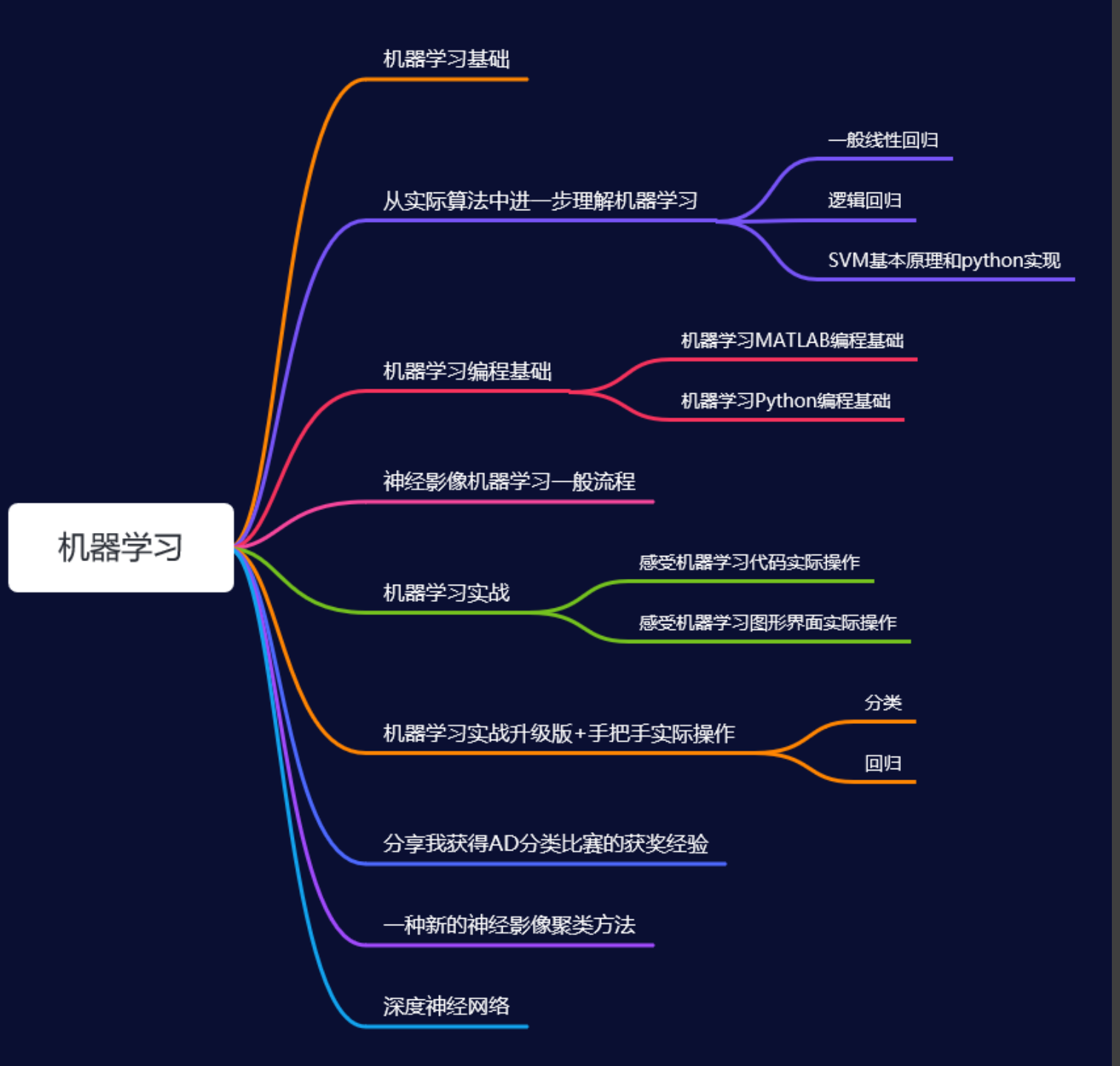
第一届机器学习实战班

黎超

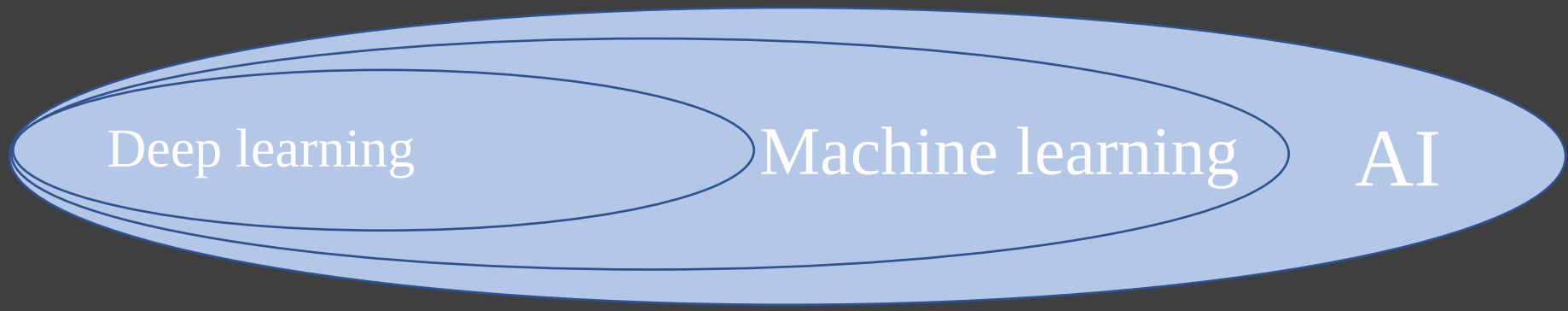
- 医学博士
- 拥有7年的机器学习与10万行以上代码编程经验
- 掌握Python与MATLAB编程
- 开发了机器学习的图像界面软件
- 已发表多篇SCI论文
- Email:lichao19870617@163.com

个人简介

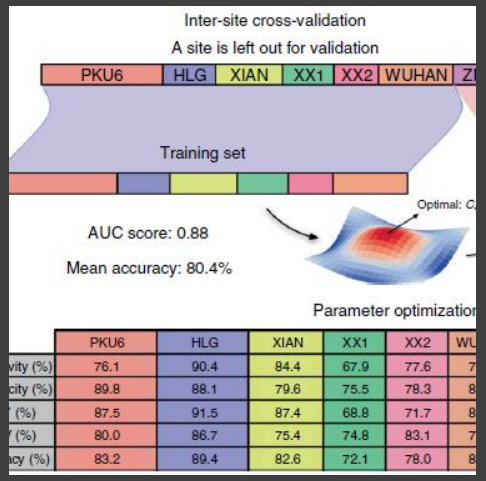
- 医学博士
- 拥有7年的机器学习与10万行以上代码编程经验
- 掌握Python与MATLAB编程
- 开发了机器学习的图像界面软件
- 已发表多篇SCI论文

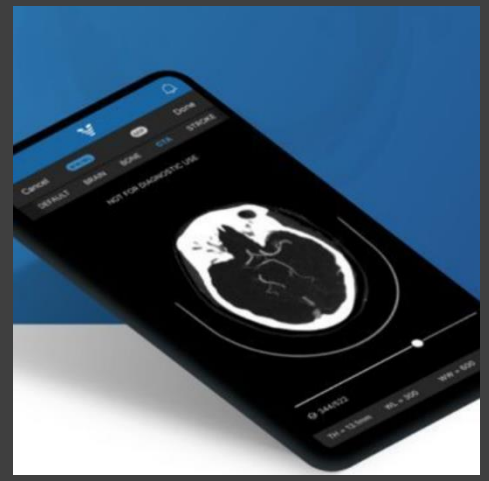

机器学习
机器学习基础
从实际算法中进一步理解机器学习
一般线性回归
逻辑回归
SVM基本原理和python实现
机器学习编程基础
机器学习MATLAB编程基础
机器学习Python编程基础
神经影像机器学习一般流程
机器学习实战
感受机器学习代码实际操作
感受机器学习图形界面实际操作
机器学习实战升级版+手把手实际操作
分类
回归
分享我获得AD分类比赛的获奖经验
一种新的神经影像聚类方法
深度神经网络

为什么要做这个课程?


Deep learning
Machine learning
AI

机器学习能做什么?





精神疾病客观诊断

机器学习算法能以超过80%的准确度诊断精神分裂症患者【1】

肿瘤的快速自动诊断

人工智能诊断脑瘤的能力与我们医院放射科主任的符合率达到99%！超越了绝大多数医生的水平

使中风患者更快得到精准治疗

2018年2月，总部位于旧金山的医疗保健公司Viz.ai宣布获得FDA对其脑卒中护理应用的营销授权。该应用提供临床决策支持，使用深度学习算法自动分析CT神经图像，以检测与脑卒中相关的指标

【1】A neuroimaging biomarker for striatal dysfunction in schizophrenia

研究人员学会使用机器学习的必要性

AI不会替代人类，但会淘汰不会使用AI的工作者

- 未来要求我们使用人工智能的行业场景会越来越多，甚至渗透到每个领域
- 脑功能研究领域的未来一定会朝着个体精准医学的方向发展，机器学习是其不可或缺的一种重要工具
- 几乎每一篇优秀的影像组学的研究都要涉及到机器学习的方法

当前发表的机器学习研究的方法问题

- 机器学习流程不规范，甚至是错误的
- 过拟合导致结果夸大

为什么不能根据测试集的
来反复调试训练参数?

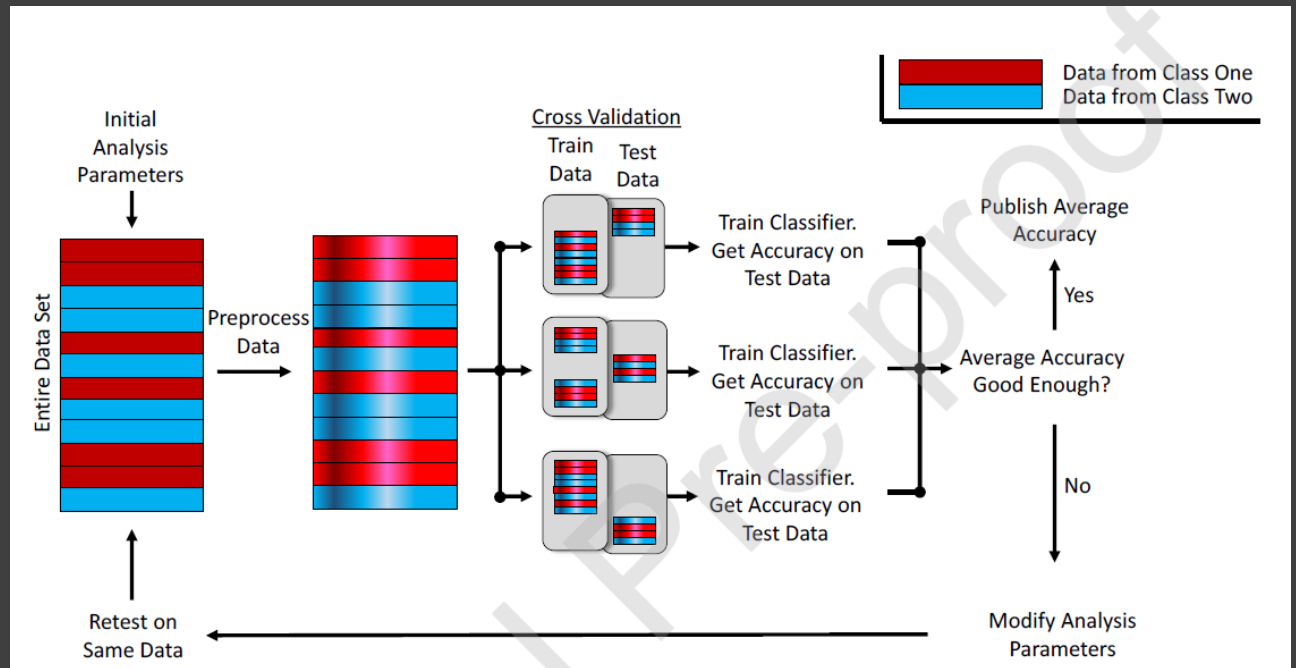


Figure 1. An example of how over-hyping can be induced by modifying hyperparameters after evaluating a system through cross-validation. The feedback loop allowing hyperparameters to be adjusted after viewing the results provides a route for analysis decisions to be made in response to the noise in the data set, despite the separation of data into training/testing sets.

