WDS

DataScience@Web

机器学习讨论班

2018年暑期

讨论班约定

- 组织者：郭雪纯
- 教材：《机器学习》周志华
- 网站：http://cse.seu.edu.cn/PersonalPage/x.zhang/wds/ml2018/
- 请假制度：向郭雪纯报备
- 每位同学报告前**至少三天**在群里发布报告PPT和相关文件
- PPT采用统一模板

报告内容约定

- PPT包含：
  - 问题描述 | 模型描述 | 概念解释
  - 应用场景举例
  - 重要细节
  - 优点与缺点 | 适用性与非适用性
  - 参考资料 | 第三方工具
- 图片为主，文字为辅（但需保证PPT的自说明性）
- 时间：每个topic（或半个topic）30-60分钟
- 请尽量保证报告的直观性、可理解性！

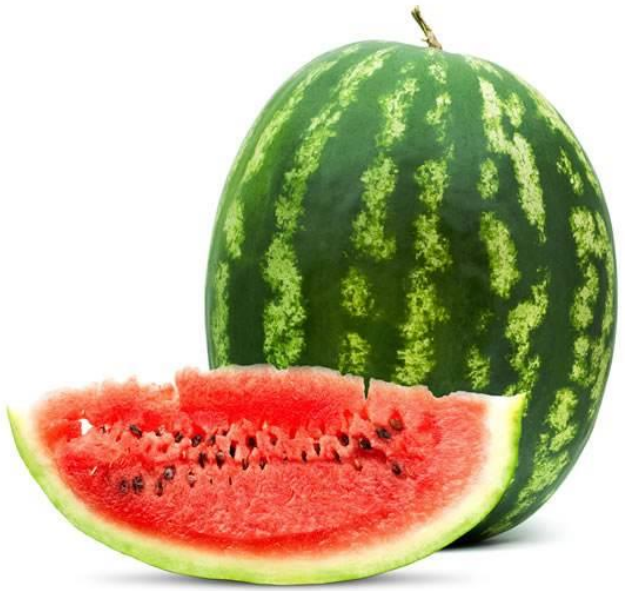
预备知识

- 西瓜书第一章：绪论
  - 假设空间
  - 归纳偏好
- 西瓜书第二章：模型评估与选择
  - 误差与拟合
  - 评估方法
  - 性能度量
  - 比较检验
  - 偏差与方差

报告安排

一：分类与回归	二：无监督学习	三：半、远程、强化、主动学习	四：统计推断
线性分类与线性回归			
k近邻			
朴素贝叶斯			
逻辑斯蒂回归			
最大熵模型			
支持向量机			
决策树			
集成学习			
多标签分类			
协同分类			
其他回归问题			

讨论班福利



1. 线性分类与回归

张祥

x.zhang@seu.edu.cn

介绍内容

- 分类与回归
- 线性与非线性模型
- 线性回归
- 对数几率回归
- 线性判别分析
- 多分类学习
- 类别不平衡问题
- 实验：房屋价格估计

60秒机器学习

“A computer program is said to learn from experience E with respect to some task T and some performance measure P if its performance on T, as measured by P, improves with experience E.”

- Andrew Ng

监督学习

训练集

2015 年福建省高一数学竞赛试题参考答案及评分标准
（考试时间：5 月 10 日上午 8：30—11：00）

一、选择题（每小题 6 分，共 36 分）

1. 集合 $A=\{x \mid |x-1|<3, x\in N\}$ 的子集有（　　）

A. 4 个　　B. 8 个　　C. 16 个　　D. 32 个

【答案】 C

【解答】 由 $|x-1|<3$，知 $-2<x<4$，结合 $x\in N$，得 $A=\{0,1,2,3\}$。

∴ A 的子集有 $2^4=16$ 个。

2. 若直线 l_2 与直线 l_1：$y=2x-1$ 关于直线 $y=x$ 对称，则 l_2 与两坐标轴围成的三角形的面积为（　　）

A. 1　　B. $\frac{2}{3}$　　C. $\frac{1}{2}$　　D. $\frac{1}{4}$

【答案】 D

【解答】 在直线 l_1：$y=2x-1$ 取点 $A(0,-1)$，则 $A(0,-1)$ 关于直线 $y=x$ 的对称点 $A'(-1,0)$ 在直线 l_2 上。

又直线 l_1 与直线 $y=x$ 的交点 $P(1,1)$ 在直线 l_2。

∴ l_2 过 $A'(-1,0)$ 和 $P(1,1)$ 两点，其方程为 $y=\frac{1}{2}x+\frac{1}{2}$。 blog.sina.com.cn/sqing

