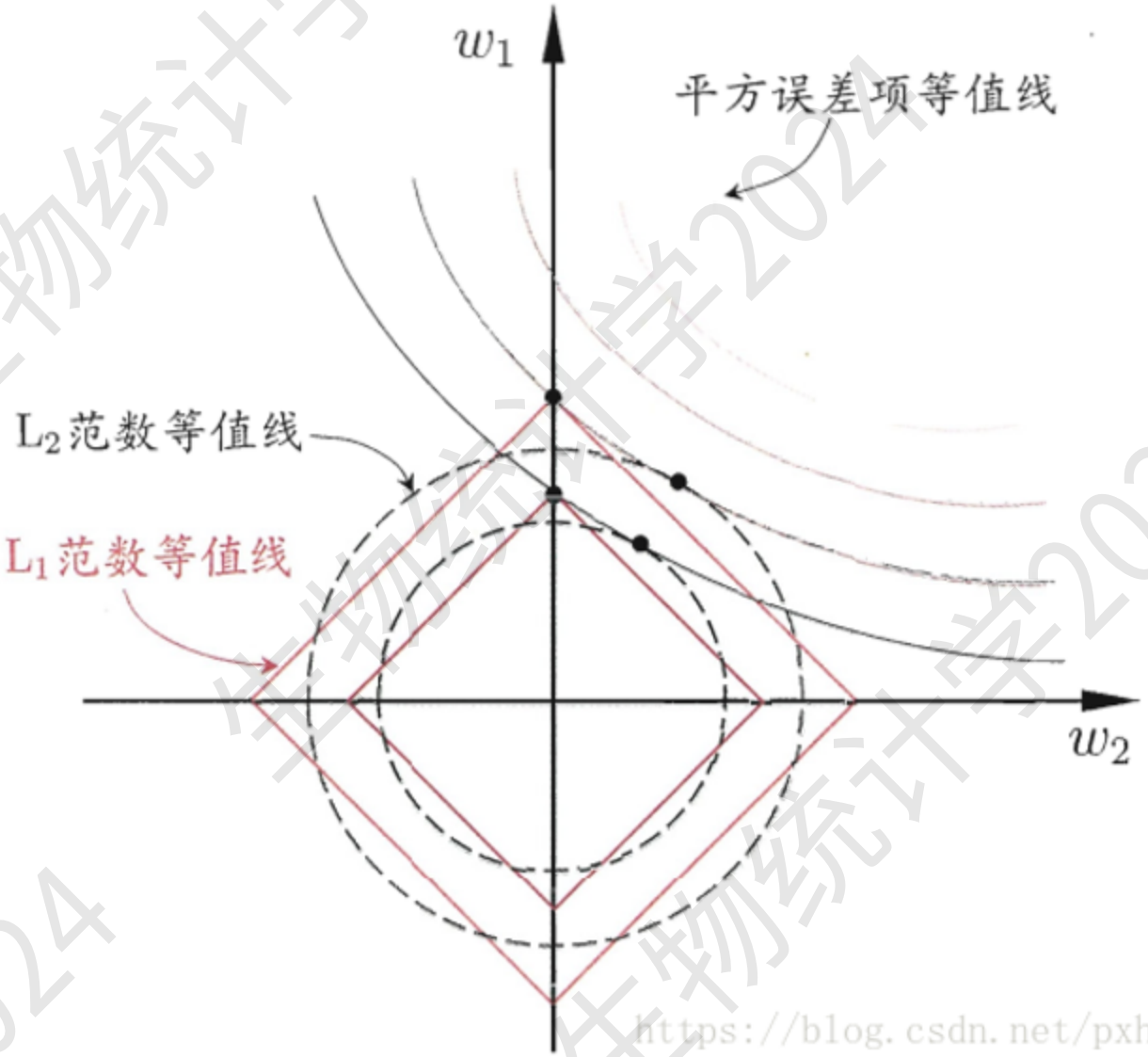
LASSO



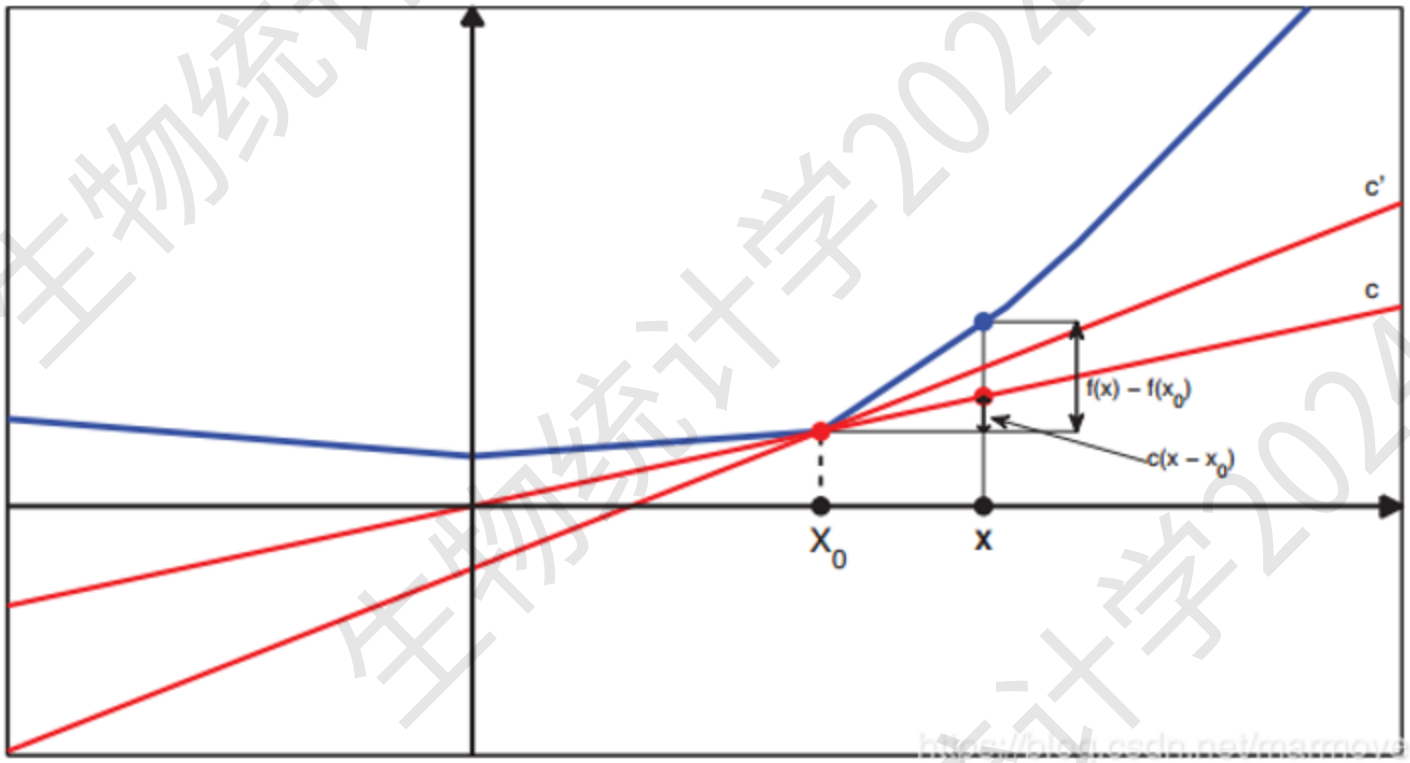
LASSO

- 以二维数据空间为例，说明lasso和Ridge两种方法的差异，左图对应于Lasso方法，右图对应于Ridge方法。
- 两个图是对应于两种方法的等高线与约束域。
- 红色的椭圆代表的是随着λ的变化所得到的残差平方和，β^为椭圆的中心点，为对应普通线性模型的最小二乘估计。
- 左右两个图的区别在于约束域，即对应的蓝色区域。
- 等高线和约束域的切点就是目标函数的最优解，Ridge方法对应的约束域是圆，其切点只会存在于圆周上，不会与坐标轴相切，则在任一维度上的取值都不为0，因此没有稀疏；对于Lasso方法,其约束域是正方形，会存在与坐标轴的切点，使得部分维度特征权重为0，因此很容易产生稀疏的结果。
- 所以，Lasso方法可以达到变量选择的效果，将不显著的变量系数压缩至0，而Ridge方法虽然也对原本的系数进行了一定程度的压缩，但是任一系数都不会压缩至0，最终模型保留了所有的变量。