I tried a bunch of things: The dangers of unexpected overfitting in classification OF BRAIN DATA

当前机器学习培训班的缺陷

- 不符合实际应用场景：没有按照标准的机器学习流程来讲解机器学习的应用
- 没有适合零编程基础学员的用户友好型图形界面软件：要求使用者编写代码，容易出错，纠错耗费大量宝贵的时间和精力

本课程的适宜人群?

- 所有从事脑功能及影像组学研究的人员
- 包括但不限于认知神经科学、心理学和神经影像学学生和科研人员
- 放射科、核医学科、精神医学科、心理科、神经内科、康复医学科的医生和医学生
- 患有“编程恐惧症”的且时间有限，想重点关注研究的设计和文章写作的医生和医学生
- 渴望在图形界面软件上，通过“点点点”的方式就能轻松完成以往必须进行复杂的编程才能完成的机器学习的群体
- 希望在“一站式”机器学习软件上轻松得到用于发表高水平SCI文章的结果报告和图片的群体

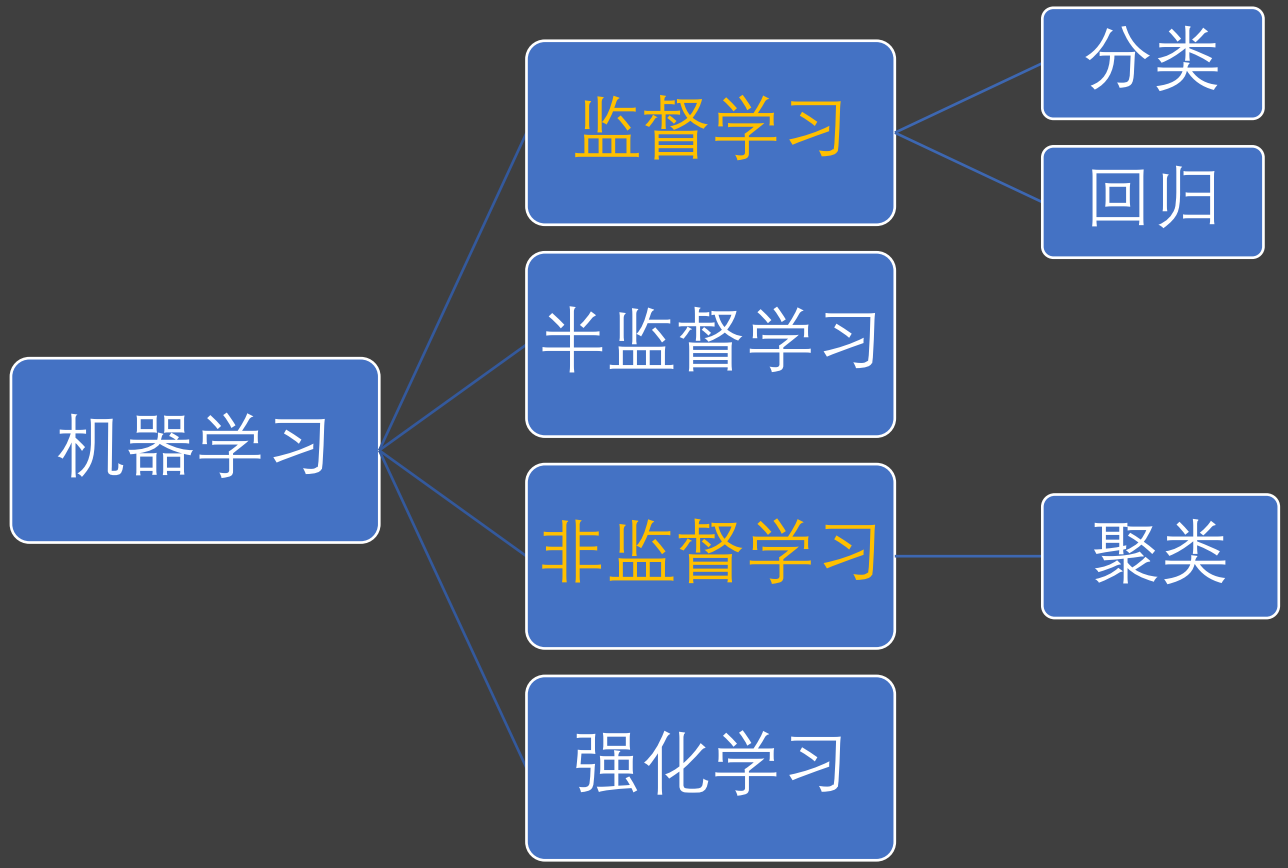
本课程的亮点?

- 以浅显易懂的方式讲解机器学习的基本理论知识，照顾零基础学员
- 以机器学习实际应用为核心，详细讲解每一个步骤的目的和实际操作
- 讲课老师讲解自己团队开发的机器学习软件，可以优先根据学员的共性要求升级软件
- 正真“零编程”实现机器学习
- 有一定基础的学员可以同时配合代码，做更加灵活的机器学习设计
- 课后微信群长期支持服务

学完课程收获

- 从概念上了解机器学习
- 学会使用脑影像或其他数据对疾病进行诊断、鉴别诊断
- 学会使用脑影像数据预测被试的脑龄、智力或其他连续性变量
- “零编程”基础学员通过简单的点击便可以完成复杂的机器学习任务
- 有一定编程基础的学员可以灵活的实现机器学习任务

机器学习基础


机器学习
监督学习
半监督学习
非监督学习
强化学习
分类
回归
聚类

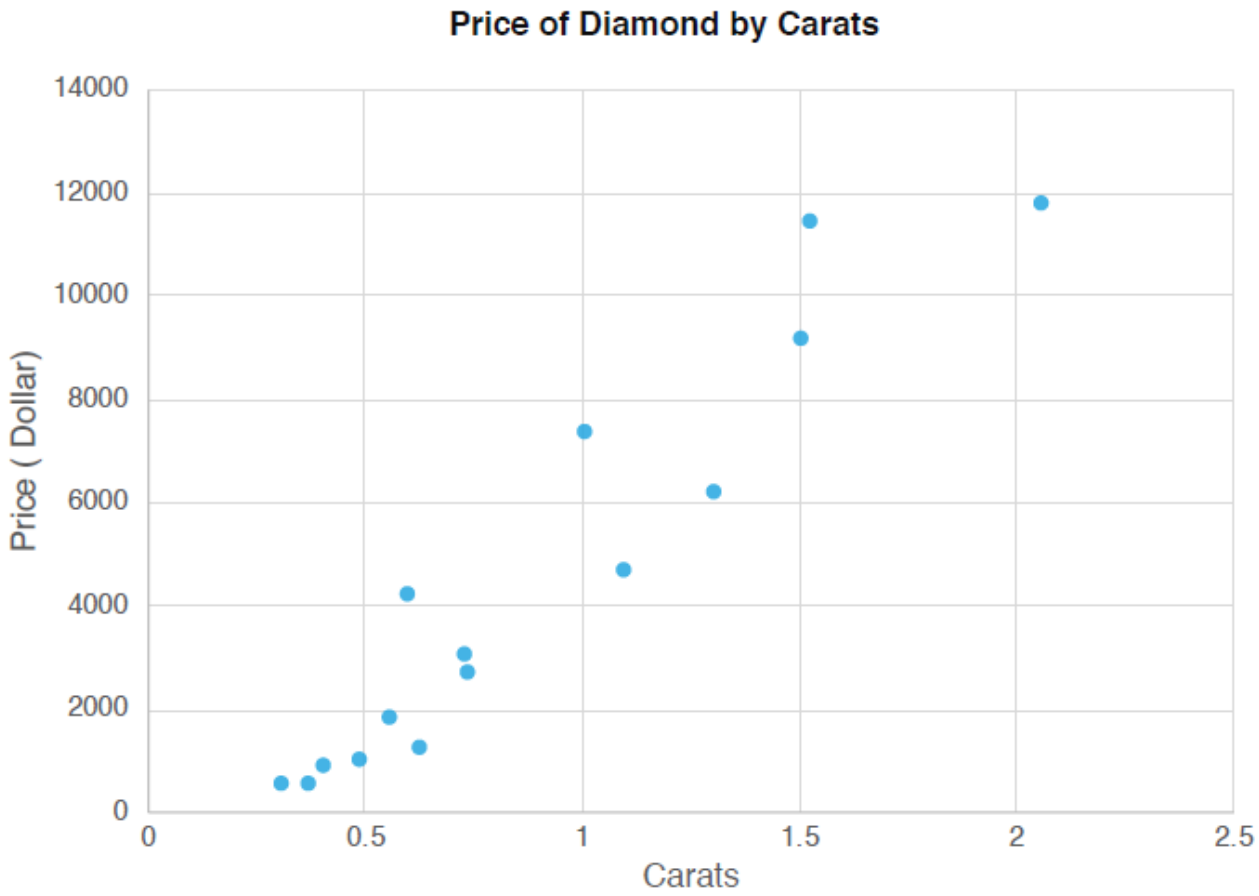
:: Example - How to predict an answer with simple model

Example: predict the price of a diamond - *continued*

Price of diamonds in jewelry store

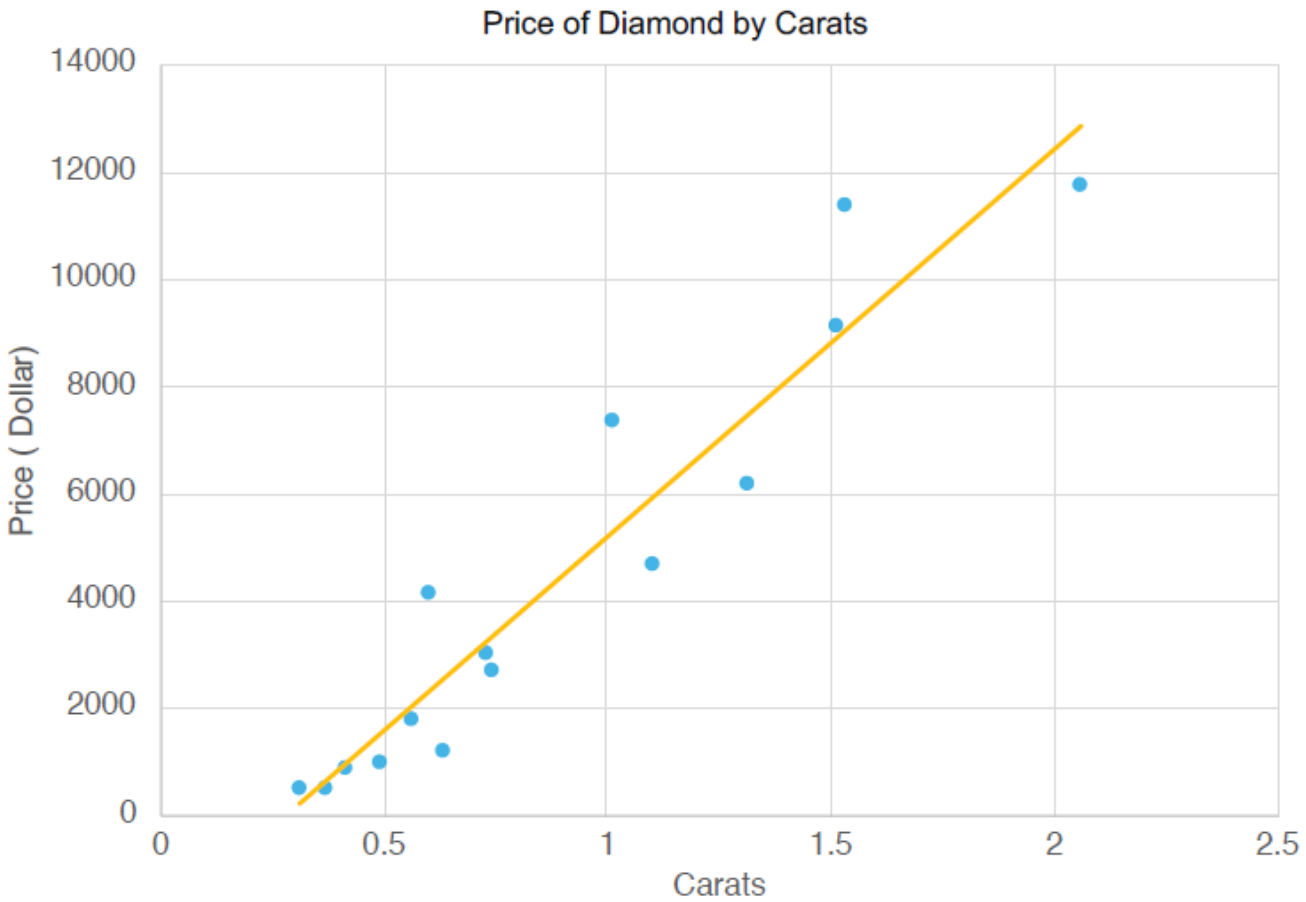
Carats	Price ($)
1.01	7,366
0.49	985
0.31	544
1.51	9,140
0.37	493
0.73	3,011
1.53	11,413
0.56	1,814
0.41	876
0.74	2,690
0.63	1,190
0.6	4,172
2.06	11,764
1.1	4,682
1.31	6,172