验证集

机器学习模型

测试集

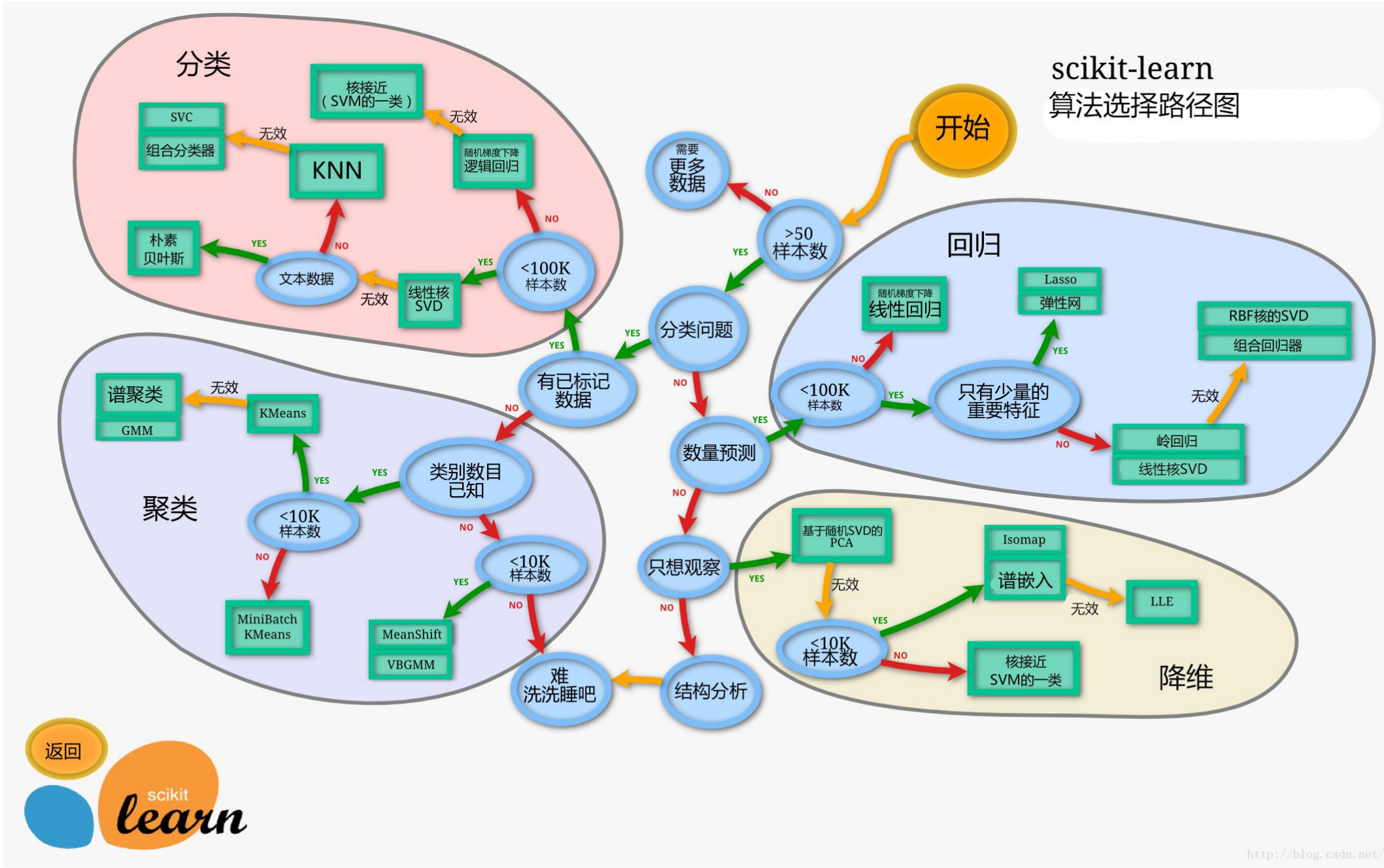
分类与回归

分类问题：
好瓜还是坏瓜？

回归问题：
今年南京横溪西瓜的产量？

分类与回归

分类	回归
• 标签预测 • 离散型	• 值预测 • 连续性



线性 | 非线性系统

“在一个系统中，如果两个不同因素的组合作用只是两个因素单独作用的简单叠加，这种关系或特性就是线性的。反之，如果一个系统中一个微小的因素能够导致用它的幅值无法衡量的结果，这种关系或特性就是非线性的。”