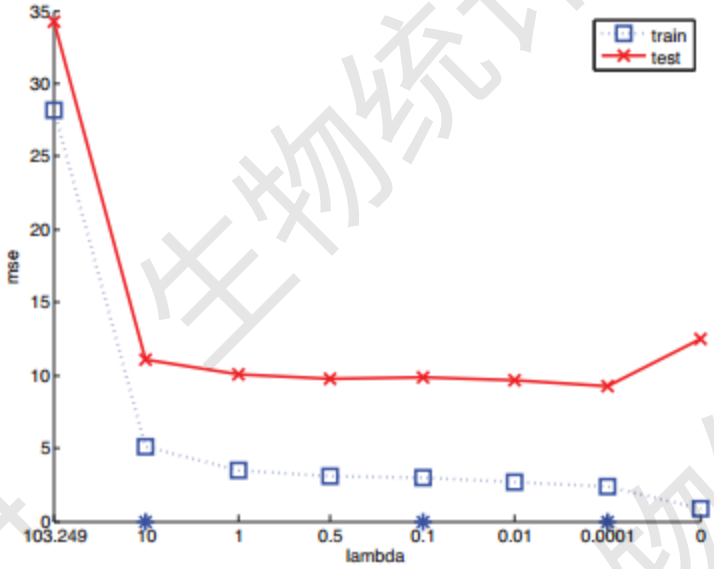
LASSO



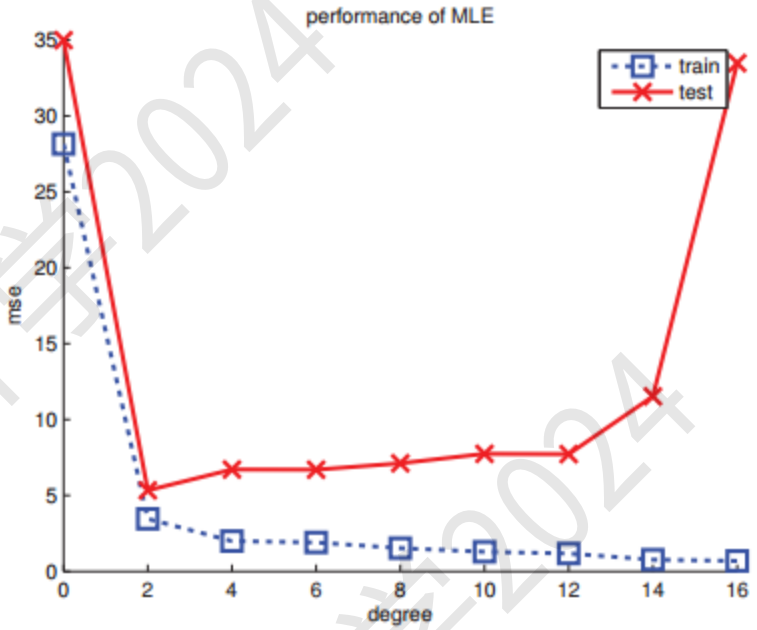
LASSO



LASSO



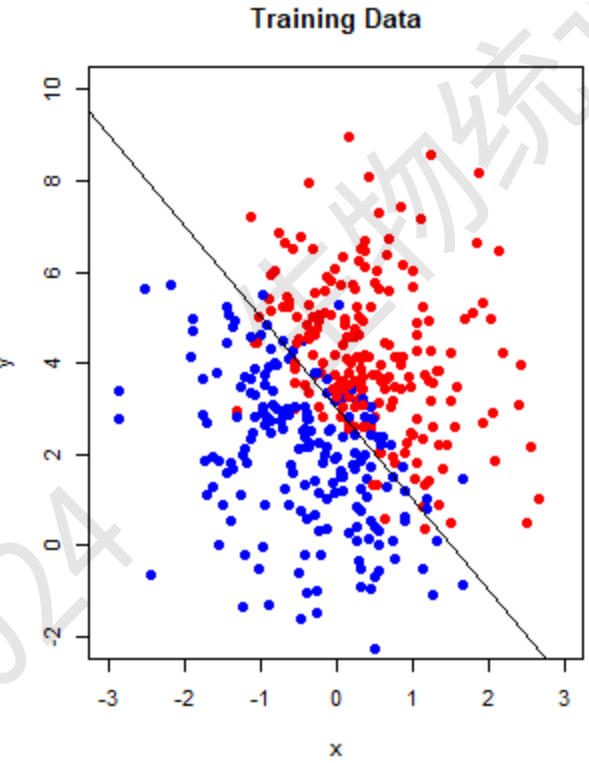
(a)

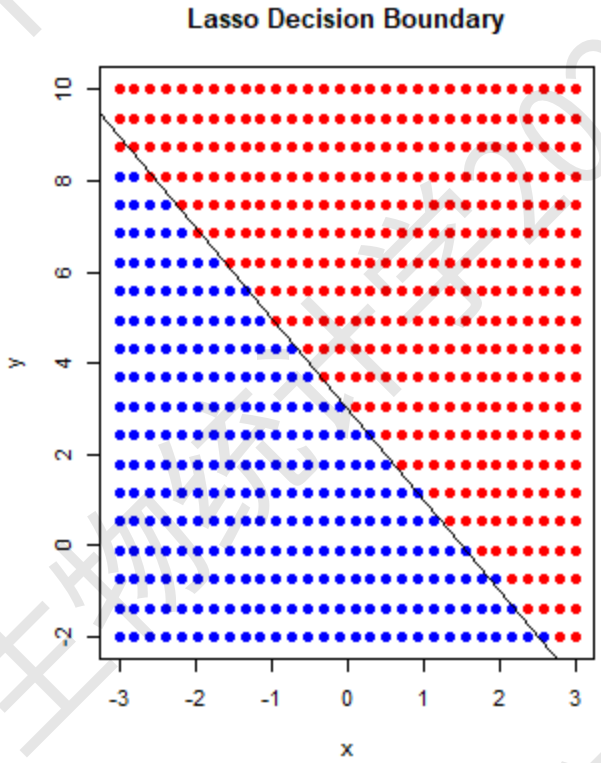


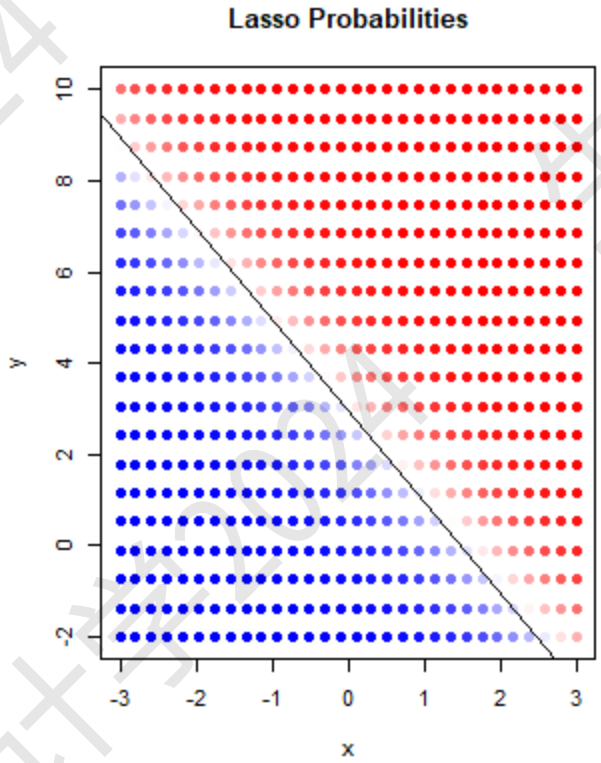
(b)