Plot the data



:: Example - How to predict an answer with simple model

Predict the price of a diamond by Linear regression



Linear regression is an approach for modeling the relationship between a continuous dependent variable y and one or more predictors X

By drawing a line, we created a model

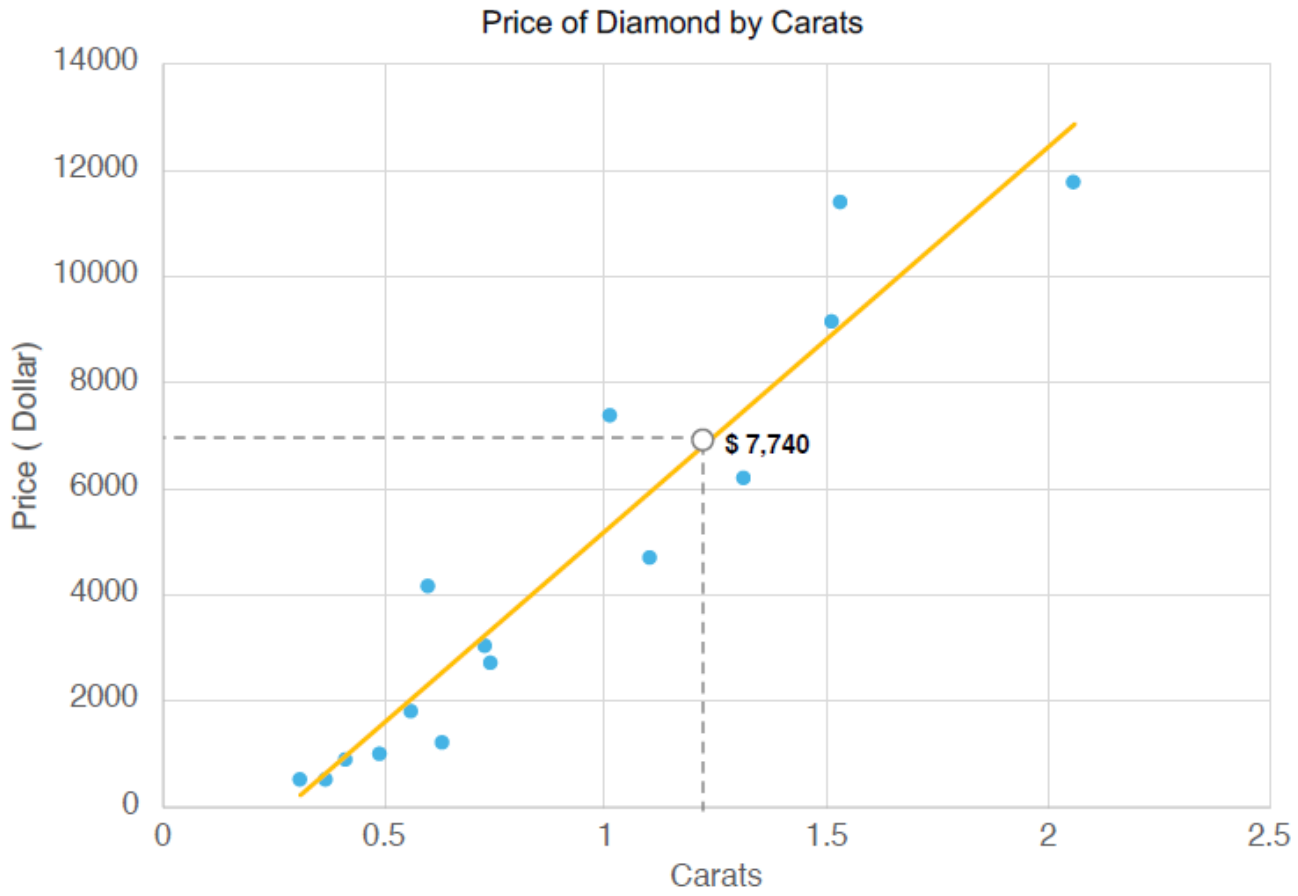
The fact that all the dots don't go exactly through the line is OK. Data scientists explain this by saying that there's the model - that's the line - and then each dot has some noise or variance associated with it.

There's the underlying perfect relationship, and then there's the gritty, real world that adds noise and uncertainty.

Because we're trying to answer the question How much? this is called a regression. And because we're using a straight line, it's a linear regression.

:: Example - How to predict an answer with simple model

Predict the price of a diamond by Linear regression



Linear regression is an approach for modeling the relationship between a continuous dependent variable y and one or more predictors X

Use the model to find the answer

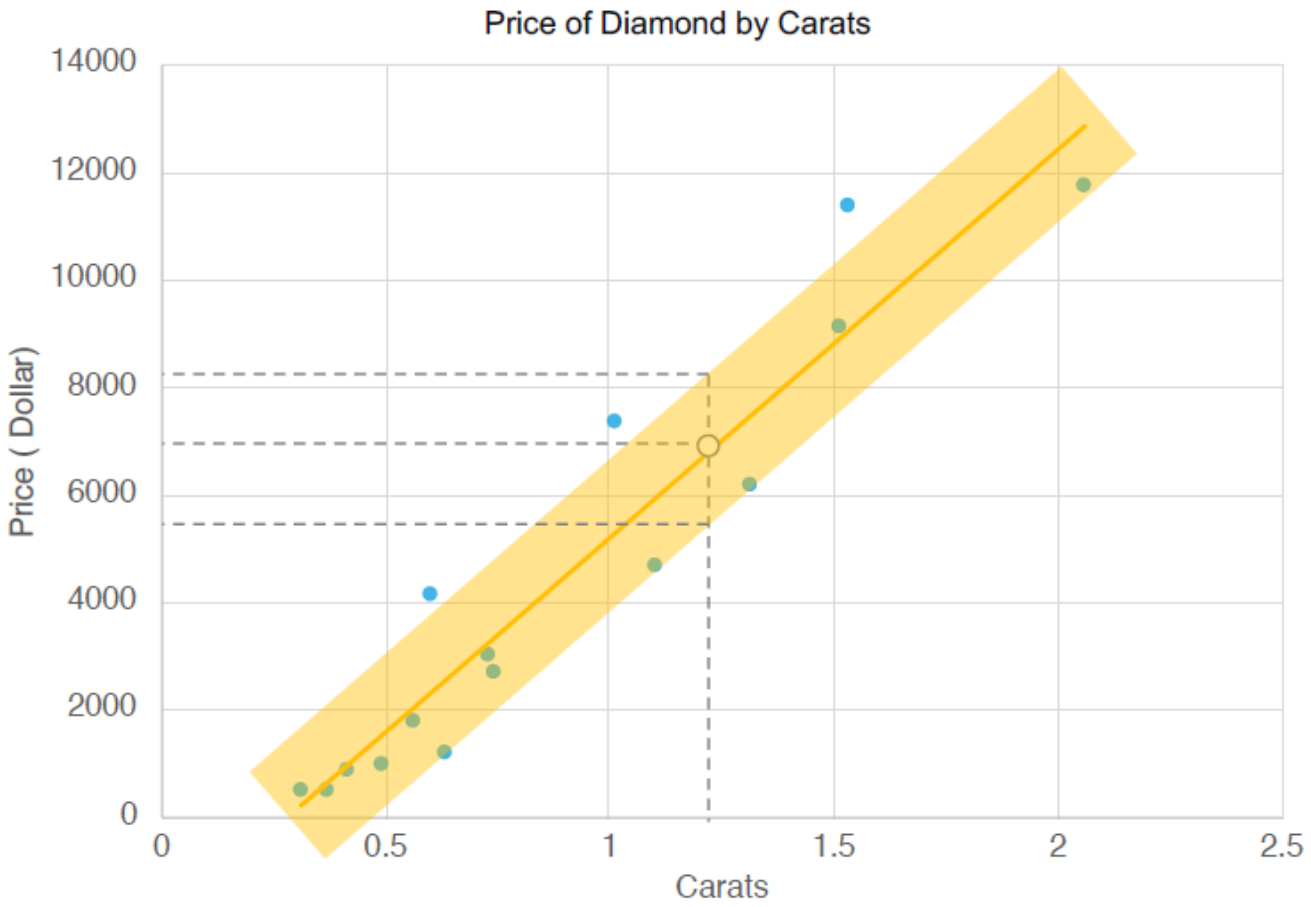
Now we have a model and we ask it our question: How much will a 1.35 carat diamond cost?

To answer our question, we eyeball 1.35 carats and draw a vertical line. Where it crosses the model line, we eyeball a horizontal line to the dollar axis. It hits right at $ 7740.

Boom! That's the answer: A 1.35 carat diamond costs about $7740.

:: Example - How to predict an answer with simple model

Predict the price of a diamond by Linear regression



Linear regression is an approach for modeling the relationship between a continuous dependent variable y and one or more predictors X

Create a confidence interval

It's natural to wonder how precise this prediction is. It's useful to know whether the 1.35 carat diamond will be very close to $7400, or a lot higher or lower.

To figure this out, let's draw an envelope around the regression line that includes most of the dots. This envelope is called our confidence interval: We're pretty confident that prices fall within this envelope, because in the past most of them have. We can draw two more horizontal lines from where the 1.35 carat line crosses the top and the bottom of that envelope.

Now we can say something about our confidence interval: We can say confidently that the price of a 1.35 carat diamond is about $7,740 - but it might be as low as $6,300 and it might be as high as $9,000.