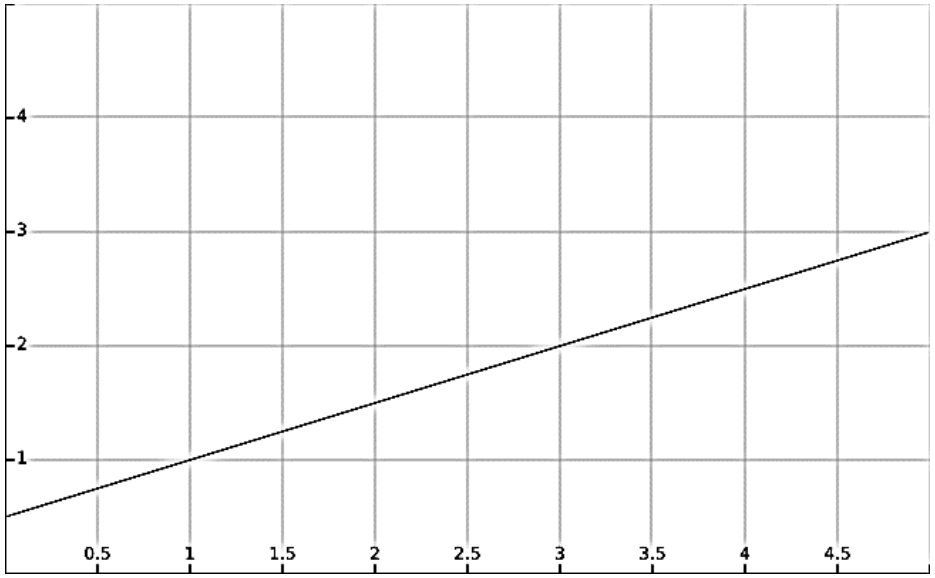
线性系统举例

非线性系统举例

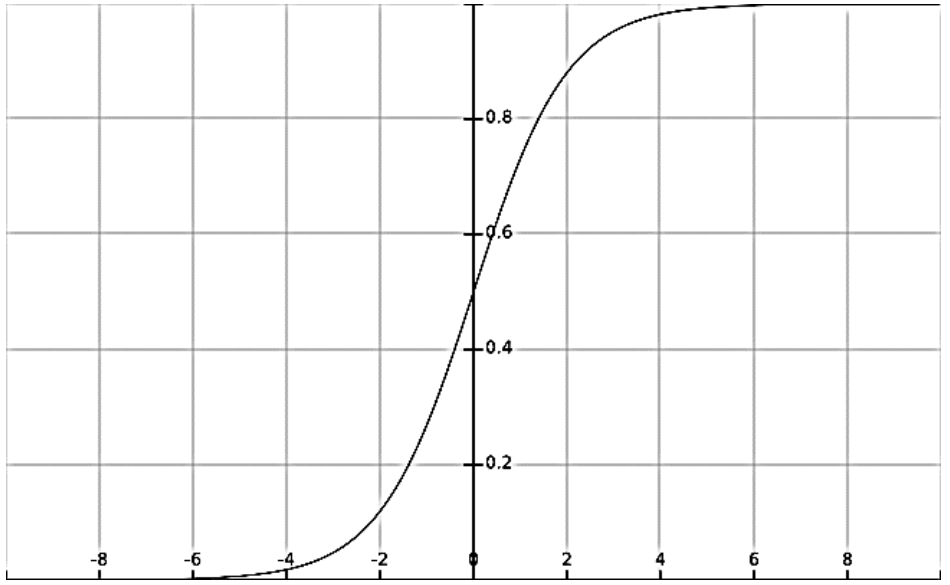
线性系统(模型)的基本特征

相加性	齐次性
$f(x_1 + x_2) = f(x_1) + f(x_2)$	$f(c \times x_1) = c \times f(x_1)$

线性与非线性的函数表示



$$f(x) = 0.5x + 0.5$$



$$sigmoid(x) = \frac{1}{1 + e^{-x}}$$

线性模型

x：由d个不同属性描述的样本，比如西瓜的色泽、根蒂、敲声

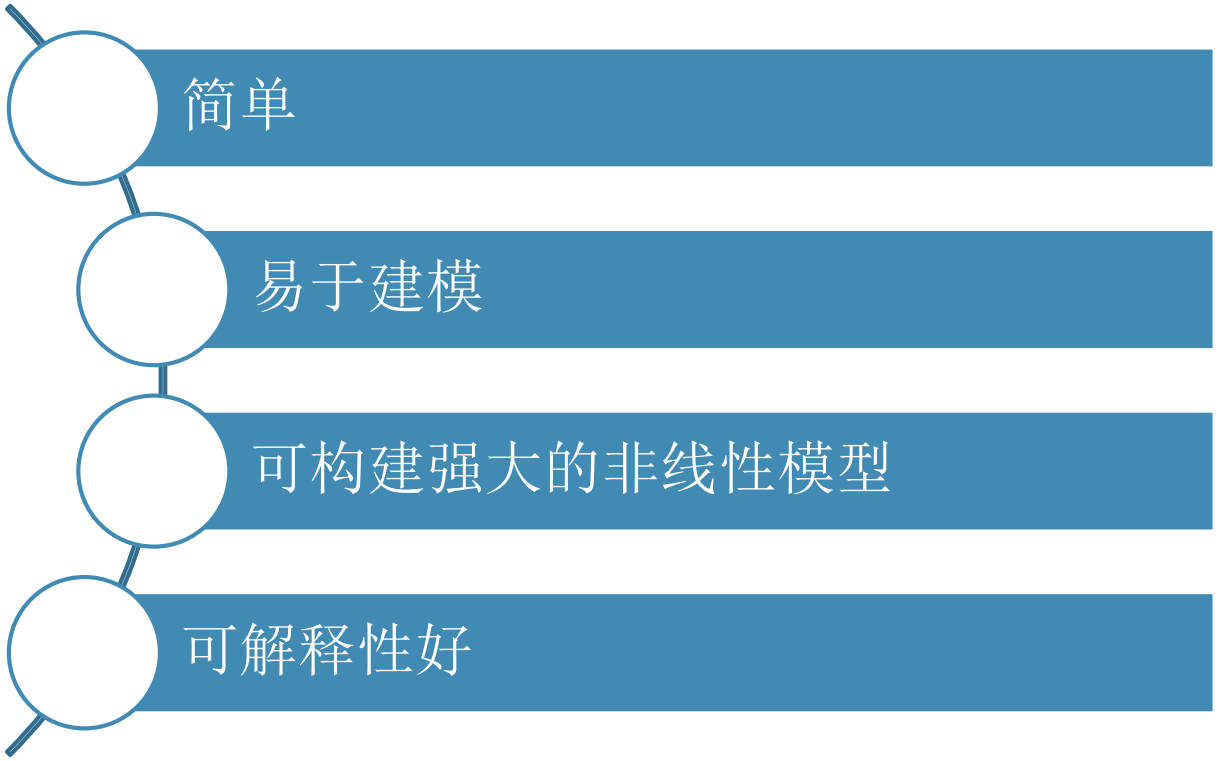
$f(x)$: 由样本的属性通过线性组合得到值，比如西瓜的质量得分

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_d \end{pmatrix}$$