- 左边是lasso，MSE随着变化的情况
- 右边则是subset selection随着K的变化
- 这两个算法的性能是比较接近的

Lasso Model

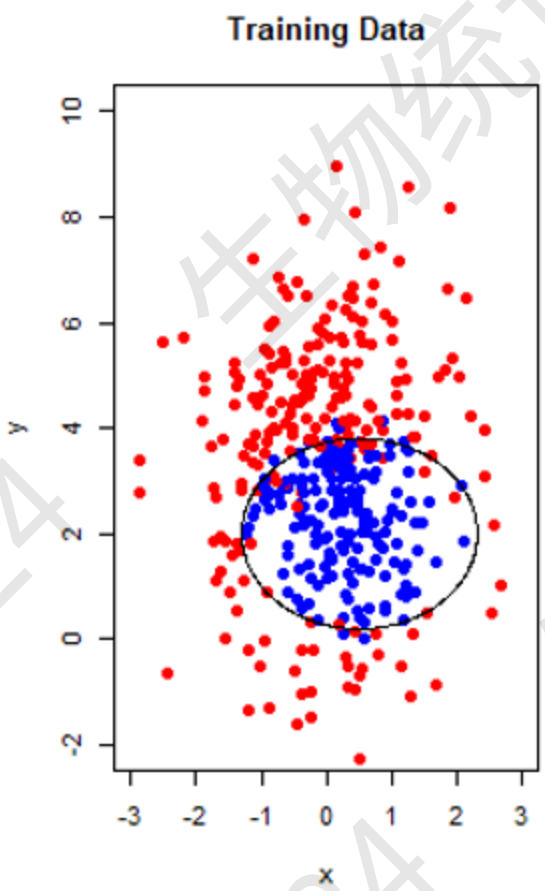


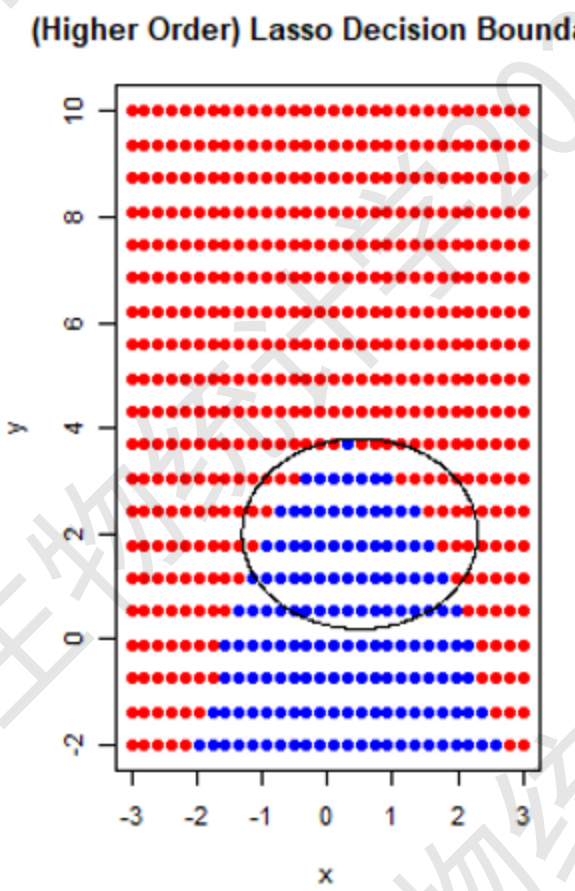


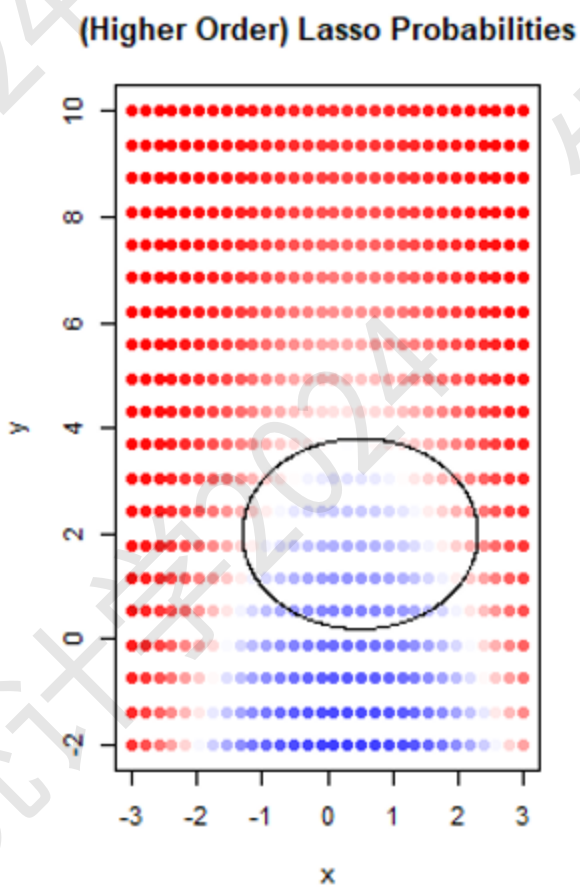