模型训练之直接优化

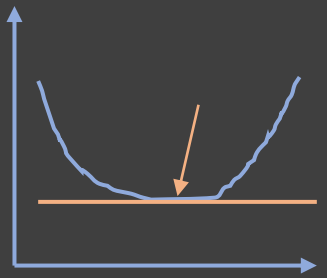
Y = W*X + b

已知X和Y

求权重W和偏置项b

$$J(w) = \frac{1}{2m}\sum_{i=1}^{m}(f(x_i) - y_i)^2$$

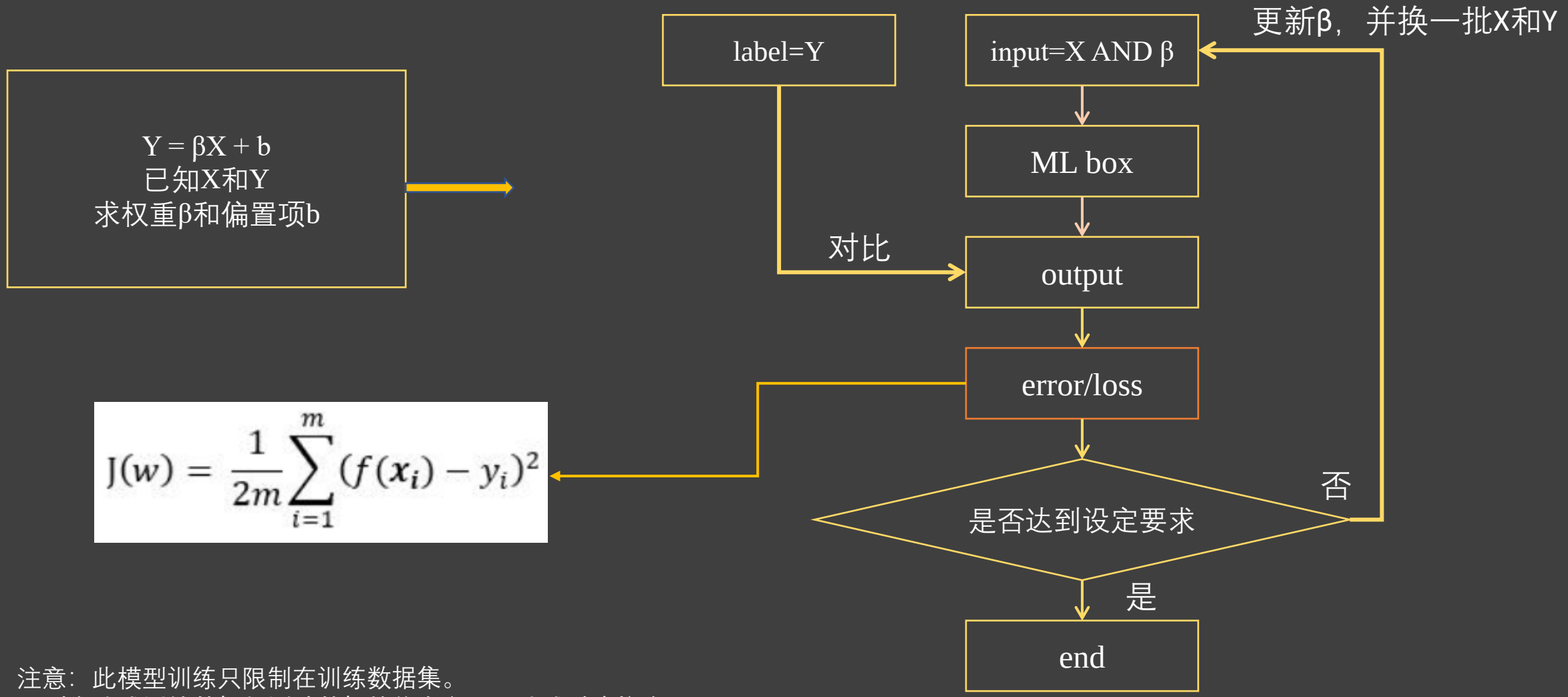
导数/偏导数为0时求得最小值



求J在w上的导数/偏导数

$$J'(\mathbf{w}) = \frac{1}{m}(\boldsymbol{Xw} - \boldsymbol{y}) \cdot \boldsymbol{X^T} = \frac{1}{m}(\boldsymbol{X^T X w} - \boldsymbol{X^T y})$$

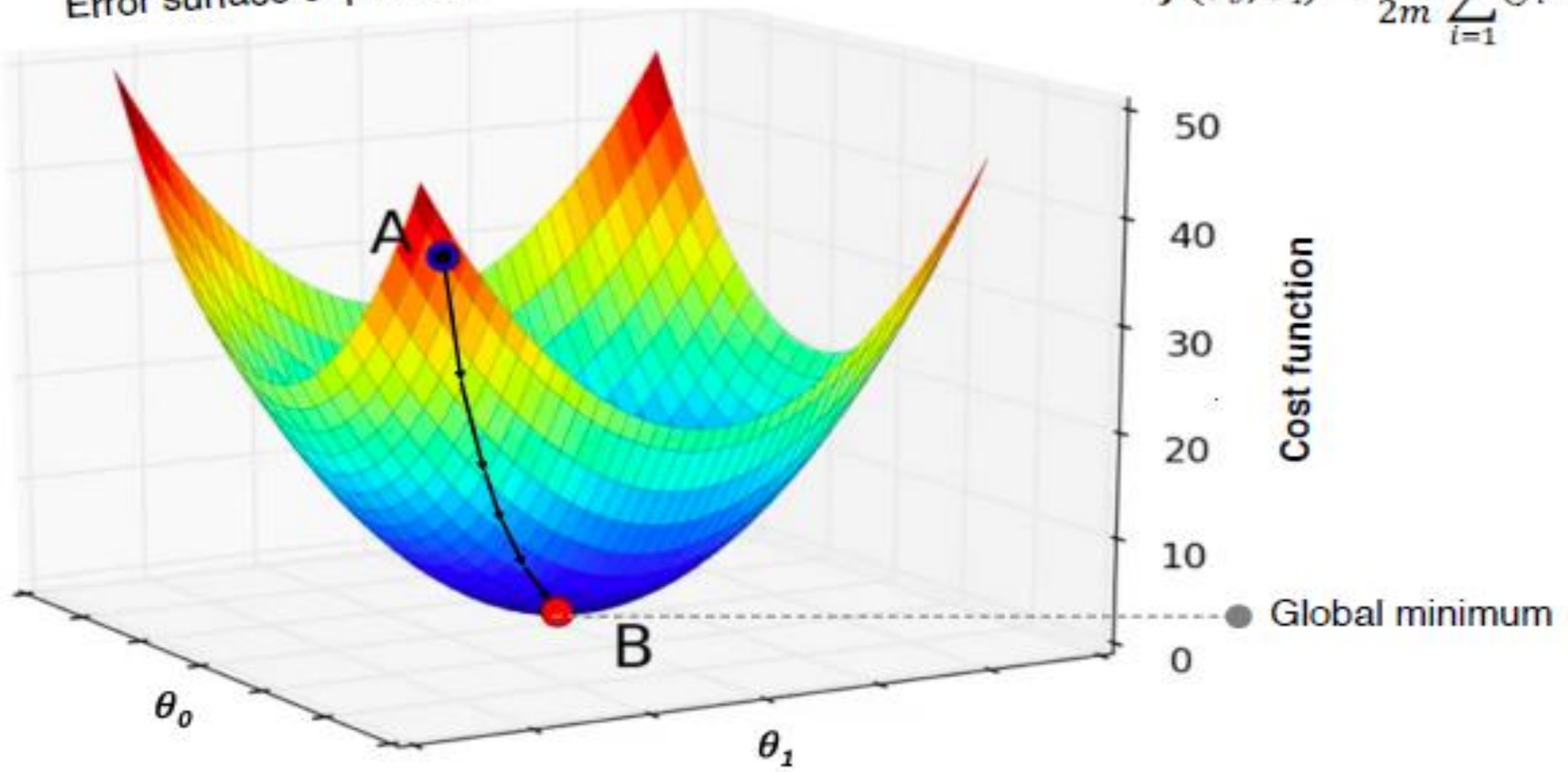
模型训练梯度下降



注意：此模型训练只限制在训练数据集。
否则会造成训练数据与测试数据的信息交互，造成过度拟合。

Error surface of predictions

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} (y_i - \hat{y}_i)^2$$

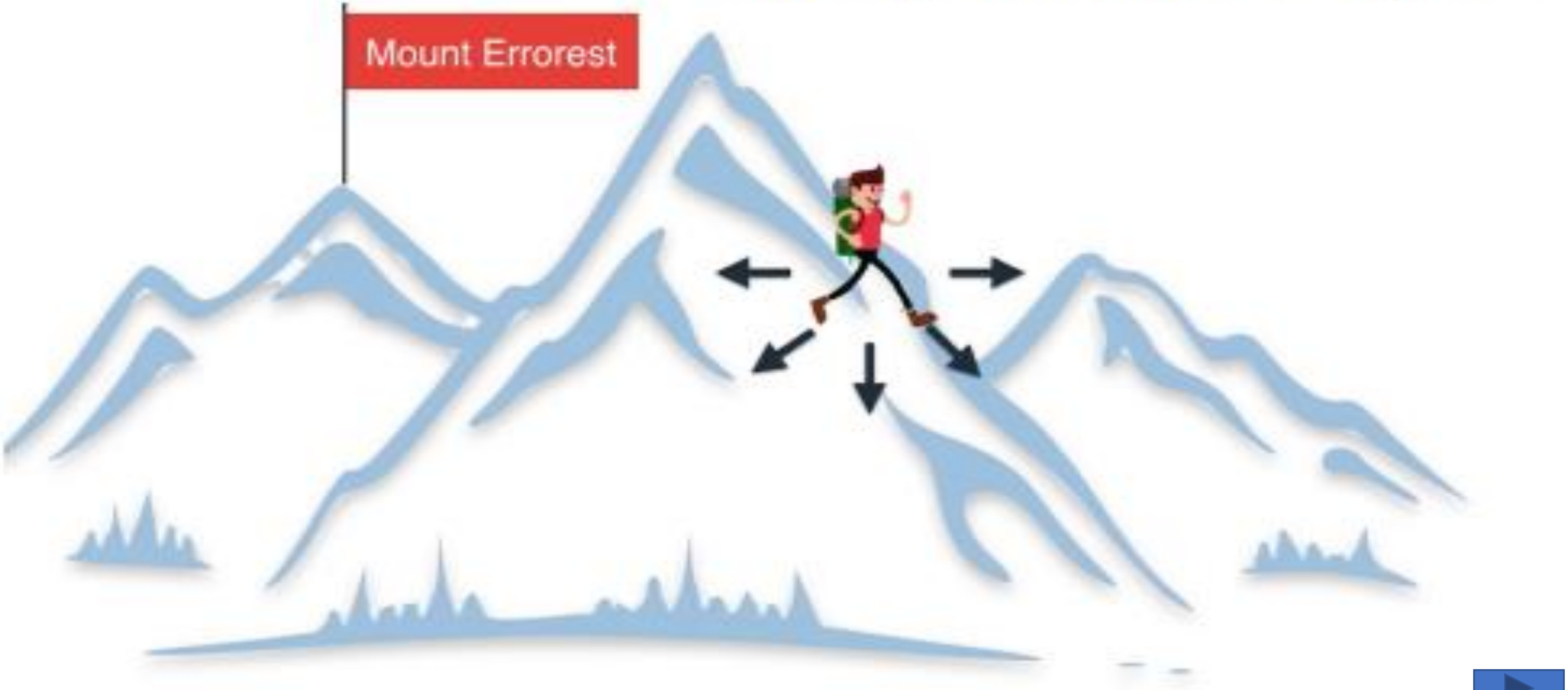


The combination of different parameters θ_0 , θ_1 leads to different error (cost function).

The error surface of the predictions corresponds to the different value of parameters.

If you're blindfolded Hiker,

how to get to the lowest place ?



:: Linear Regression

How does Gradient descent work ?

When there are multiple parameters, the gradient is a vector of partial derivatives with respect to the parameters. For the sake of simplicity, consider the cost function $J(\theta)$ that depends on only one parameter θ.