$$f(x) = w_1 x_1 + \cdots + w_d x_d + b$$

$$f(x) = w^{\mathrm{T}} x + b$$

线性模型的优点



线性回归

给定数据集$D = \{(x_1, y_1), \dots, (x_m, y_m)\}$，其中$x_i = (x_{i1}, \dots x_{id})$，$y_1 \in \mathbb{R}$。 线性回归试图学习一个线性模型以尽可能准确地预测实值输出标记。

问题简化

假设输入属性的数目只有一个

- $D = \{(x_i, y_i)\}^m$

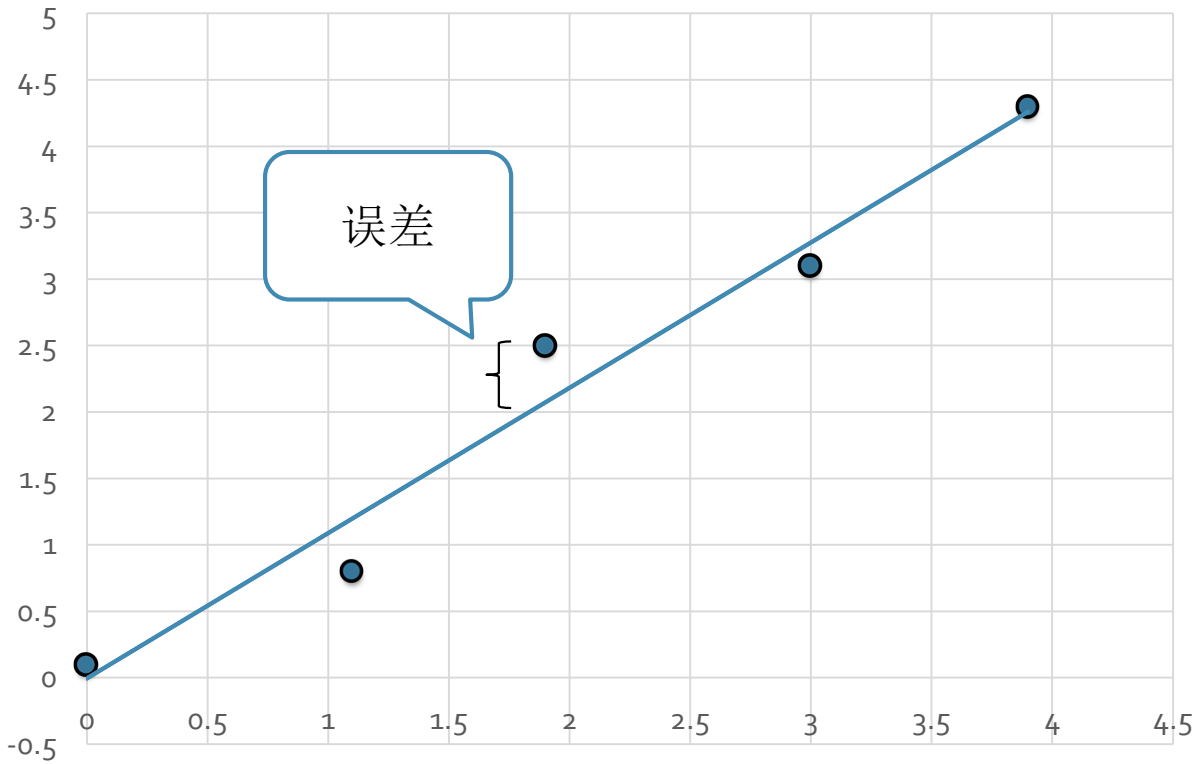
离散属性的连续化

- 存在“序”关系：身高的“高”“矮”→ $\{1.0, 0.0\}$
- 不存在“序”关系：瓜类的“西瓜”、“南瓜”、“黄瓜”→ $(0,0,1), (0,1,0), (1,0,0)$
- 无序属性的连续化可能会不恰当的引入序关系，对后续处理如距离计算造成误导

