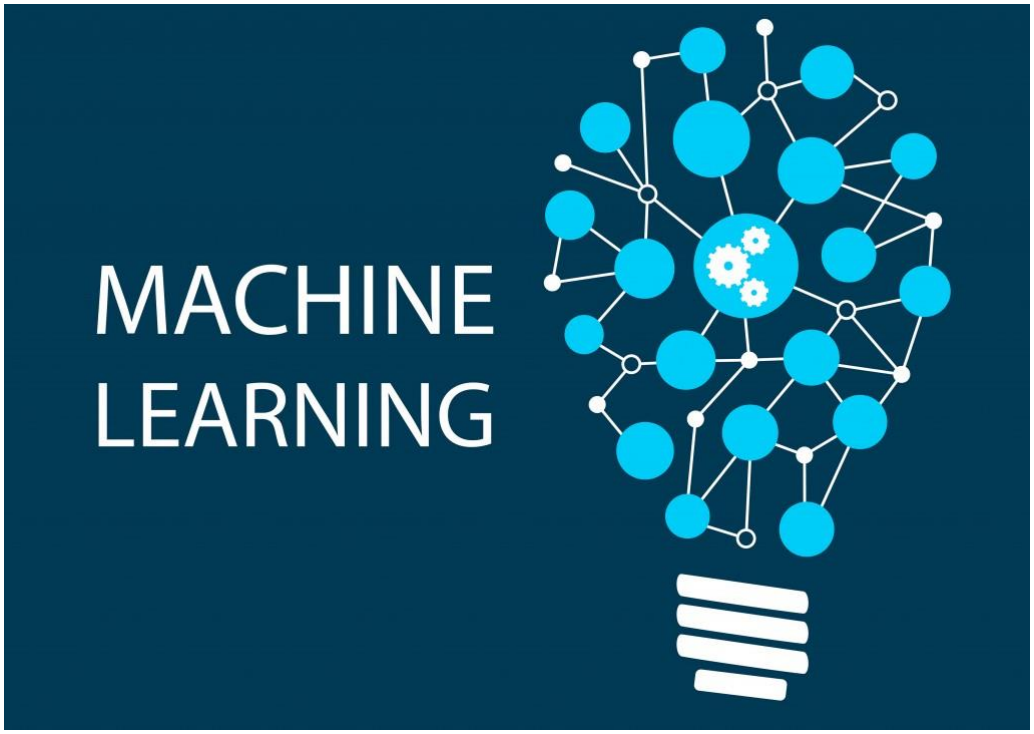
法律声明

- 本课件包括：演示文稿，示例，代码，题库，视频和声音等，小象学院拥有完全知识产权的权利；只限于善意学习者在本课程使用，不得在课程范围外向任何第三方散播。任何其他人或机构不得盗版、复制、仿造其中的创意，我们将保留一切通过法律手段追究违反者的权利。

关注 小象学院

第四讲



Python机器学习（2）

--Robin

目录

目录

决策树

- 例子1

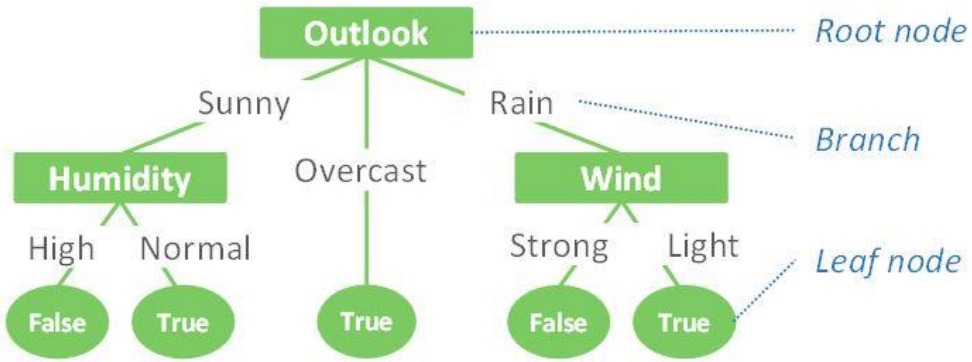

决策树例子
Is it alive?
No
Yes
Does it fly?
No
Yes
Carry > 10 people?
No
Yes
Objects with: Alive == No AND Fly == No AND Carry > 10 == No