Reference: https://healthcare.ai/visual-tour-lasso-random-forest/

Lasso Model



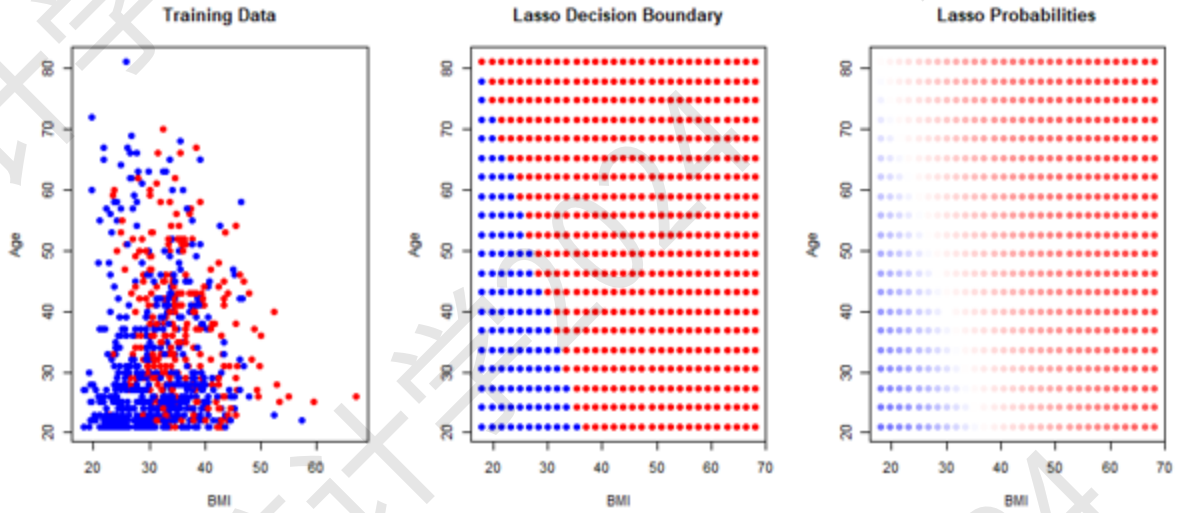




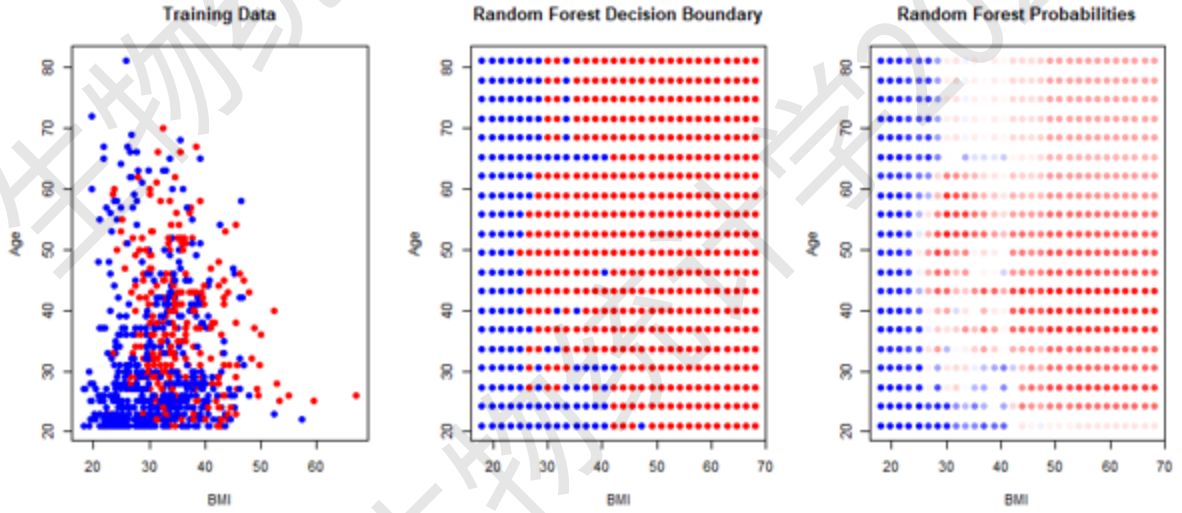
Reference: https://healthcare.ai/visual-tour-lasso-random-forest/