参数估计

- $f(x_i) = wx_i + b$，使得$f(x_i) \approx y_i$

参数估计方法：线性最小二乘法



最小二乘法：基于均方误差（又称为平方损失）最小化来进行模型求解：

$$(w^*, b^*) = \arg\min_{(w,b)} \sum_{i=1}^{m} (y_i - wx_i - b)^2$$

$$(w^*, b^*) = \arg\min_{(w,b)} L_{(w,b)}$$

求解方法：分别将$L_{(w,b)}$对w和b求导，使导数为0，得到线性方程组，计算出w和b的最优解。

插播：误差与残差

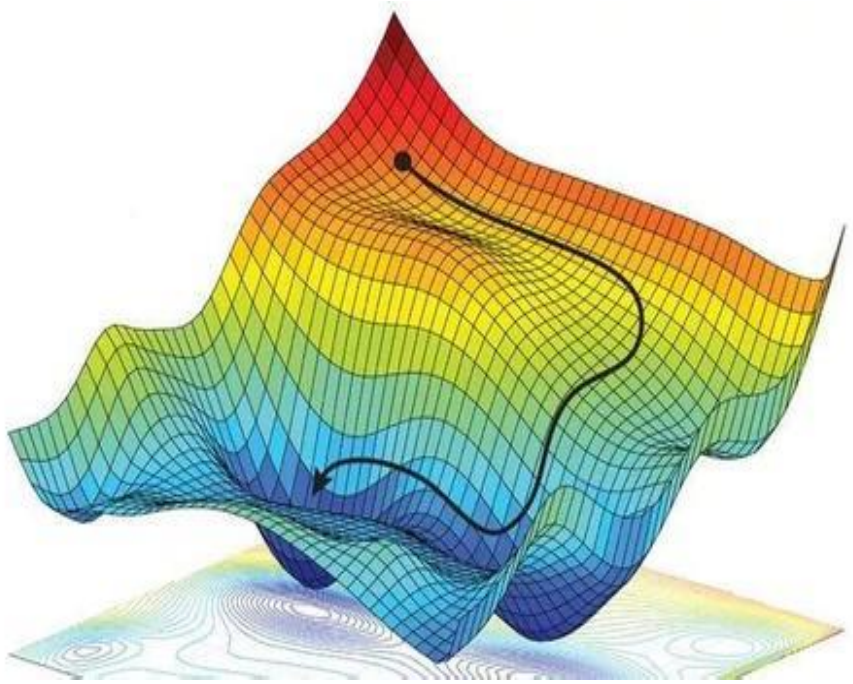
误差	残差
• Error • 观测值与真实值的偏离 • 反映测量的准确性	• Residual • 观测值与拟合值的偏离 • 反映预测的准确性

参数估计方法：线性最小二乘法

https://www.zhihu.com/question/37031188

参数估计方法：梯度下降法

$p = \begin{pmatrix} p_1 \\ p_2 \end{pmatrix}$，为一组参数，$L$为损失函数



假设最优化问题为：$p^* = \underset{(p)}{\operatorname{argmin}} L_{(p)}$

1. 随机指定p的初始值：p_1^0，p_2^0
2. $$\begin{bmatrix} p_1^1 \\ p_2^1 \end{bmatrix} = \begin{bmatrix} p_1^0 \\ p_2^0 \end{bmatrix} - \eta \begin{bmatrix} \Delta L(p_1^0) \\ \Delta L(p_2^0) \end{bmatrix}$$
3. …迭代
4. $$\begin{bmatrix} p_1^m \\ p_2^m \end{bmatrix} = \begin{bmatrix} p_1^{m-1} \\ p_2^{m-1} \end{bmatrix} - \eta \begin{bmatrix} \Delta L(p_1^{m-1}) \\ \Delta L(p_2^{m-1}) \end{bmatrix}$$

ΔL是损失函数对某个参数的偏导，称为梯度

梯度下降的相关主题

- 梯度下降的学习率（Learning Rate）
- 随机梯度下降法（Stochastic Gradient Descent）
- 特征缩放（Scaling）
- 参考：https://yoferzhang.gitbooks.io/machinelearningstudy/content/20170327ML04GradientDescent.html

线性（狭义）最小二乘 vs. 梯度下降

所谓封闭解(Closed-formed)：根据严格的公式推导，给出任意的自变量就可以求出其因变量

线性(狭义)最小二乘	梯度下降
• 封闭解 • 一元：线性方程组 • 多元：矩阵形式 • 全局最优解	• 迭代法 • 假设条件更宽泛（可解决的最优化问题更多） • 局部最优

多元线性回归

更一般的情况是：一个样本x_i由d个属性（多个自变量）描述。此时试图学得一个线性模型来刻画数据集，这称为多元线性回归。参数估计方法依然为最小二乘法。

线性回归的拟合度检验

被翻译为：决定系数、判定系数、拟合优度

R^2检验

$$R^2 = \frac{\sum_i (f(x_i) - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}$$

取值范围为0-1，越大越好

误差平方和

总离差平方和

其他模型检验方法

- 调整后的R^2
- t检验
- SE检验
- F检验
- 显著性度量: 临界值表

https://www.cnblogs.com/tangxianwei/p/8323495.html

线性对数回归

可以令模型预测值逼近y的衍生物，例如$\ln y = wx + b$。形式上仍然是线性的，但实质上是在试图让e^{wx+b}逼近y，因此是输入空间（特征）通过非线性函数映射到输出空间（预测项）。