决策树

- 例子2

Playing Tennis Example

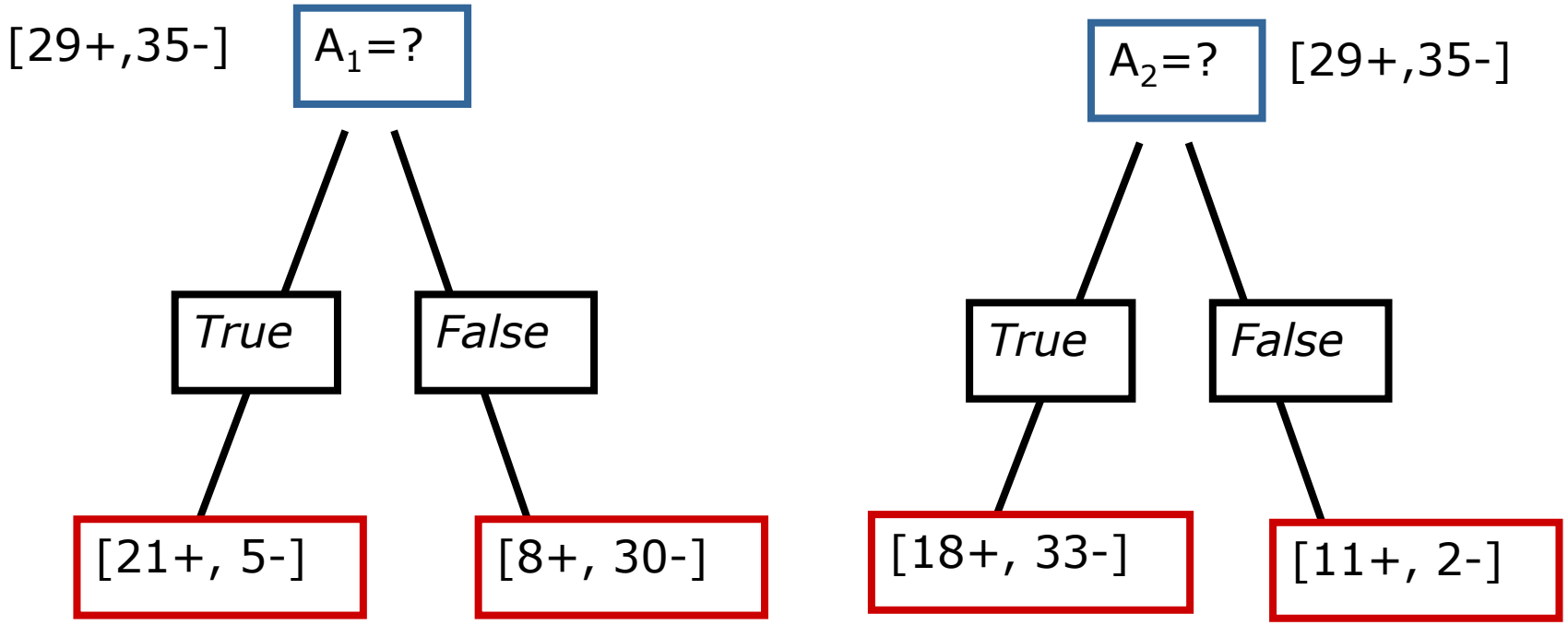
Day	Outlook	Temp	Humid	Wind	Play?
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No