Lasso vs. Random forest

Lasso

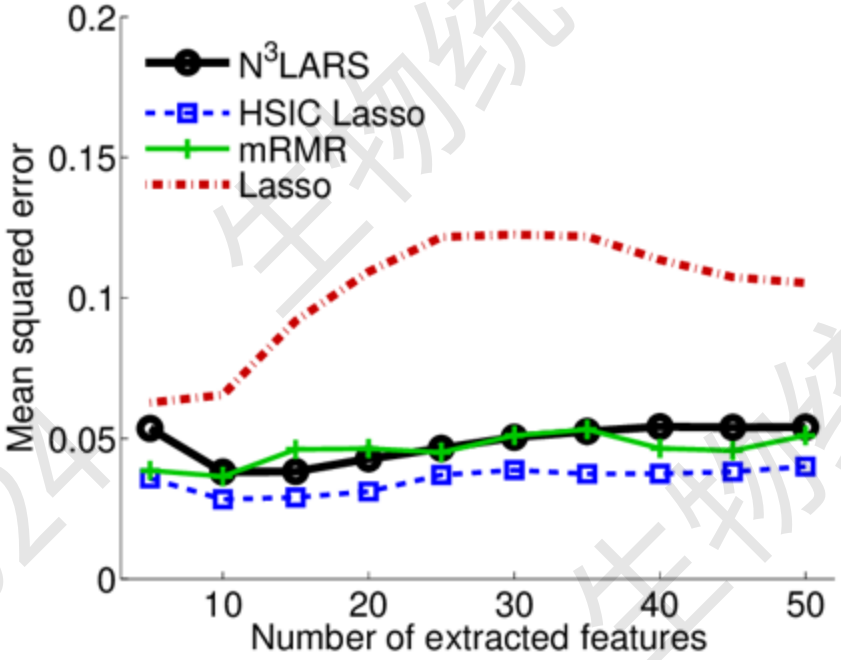


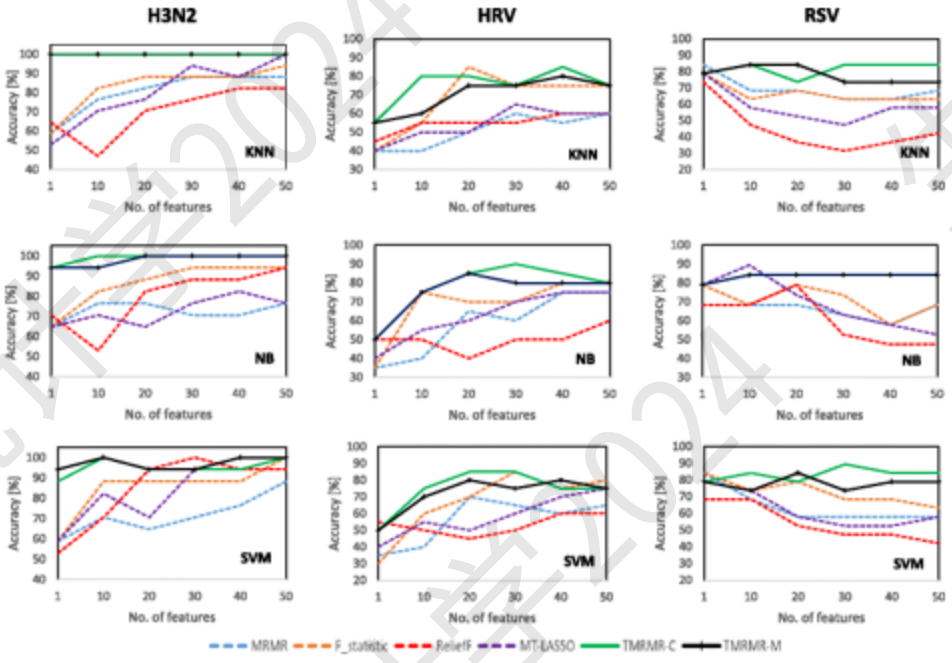
Random forest



Reference: https://healthcare.ai/visual-tour-lasso-random-forest/

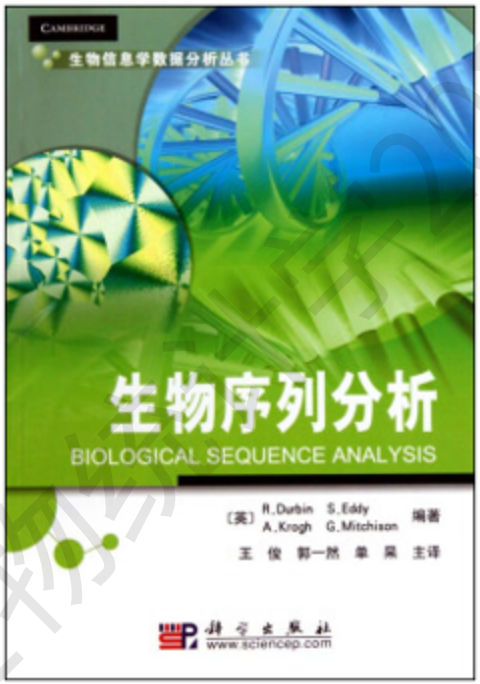
Other variable selection methods





References


2016
基因组学
GENOMICS


CAMBRIDGE
生物信息学数据分析丛书
生物序列分析
BIOLOGICAL SEQUENCE ANALYSIS
[英] R.Durbin S.Eddy A.Krogh G.Mitchison 编著
王俊 郭一然 单梁 主译
www.sciencep.com