广义线性模型

$g(y) = wx + b$

$g(\cdot)$必须单调可微

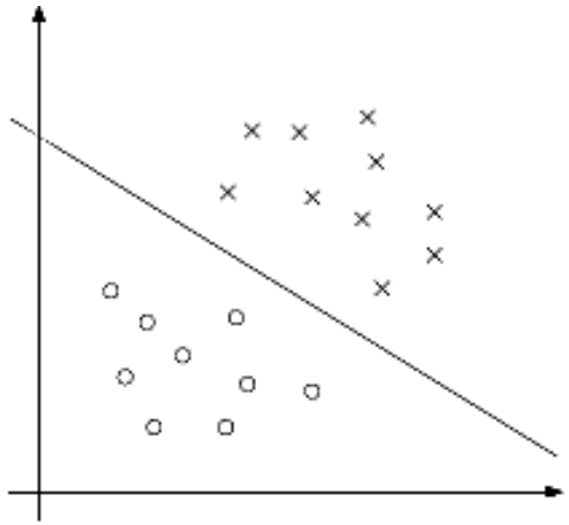
$y = g^{-1}(wx + b)$

线性对数回归的主要用途：经济计量学。取对数意味着原被解释变量对解释变量的弹性。何时取对数的经验：

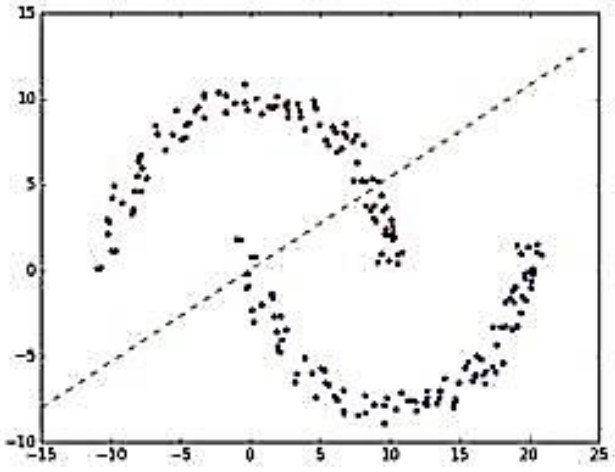
1. **与市场价值相关**的，例如，价格、销售额、工资等都可以取对数；
2. **以年度量的变量**，如受教育年限、工作经历等通常不取对数；
3. **比例变量**，如失业率、参与率等，两者均可

- 伍德里奇:《计量经济学导论》

线性分类器

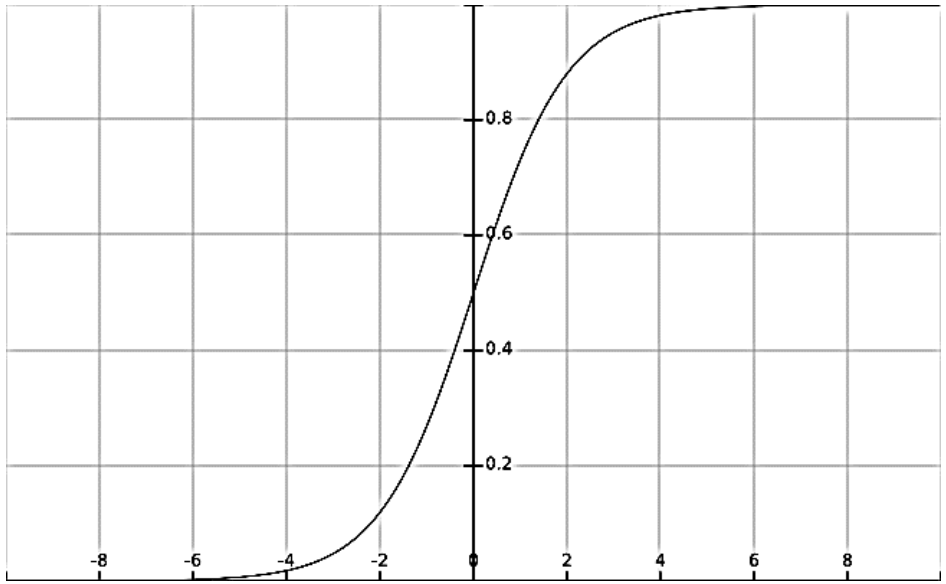


某些数据集线性可分



某些数据集线性不可分

线性分类器 – 对数几率回归（逻辑斯蒂回归）

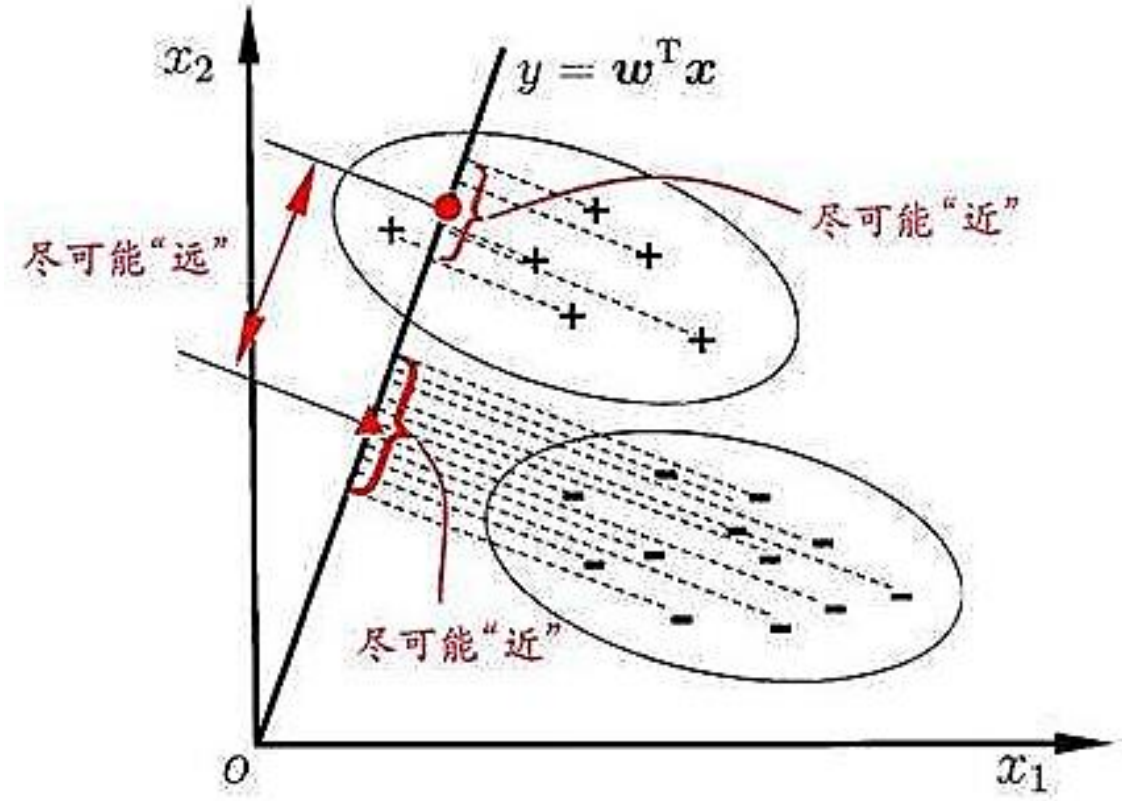


对于二分类问题，其输出标记通常为$y \in \{0,1\}$（例如“好瓜”或“坏瓜”）。需要将回归得到的连续值预测结果（例如某瓜质量评分=2）映射为2值离散结果，具体做法可以用单调可微的sigmoid函数将连续值映射至[0,1]区间。那么小于0.5可认为是好瓜，大于0.5认为是坏瓜，等于0.5则不好不坏。

$$y' = \frac{1}{1 + e^{-y}}$$

对y做了一个sigmoid函数变换，使得y的区间在0-1之间

线性分类器 – 线性判别分类LDA



最简线性分类器 – 决策树桩

身高	体重	性别
170	55	男
160	45	女
180	65	男
165	50	女
170	50	女

1. 决定一个阈值；
2. 大于这个阈值的是第一类，小于这个阈值的是第二类。
3. 例如用体重（特征）预测性别（预测项），则决策树桩需要学习出：“>50为男性，否则为女性”（决策树桩通常为一维线性模型）