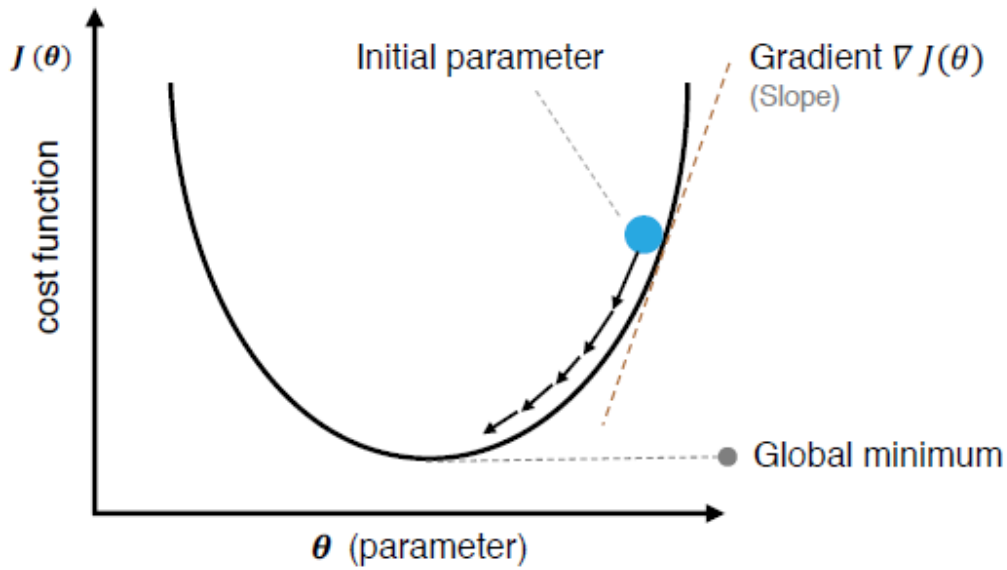
Gradient Descent algorithm

The parameter are iteratively updated in the following equation:

Learning rate (Step size)

$$\theta_{new} = \theta_{old} - \alpha \cdot \nabla_\theta J(\theta)$$

Gradient $\nabla_\theta J(\theta) = \frac{d}{d\theta} J(\theta)$



1. Pick a value for the learning rate α
2. Start with a random point θ
3. Calculate the gradient $\nabla J(\theta)$ at the point θ. Follow the opposite direction of gradient to get new parameter θ_{new}
4. Repeat until the cost function converges to the minimum

In this example, initially the slope is large and positive. So, in the update equation, θ is reduced. As θ keeps getting reduced, notice that the gradient also reduces, and hence the updates become smaller and smaller and eventually, it converges to the minimum.[1]

一般线性回归

- 实际操作
- linear_regression.ipynb

$$J(w) = \frac{1}{2m}\sum_{i=1}^{m}(f(x_i) - y_i)^2$$

$$J'(\mathbf{w}) = \frac{1}{m}(\boldsymbol{Xw} - \boldsymbol{y}) \cdot \boldsymbol{X^T} = \frac{1}{m}(\boldsymbol{X^T Xw} - \boldsymbol{X^T y})$$

- 实际操作:
- 自己动手调整初始的权重w_fit，以及学习率alpha， 观察模型训练和损失收敛情况

逻辑回归

课后作业：动手完成一个梯度下降的例子

- 根据一般线性回归的代码和如下的损失函数和损失函数对权重的导数来完成逻辑回归的训练（要求模型有两个权重[w1, w2]）

逻辑回归表达式：

$$\frac{1}{1+\exp(-\theta^{\top}\mathbf{x})}$$

$$J(\theta) = -\frac{1}{m}\left[\sum_{i=1}^{m} y^{(i)} log(h_\theta(x^{(i)})) + (1 - y^{(i)}) log(1 - h_\theta(x^{(i)}))\right]$$

$$\frac{\partial J(\theta)}{\partial \theta_j} = \frac{1}{m}\left[\sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)}\right]$$

逻辑回归实战：鸢尾花分类

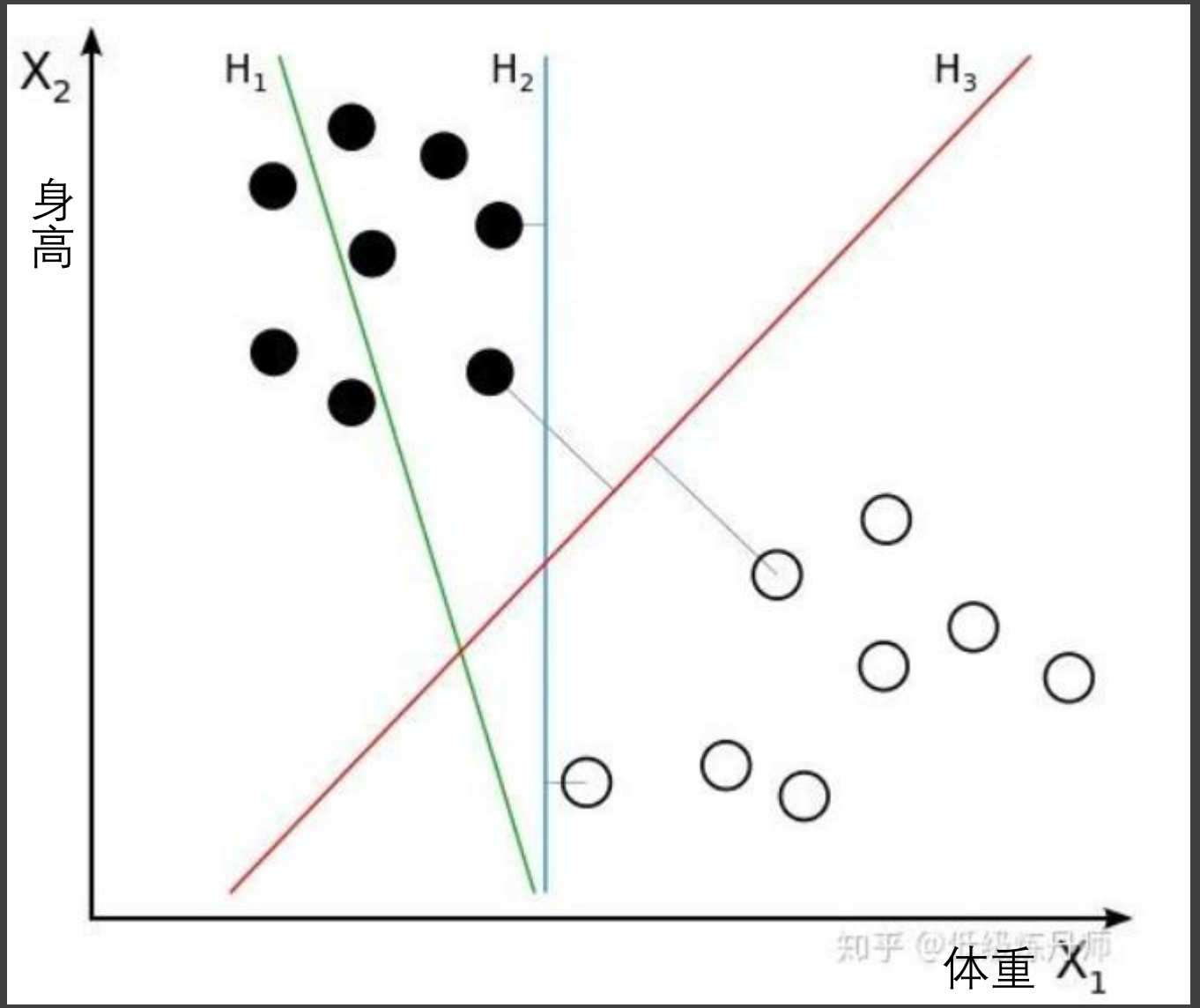
根据花瓣的长度和宽度，花萼的长度和宽度来区分setosa, versicolor两类鸢尾花

支持向量机SVM

用另一种方法来根据体重和身高来计算BMI

体质指数（BMI）=体重（kg）÷身高（m）的平方

bmi大于28为肥胖, 小于28为非肥胖



https://zhuanlan.zhihu.com/p/57648645

条件:

$$\begin{cases} \omega^T x_i + b \geq +1, & y_i = +1 \\ \omega^T x_i + b \leq -1, & y_i = -1 \end{cases} \tag{1.1}$$

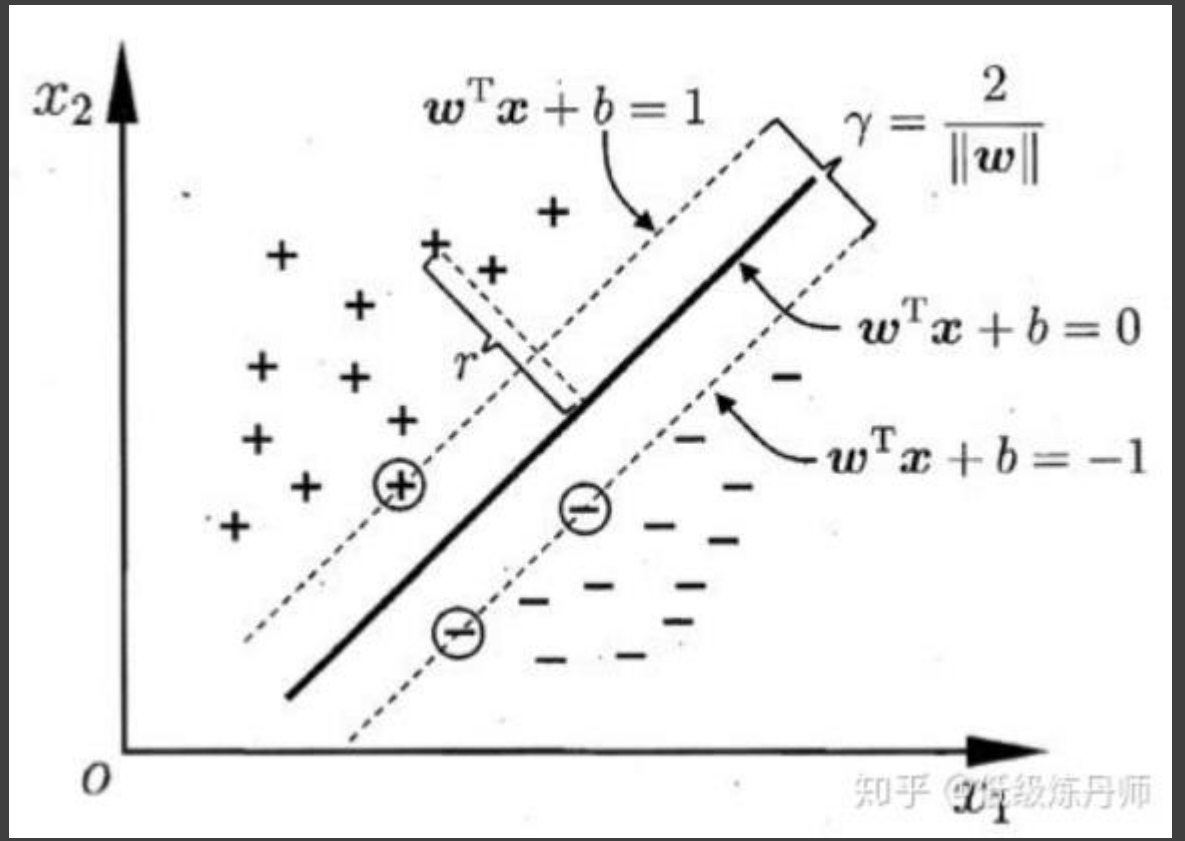
任意点到超平面（分界线）的距离:

$$r = \frac{|\omega^T x + b|}{||\omega||} \tag{1.2}$$

向量的模长

支持向量到超平面（分界线）的距离:

$$r = \frac{1}{||\omega||} \tag{1.3}$$



为什么经过支持向量的直线一定可以表示为 wx+b = 1/-1

设直线 L 的方程为Ax+By+C=0，点 P 的坐标为（x0,y0），则点 P 到直线 L 的距离为：$\frac{|Ax_0 + By_0 + C|}{\sqrt{A^2 + B^2}}$

一下三条直线是等价的:

$X1 + 2*X2 + 1/2 = 0$

$2X1 + 4*X2 + 1 = 0*2$

$2X1 + 4*X2 + 1 + 1 = 0*2 + 1$

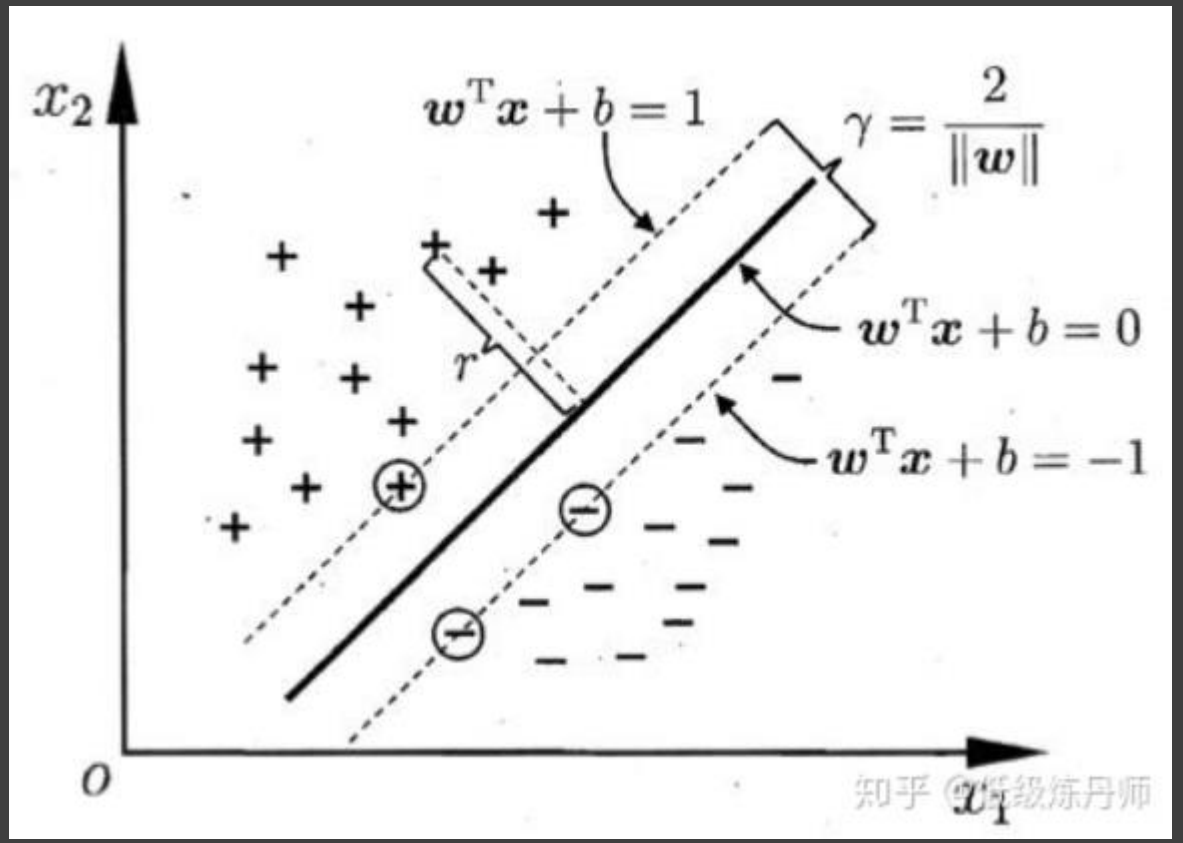
为什么经过支持向量的
直线一定可以表示为
$wx+b = 1/-1$

支持向量机的思想:
最大化支持向量到间隔的距离

$$r = \frac{1}{||\omega||} \tag{1.3}$$

等价于

$$min_{\omega,b} \frac{1}{2}||\omega||^2, s.t.\, y_i(\omega^T x_i + b) \geq 1, i = 1, 2, \ldots, m \tag{1.4}$$



$$min_{\omega,b}\frac{1}{2}||\omega||^2, s.t.\, y_i(\omega^T x_i + b) \geq 1, i = 1, 2, \ldots, m \qquad (1.4)$$

先考虑等式约束的问题，我们要使得目标函数 f(x)最小且同时满足 g(x)=0的约束。那我们可以转换目标函数为拉格朗日函数

L(x, α) = f(x) + α*g(x)

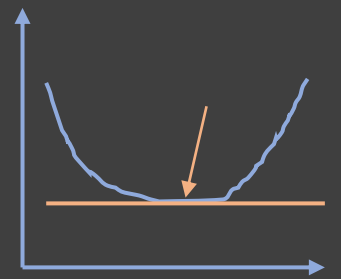
因为1.4的约束条件是不等式，所以额外引入KKT条件即可

$$\begin{cases} \alpha_i \geq 0 \\ y_i f(x_i) - 1 \geq 0 \\ \alpha_i (y_i f(x_i) - 1) = 0 \end{cases}$$

服从条件

因此SVM的最终的拉格朗日函数为如下函数，我们求解L(w,b,a)的最小值即可

$$L(\omega, b, \alpha) = \frac{1}{2}||\omega||^2 + \sum_{i=1}^{m} \alpha_i (1 - y_i(\omega^T x_i + b))$$



SVM实战

- 根据身高和体重区分BMI是否大于28，即是否肥胖

- linear_svm_bmi.ipynb

机器学习之MATLAB编程基础

常用的数据类型

- 一、数值类型
- 二、字符与字符串
- 三、结构体
- 四、单元数组/元胞
- 五、映射容器

https://www.cnblogs.com/jwg-fendi/p/10052898.html

数值类型

类型	子类型	符号	位数	用法
整形	有符号的整形	int8	8	a=int32(12)
		int16	16	
		int32	32	
		int64	64	
	无符号的整形	uint8	8	a=uint32(12)
		uint16	16	
		uint32	32	
		uint64	64	
单精度浮点		single	32	a=single(12.34)
双精度浮点		double	64	a=12.23

注：单精度保留到小数点后7位，双精度保留到15位

数值类型

- Inf：无穷数
- NaN：非数值量

- 用isinf识别Inf
- 用isnan识别NaN

数值类型

- 练习：识别并把d=[Inf, NaN]中的无穷量和非数值量赋值为1和0

字符与字符串

- 字符串： a = 'matlab'
- 字符： b = 'm'
- a(1) == b

结构体

- 描述的结构体：

- zhangsan.gender = '男';
- zhangsan.height = 175;
-
- zhangsan = struct('gender', '男', 'height', 175)

- 自己编写一个结构体，储存韩梅梅的信息，韩梅梅是女生，身高165，郑州人。

元胞

- d = {[1,2,3], 'hello’}

- 1、试试d(1)和d{1}的结果有什么区别?

映射容器

- zhangsan = containers.Map({'gender', 'height'},{'男',175});

程序控制结构之条件控制

- 学习ifelse.m

程序控制结构之分支控制

- switch case otherwise
- 学习switch1.m
- 学习switch2.m

- 练习：用switch写一段程序，改程序可以判断用户输入的是否是正数，如果是打印出‘您输入的是正数’，否则打印‘您输入的不是正数’

程序控制结构之试探结构

- try catch end
- 一般用于探试，有一定容错能力

学习try_catch.m

程序控制结构之循环控制

- 学习for_.mat
- 学习while_.mat

- 练习： 用for和while循环依次打印出'i’, ‘love’, matlab'

MATLAB编程作业1

- 用matlab语言实现linear_regression.py

- 给定知识：
- 1、生成正态分布随机数函数randn
- 2、设定随机种子点rng(seed)
- 3、求x的平方函数为x^2

- 注：遇到不会的函数可在matlab命令窗口输入 help “function” 来查看function的使用

MATLAB编程作业2

- 用matlab实现linear_svm_bmi.py
- 训练： model = fitcsvm(x_train, y_train)
- 测试： yhat = model.predict(x_test)

机器学习之Python编程基础

常用数值类型

- 整型
- 浮点型

- 练习number.py

字符串

- 练习string.py

布尔类型

- 练习bool.py

列表

- 练习list.py

元组

- 练习tuple.py

字典

- 练习dict.py

- 字典一般形式
- d = {key1 : value1, key2 : value2 }

程序控制结构之条件控制

- 练习ifelse.py

程序控制结构之探试结构

- 练习try_exception.py

程序控制结构之while循环

- 练习while.py
- 作业：用while循环求从1到100的和

程序控制结构之for循环

- 练习for.py
- 作业：用for循环求从1到100的和

csv/excel文件读取

- 练习io_csv_redwine.py
- 为后面的学习做准备

神经影像机器学习一般流程

https://www.processon.com/mindmap/5fa3c09d5653bb256348fb36

机器学习代码实际操作

红葡萄酒质量预测

- 练习redwine.py

1-固定酸度
2-挥发性酸度
3-柠檬酸
4-残糖
5-氯化物
6-游离二氧化硫
7-总二氧化硫
8-密度
9相
10-硫酸盐
11-酒精

- 练习题：根据如下公式计算模型对红葡萄酒测试集的敏感度、特异度和召唤率。

- **TPR**：true positive rate，真阳性样本在实际阳性样本中的占比，又称为敏感性。
  计算公式为：TPR=TP/ (TP+ FN)
- **FPR**：false positive rate，假阳性样本在实际阴性样本中的占比，又称为召唤率。
  计算公式为：FPR= FP / (FP + TN)
- **TNR**：true negative rate，假阳性样本在实际阴性样本中的占比，又称为特异度。
  计算公式为：TNR= TN / (FP + TN)