生物统计学
(第五版)

Slides credits

——生物信息学研究方法概述：北京大学生物信息中心

——生物统计学：卜东波@中国科学院计算技术研究所，邓明华@北京大学

——神经网络与深度学习: 邱锡鹏@复旦大学

——Introduction to Computational Biology and Bioinformatics: Xiaole Shirley Liu Lab@Havard University

——Combinatorial Methods in Computation Biology: Ken Sung Lab@NUS

——Deep Learning in the Life Sciences: MIT

——Probabilistic Graphical Models: Eric Xing@CMU

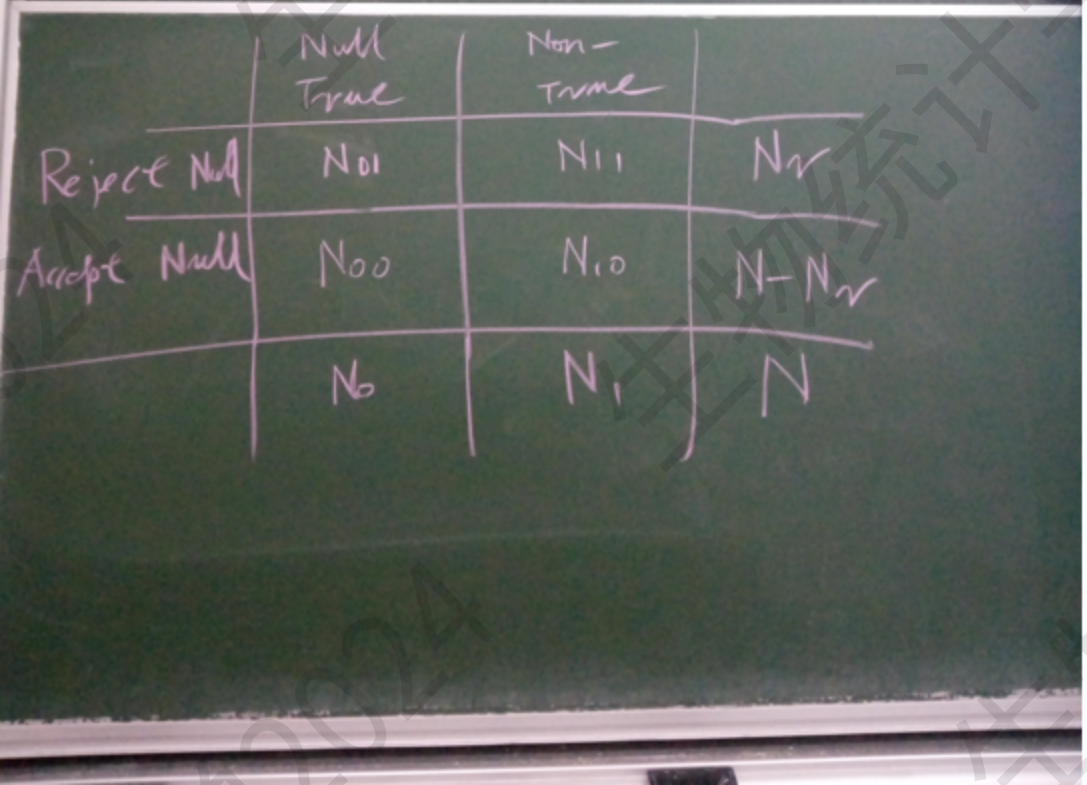
——Numerous other leading researchers and leading labs.....

补充知识

- TP, TN, FP, FN
- FPR, FDR

Measuring the Accuracy

	Real Negative	Real Positive
Claimed Positive	False Positve (FP)	True Positive (TP)
Claimed Negative	True Negative (TN)	False Negative (FN)