其他线性分类器

- 线性SVM
- 感知机
- …

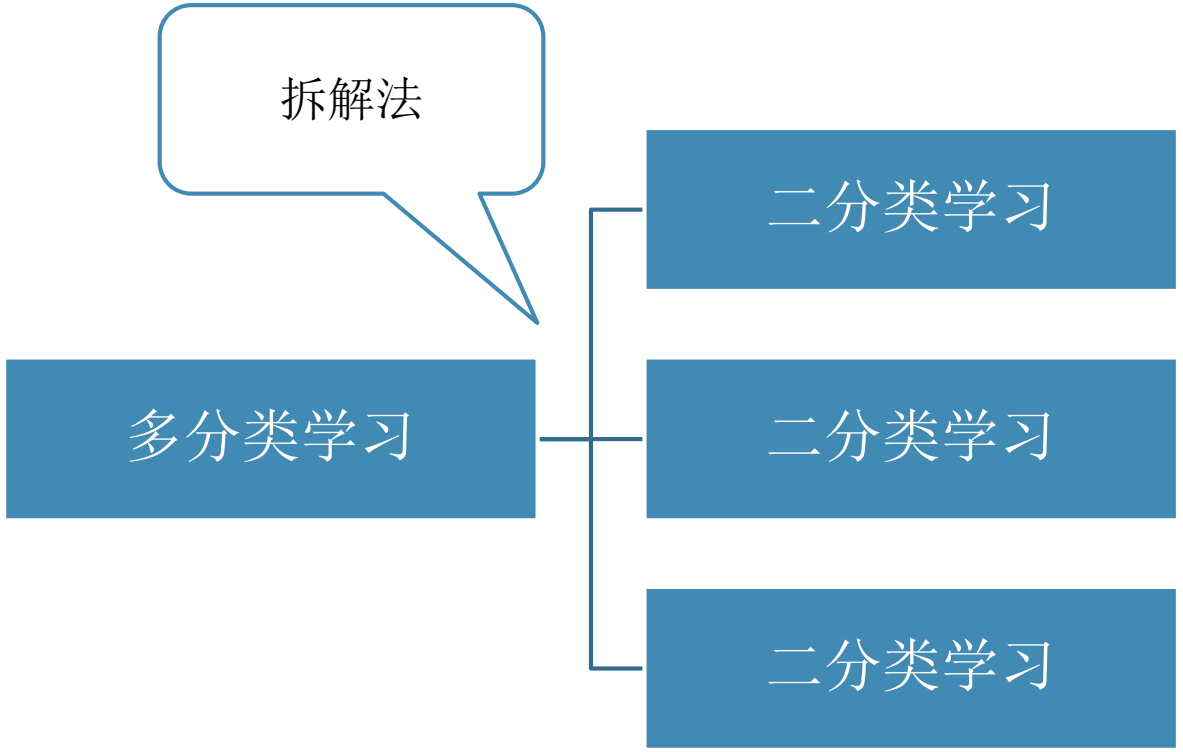
多分类学习

蓝天

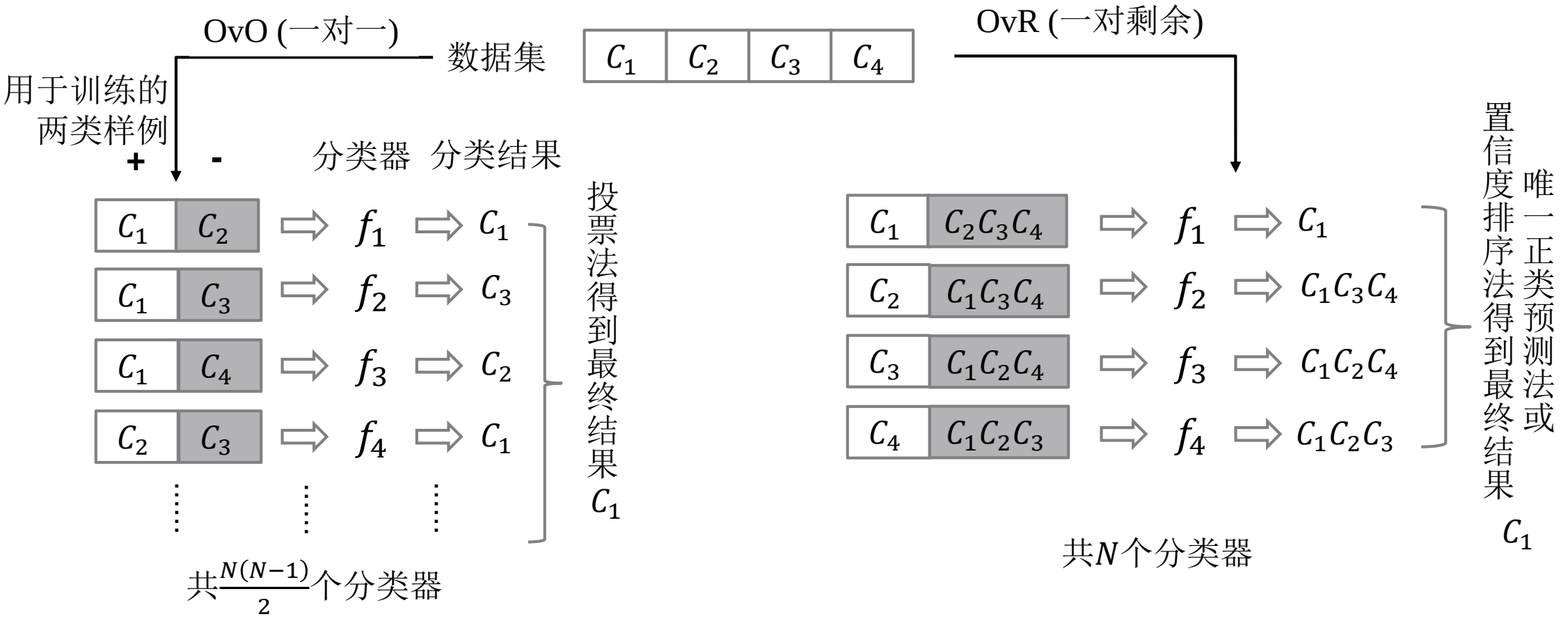
草地

羊群

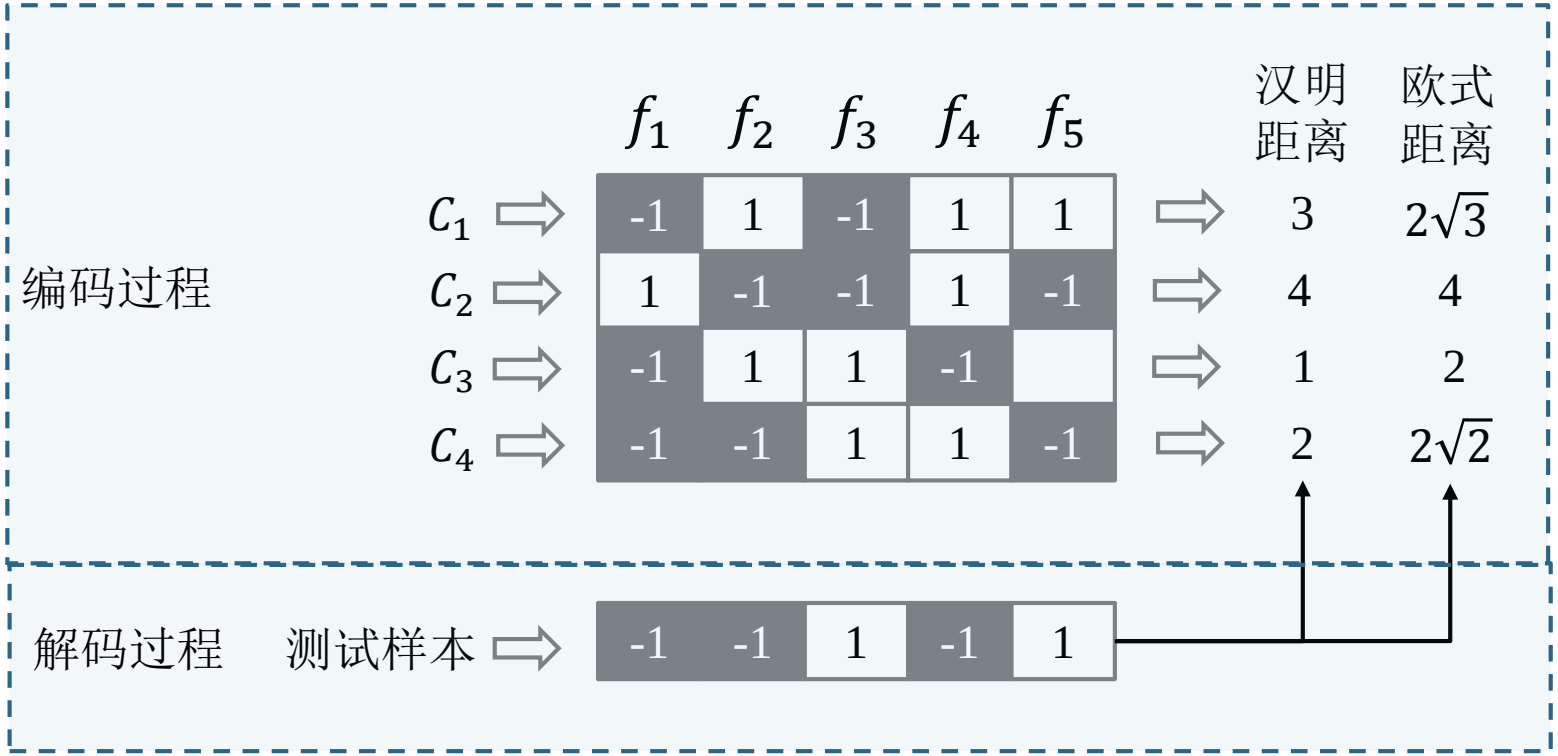
多分类学习



多分类学习的拆解策略：OvO | OvR



多分类学习的拆解策略：MvM



二元ECOC码

优点：能够容忍单个分类器的错误，ECOC码越长，纠错能力越强。

多分类学习≠多标签学习

蓝天

草地

羊群

类别不平衡问题

问题：如果训练样本中正反样本数量差别很大，例如有998个反例，但正例只有2个，那么学习方法只需要返回永远将新样本预测为反例的学习器，就能达到99.8%的精度。但这毫无意义。

假设：训练集是对真实样本总体的无偏采样。即拿来训练的“好瓜”与“坏瓜”的比例和市场所有西瓜中的比例基本一致。

再缩放 (Rescaling)