练习题：根据红葡萄酒代码，完成白葡萄酒的质量预测

白葡萄酒质量预测答案

- whitewine.py

机器学习软件实际操作

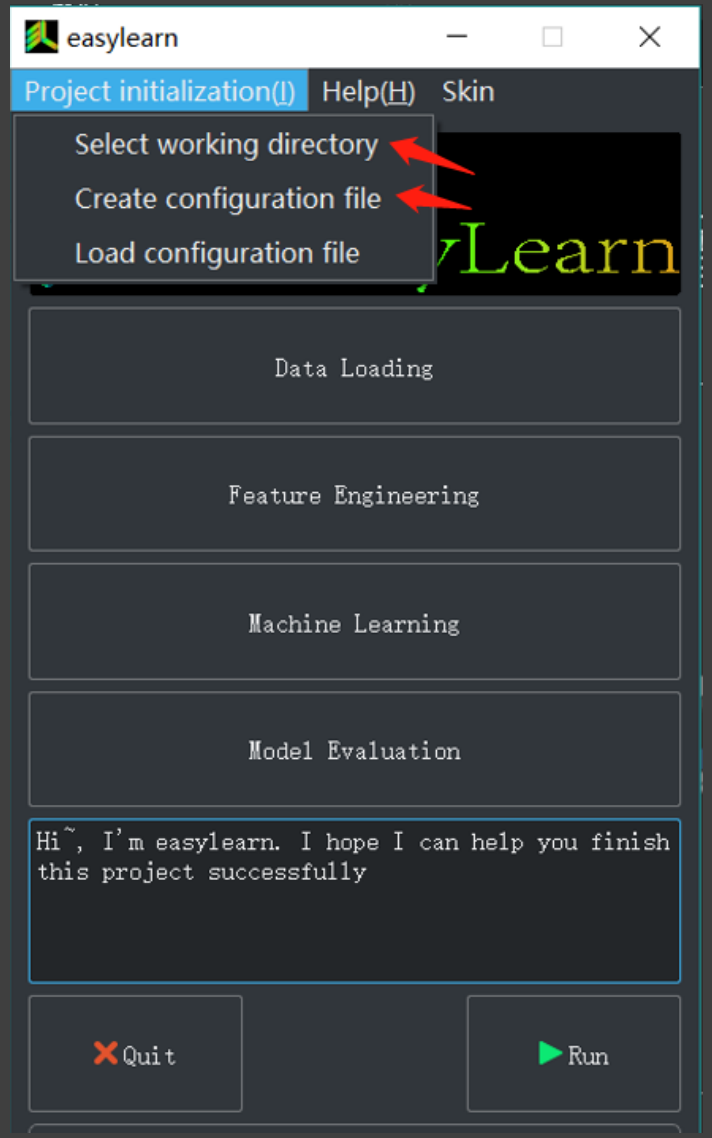
软件地址

- https://github.com/lichao312214129/easylearn

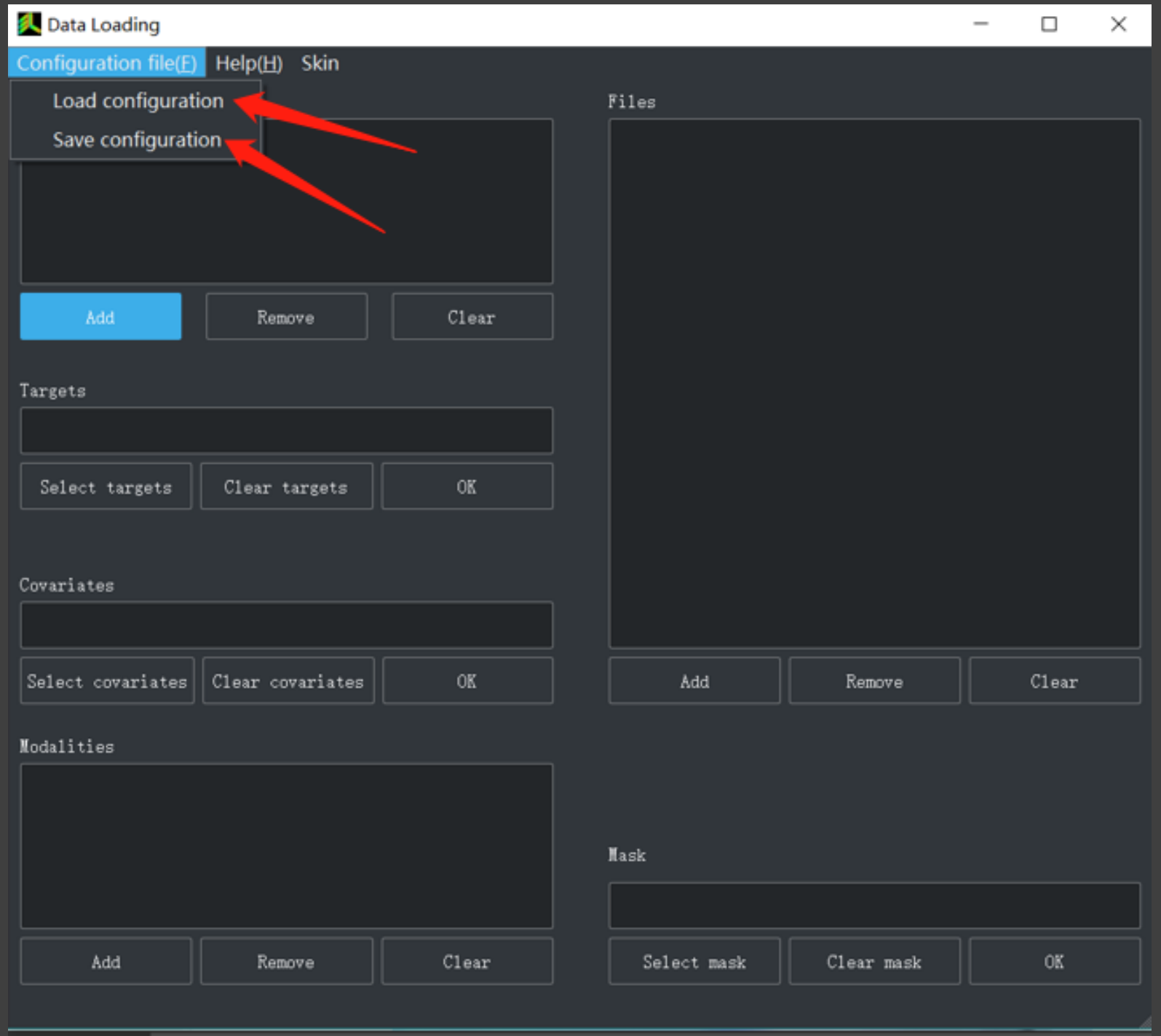
开启软件

- 两行代码开启软件界面：
- from eslearn import app
- app.run()

选择工作目录并添加配置文件



数据加载


Data Loading
Configuration file(F)
Help(H)
Skin
Load configuration
Save configuration
Files
Add
Remove
Clear
Targets
Select targets
Clear targets
OK
Covariates
Select covariates
Clear covariates
OK
Add
Remove
Clear
Modalities
Add
Remove
Clear
Mask
Select mask
Clear mask
OK

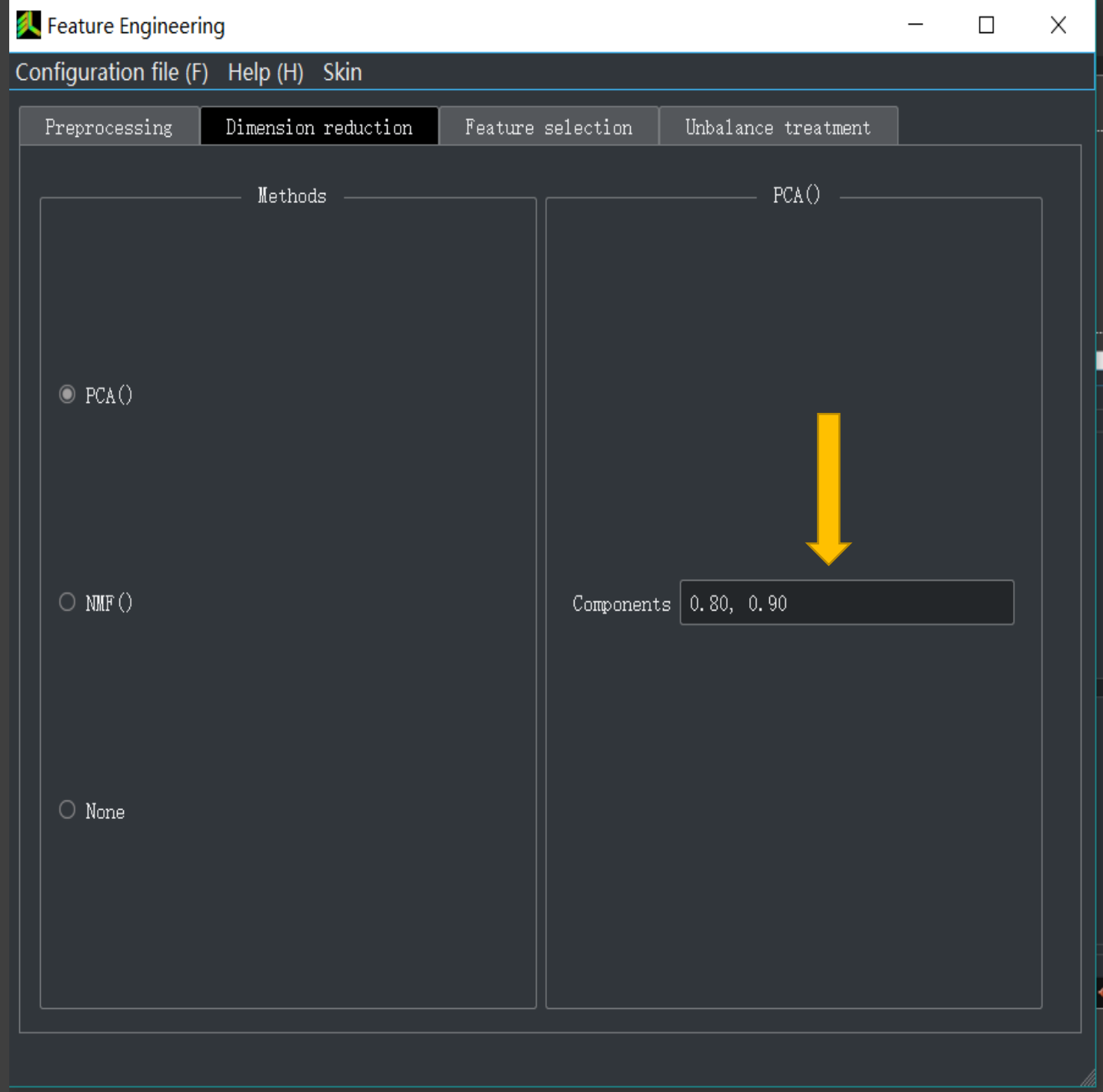
数据加载注意事项

- 注意：
- 1 如果输入Files是多个文件，即一个case一个文件，那么文件名中必须有”sub-xxx”或者”subxxx”，“xxx”是阿拉伯数字。这是为了将特征和目的匹配，特别是多模态时。如果是单个文件，则其中的数据是n_sample * （1+ n_features）的矩阵，其中有一列的名字最好是一__ID__,否则eslearn把第一列看作ID
-
- 2 如果输入的targets是文件，那么文件中数据应该有两列，名字分别是__ID__”和__Targets__。如果没有这两个列名，那么eslearn将把第一列当作是__ID__ 第2列当作是__Targets__。