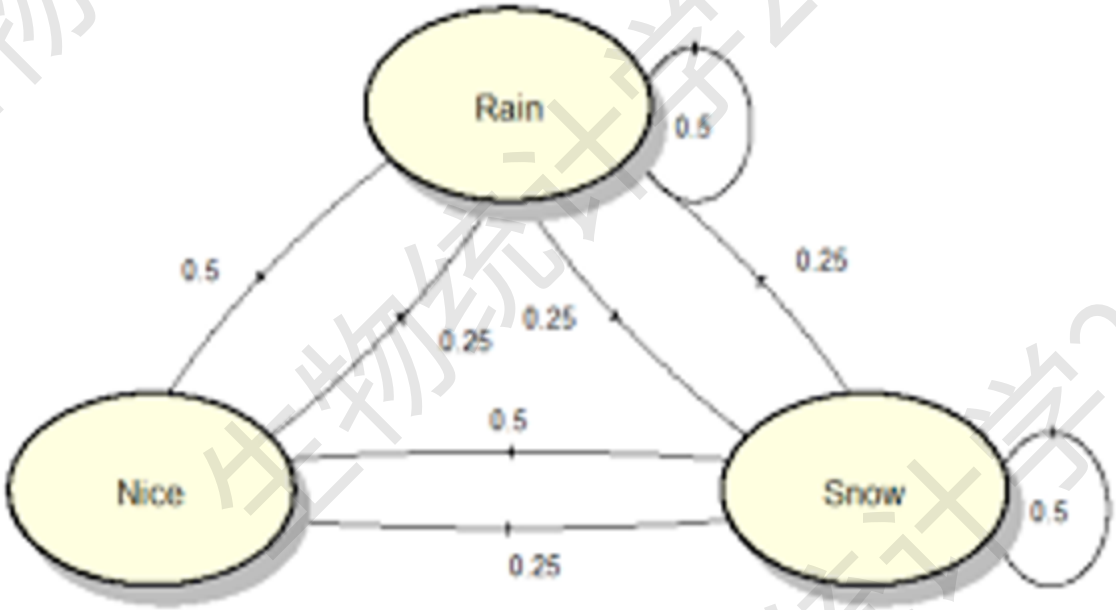


$$FPR = \frac{FP}{TN + FP}$$

$$FDR \approx \frac{N_{01}}{N_r}$$

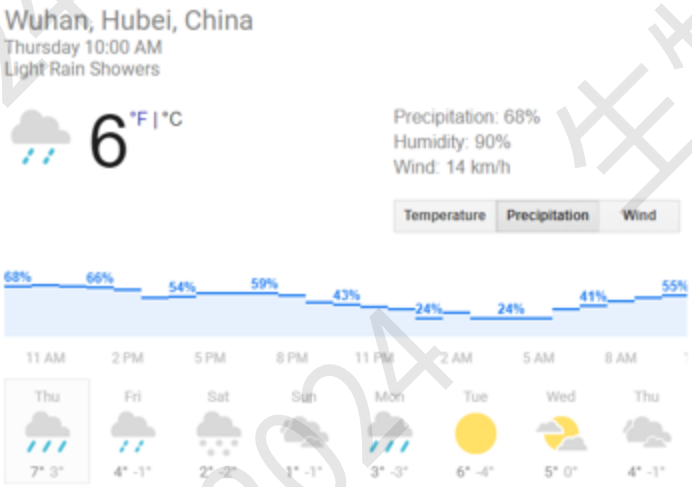
It is snowing

Markov Model to compute the probability on the nth day



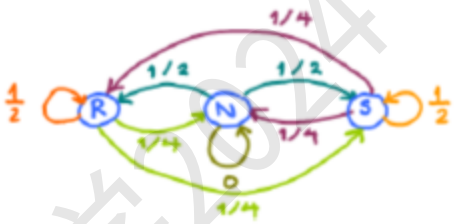
It is snowing

Markov Model to compute the probability on the nth day...



4. **(Markov Chain)** "The Land of Oz is blessed by many things, but not by good weather. They never have two nice days in a row. If they have a nice day, they are just as likely to have snow as to have rain on the next day. If they have snow or rain, they have an even chance of having the same the next day. If there is change from snow or rain, only half of the time is this a change to a nice day." [Grinstead and Snell, Ex 11.1][Kemeny, Snell, and Thompson, 1974]

   a. Draw the Markov chain corresponding to how the weather in the Land of Oz changes from one day to the next.
   Hint: This Markov chain will have three states: rain (R), nice (N), and snow (S).



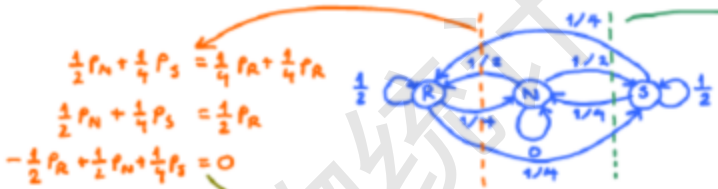
Note that we set the color of each arrow and its label to match with the corresponding part of the description which is underlined with the same color.

   b. Find the corresponding (probability) transition matrix **P**.

$$P = \begin{array}{c} \\ R \\ N \\ S \end{array} \begin{array}{c} \begin{array}{ccc} R & N & S \end{array} \\ \begin{bmatrix} 1/2 & 1/4 & 1/4 \\ 1/2 & 0 & 1/2 \\ 1/4 & 1/4 & 1/2 \end{bmatrix} \end{array}$$

   c. Find the steady-state probabilities by using balance equations

We first write two balance equations from the two boundaries shown below:



$\frac{1}{2}P_N + \frac{1}{4}P_S = \frac{1}{4}P_R + \frac{1}{4}P_R$

$\frac{1}{2}P_N + \frac{1}{4}P_S = \frac{1}{2}P_R$

$-\frac{1}{2}P_R + \frac{1}{2}P_N + \frac{1}{4}P_S = 0$

$\frac{1}{4}P_S + \frac{1}{4}P_S = \frac{1}{2}P_N + \frac{1}{4}P_R$

$\frac{1}{2}P_S = \frac{1}{2}P_N + \frac{1}{4}P_R$

$\frac{1}{4}P_R + \frac{1}{2}P_N - \frac{1}{2}P_S = 0$

One more equation: $P_R + P_N + P_S = 1$

Solve 3 eqns, 3 unknowns.

$P_R = 0.4$, $P_N = 0.2$, and $P_S = 0.4$