利用线性回归进行二分类时，y经历了sigmoid后表示的是样本属于正例的概率

$$y = w^{\mathrm{T}}x + b$$

⬇ 二分类中

$$若\ \frac{y}{1-y} > 1\ 则预测为正例$$

⬇ 调整为

训练集中正例个数

$$若\ \frac{y}{1-y} > \frac{m^+}{m^-}\ 则预测为正例$$

训练集中负例个数

⬇ 等价于

y'是对y的再缩放结果，使得分类的阈值与观测值中的正负例比例保持一致。

$$若\ \frac{y'}{1-y'} = \frac{y}{1-y} \times \frac{m^-}{m^+} > 1\ 则预测为正例$$

无偏采样不成立的情况

欠采样

- 对训练集中的反例采样
- 去除一些反例使得正反例数目接近
- 时间开销小

过采样

- 对训练集中的正例采样
- 增加正例的采样使得使得正反例数目接近
- 时间开销大
- 不能重复采用造成过拟合

阈值移动

- 使用原始训练集，不采样
- 预测后，采用再缩放
- $\frac{m^-}{m^+}$ 应当尽量使用真实情况中的比例

实验：房屋价格估计

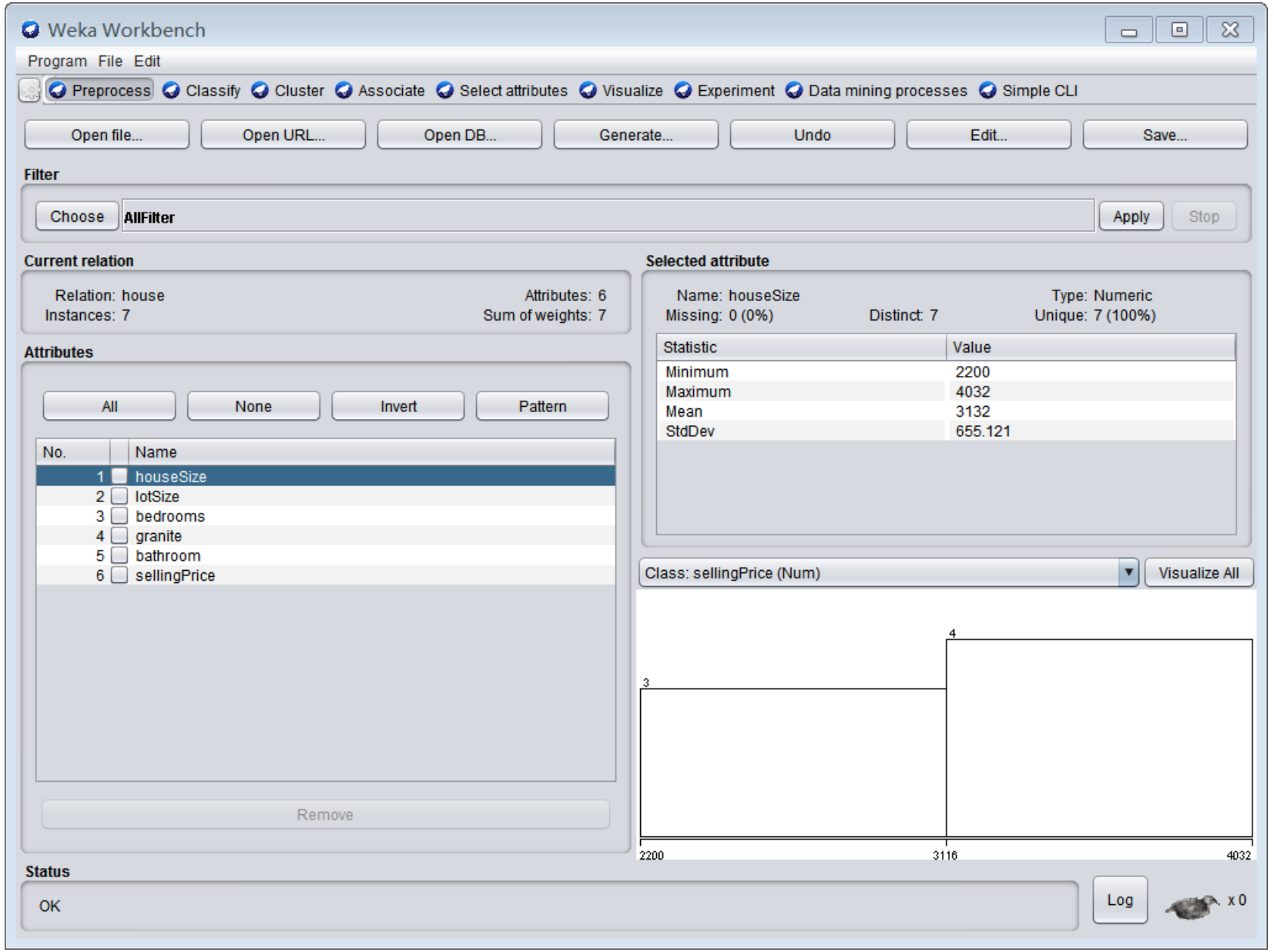
房子面积（平方英尺）	占地的大小	卧室	花岗岩	卫生间有无重装？	销售价格
3529	9191	6	0	0	$205,000
3247	10061	5	1	1	$224,900
4032	10150	5	0	1	$197,900
2397	14156	4	1	0	$189,900
2200	9600	4	0	1`	$195,000
3536	19994	6	1	1	$325,000
2983	9365	5	0	1	$230,000
3198	9669	5	1	1	????

实验环境

- Weka 3.8：https://www.cs.waikato.ac.nz/ml/weka/
- 线性回归
- https://www.ibm.com/developerworks/cn/opensource/os-weka1/

Weka数据格式

```
1  @RELATION house
2
3  @ATTRIBUTE houseSize NUMERIC
4  @ATTRIBUTE lotSize NUMERIC
5  @ATTRIBUTE bedrooms NUMERIC
6  @ATTRIBUTE granite NUMERIC
7  @ATTRIBUTE bathroom NUMERIC
8  @ATTRIBUTE sellingPrice NUMERIC
9
10 @DATA
11 3529,9191,6,0,0,205000
12 3247,10061,5,1,1,224900
13 4032,10150,5,0,1,197900
14 2397,14156,4,1,0,189900
15 2200,9600,4,0,1,195000
16 3536,19994,6,1,1,325000
17 2983,9365,5,0,1,230000
```


Weka Workbench
Program File Edit
Preprocess Classify Cluster Associate Select attributes Visualize Experiment Data mining processes Simple CLI
Open file... Open URL... Open DB... Generate... Undo Edit... Save...
Filter
Choose AllFilter Apply Stop
Current relation
Relation: house
Instances: 7
Attributes: 6
Sum of weights: 7
Selected attribute
Name: houseSize
Missing: 0 (0%)
Distinct: 7
Type: Numeric
Unique: 7 (100%)
Statistic Value
Minimum 2200
Maximum 4032
Mean 3132
StdDev 655.121
Attributes
All None Invert Pattern
No. Name
1 houseSize
2 lotSize
3 bedrooms
4 granite
5 bathroom
6 sellingPrice
Remove
Class: sellingPrice (Num)
Visualize All
4
3
2200
3116
4032
Status
OK
Log
x 0

Linear Regression Model

sellingPrice =

-26.6882 * houseSize +
7.0551 * lotSize +
43166.0767 * bedrooms +
42292.0901 * bathroom +
-21661.1208

问题1：

花岗岩去哪儿了？

问题一答案：

花岗岩对于线性回归拟合房价不起作用，也就是说Weka分析出有无花岗岩不影响房价。Weka会自动过滤掉R^2检验值较低的属性。

问题二：

精装修的卫生间会提高房价吗？

问题二答案

会。一个精装修的卫生间值$42,295。

问题三：

为什么房子面积越大越不值钱？

```
=== Summary ===

Correlation coefficient                  0.9945
Mean absolute error                   4053.821
Root mean squared error               4578.4125
Relative absolute error                 13.1339 %
Root relative squared error             10.51   %
Total Number of Instances                7
```

课后习题

通过汽缸、排量、马力、重量、加速度、年份、产地及制造商预测汽车油耗

https://www.ibm.com/developerworks/cn/opensource/os-weka1/

谢 谢