- 对于covariates来说同样如此。

特征工程

- 参数可以设置多个eslearn自动通过nest-cv的方式寻找最优参数



Nest CV

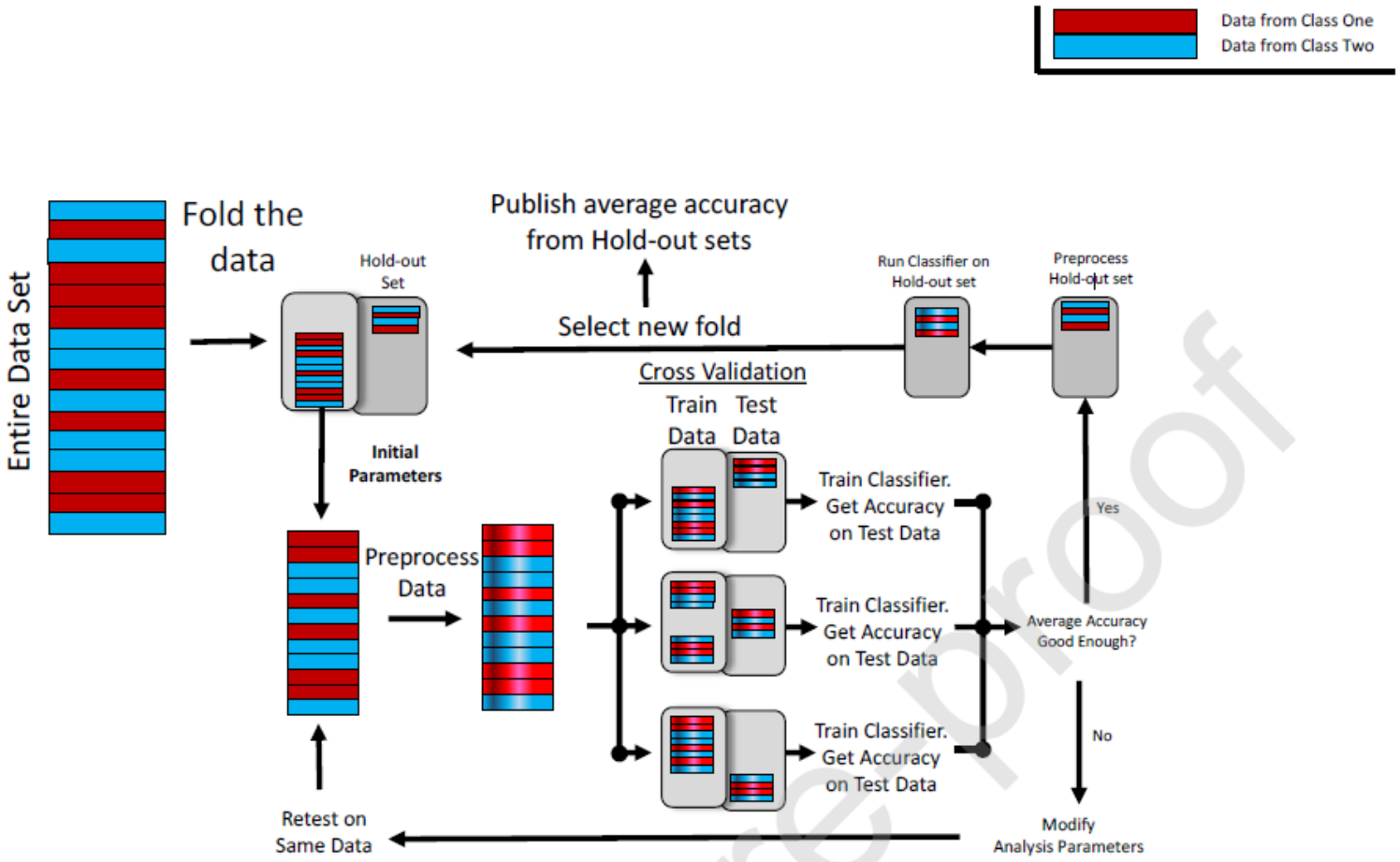
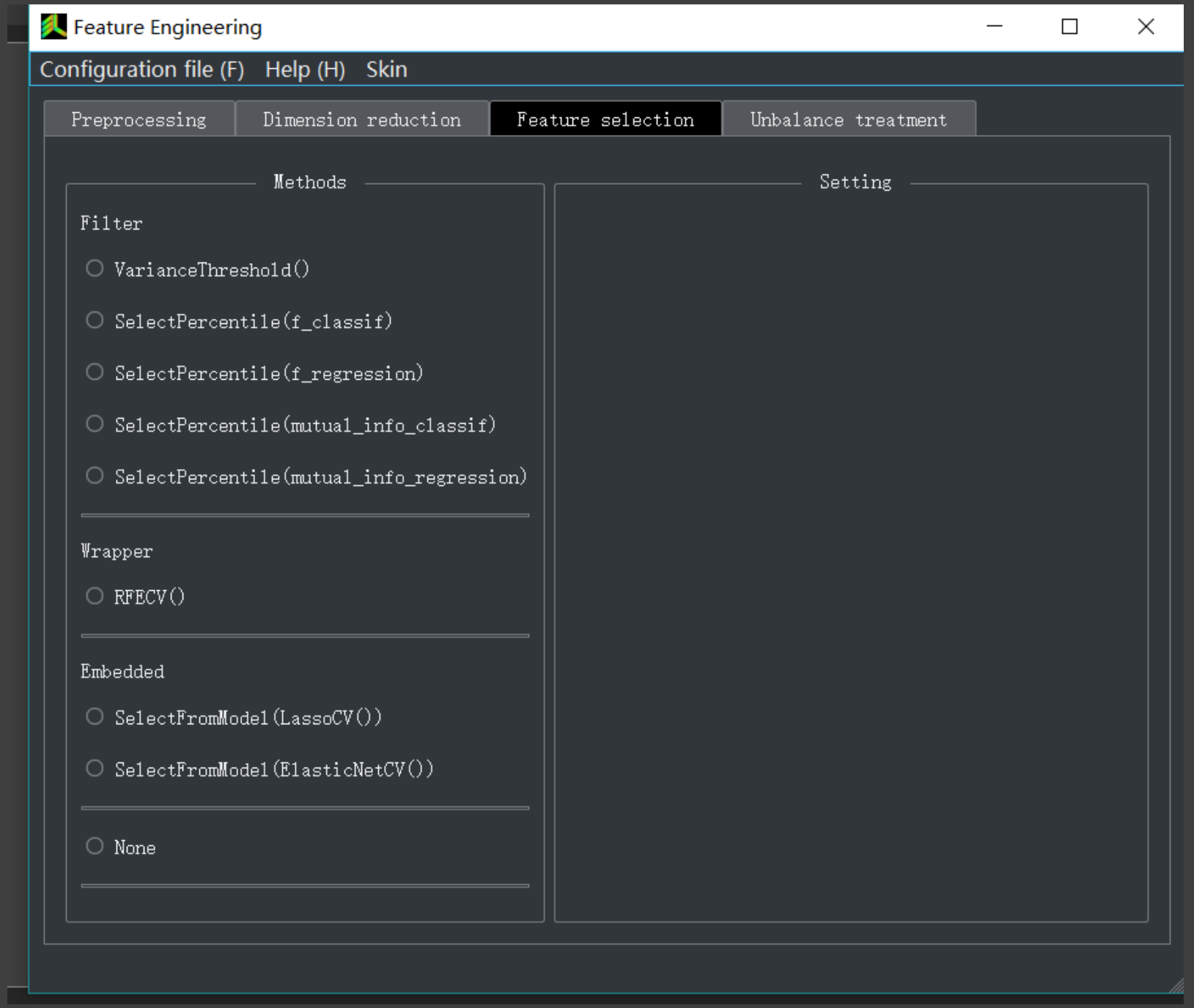


Figure 6. Here the workflow of nested cross-validation is demonstrated in illustrative form. The data set is folded into multiple combinations of hold-out set and inner optimization set. Each of these folds is essentially similar to the lock box approach described above and can be optimized. The final accuracy would be the average accuracy computed across all of the hold-out sets.

特征工程

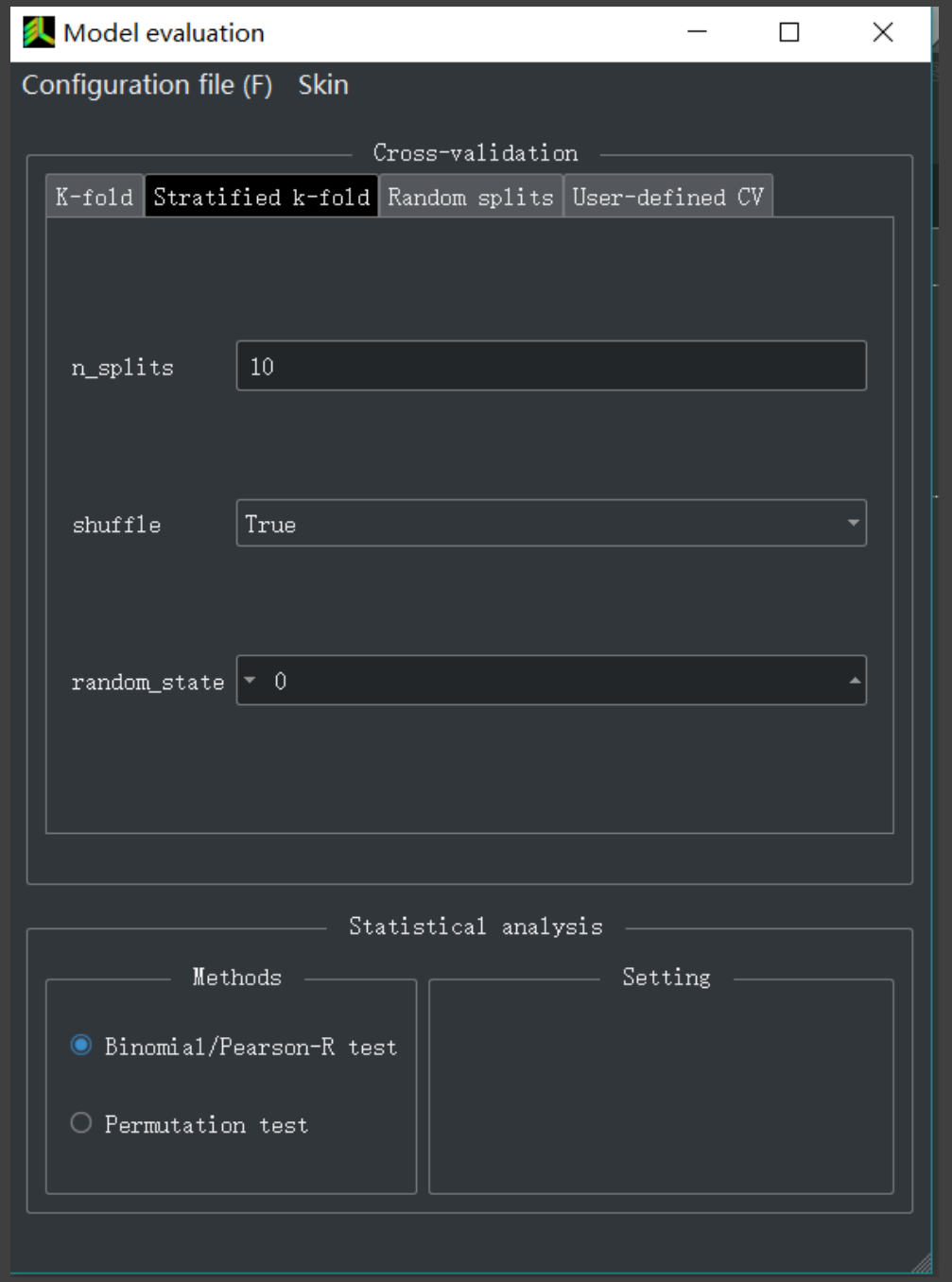
- 请注意：
分类和回归在特征筛选时对应不同的方法



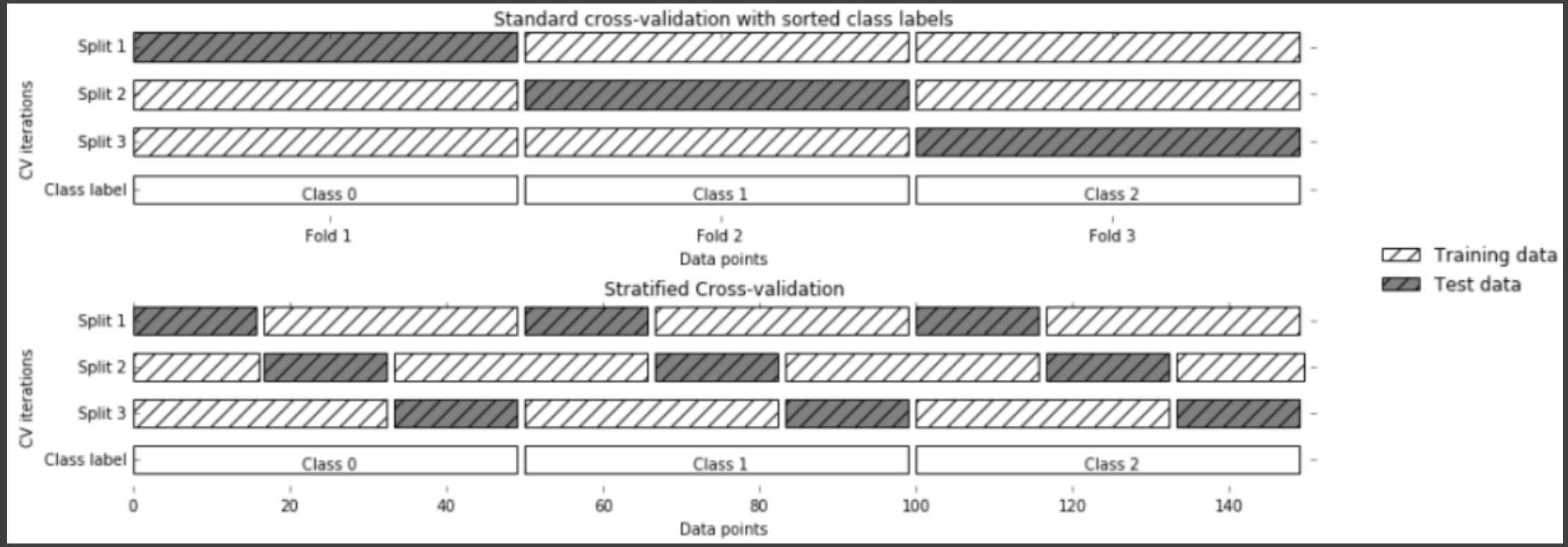
特征工程

- 请注意：

回归时不要用分层交叉验证



什么是分层交叉验证


Standard cross-validation with sorted class labels
Split 1
Split 2
Split 3
Class label
CV iterations
Class 0
Class 1
Class 2
Fold 1
Fold 2
Fold 3
Data points
Stratified Cross-validation
Split 1
Split 2
Split 3
Class label
CV iterations
Class 0
Class 1
Class 2
0
20
40
60
80
100
120
140
Data points
Training data
Test data

- 练习红葡萄就质量预测

- 查看结果view_results.py

- 自己操作白葡萄就质量预测

手把手实际操作

精分患者分类

失眠患者分类

手把手实际操作

AD比赛获奖经验

- 数据清洗:
- 数据处理:
- 特征工程:
- 参数寻优：调参神器Hyperopt
- 模型融合：提升准确度终极武器

调参神奇hyperopt

- 练习hyperopt1.py

- 练习红葡萄酒质量预测，并用pyperopt调参。redwine_hyperopt.py

神经影像聚类

深度神经网络

- 卷积神经网络

- 循环神经网络