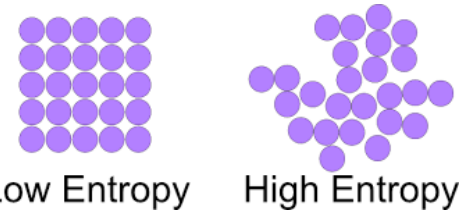
决策树

思考

- 按什么标准选择特征？
- 选择A_1还是A_2作为分割点？



决策树



预备知识

- 熵（Entropy）

在信息论中，设离散随机变量 的概率分布为 ，则概率分布的熵(Entropy)的定义为：

$$H(p) = -\sum_{i=1}^{n} p_i \log p_i$$

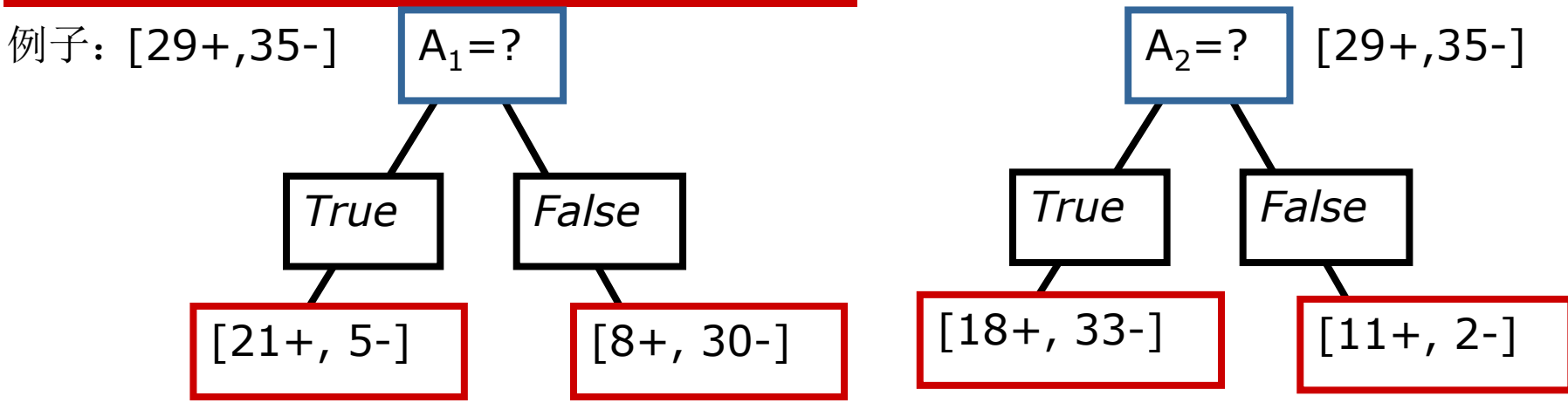
scipy.stats.entropy()

- 信息增益（Information Gain）

描述了当使用Q进行编码时，再使用P进行编码的差异。在决策树算法中，信息增益是针对某个特征而言的，就是看一个特征A，系统有它和没它的时候信息量各是多少，两者的差值就是这个特征给系统带来的信息量，即增益

$$Gain(S, A) \equiv Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

决策树



Entropy([29+,35-]) = -29/64 $\log_2$ 29/64 – 35/64 $\log_2$ 35/64 = 0.99

Entropy([21+,5-]) = 0.71

Entropy([8+,30-]) = 0.74

Gain(S, A_1)=Entropy(S)

-26/64*Entropy([21+,5-])

-38/64*Entropy([8+,30-])

=0.27

Entropy([18+,33-]) = 0.94

Entropy([11+,2-]) = 0.62

Gain(S,A_2)=Entropy(S)

-51/64*Entropy([18+,33-])

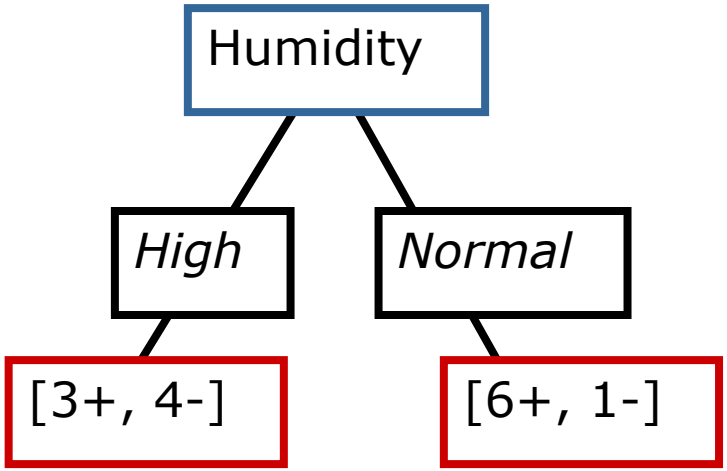
-13/64*Entropy([11+,2-])

=0.12

决策树

例子：

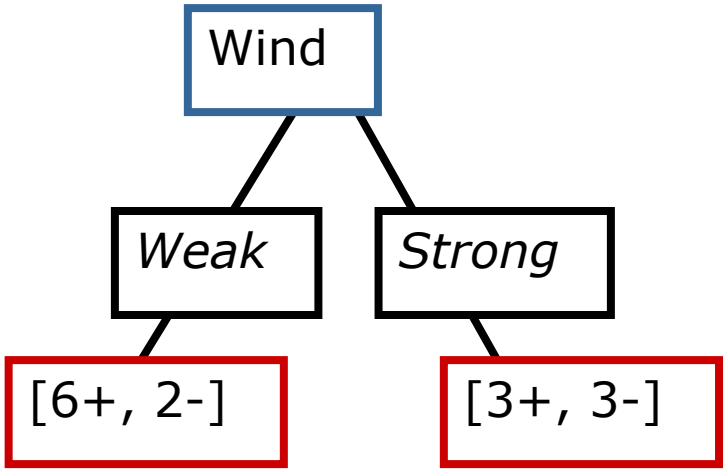
S=[9+,5-]
E=0.940



E=0.985　　E=0.592

Gain(S,Humidity)
=0.940-(7/14)*0.985
– (7/14)*0.592
=0.151

S=[9+,5-]
E=0.940

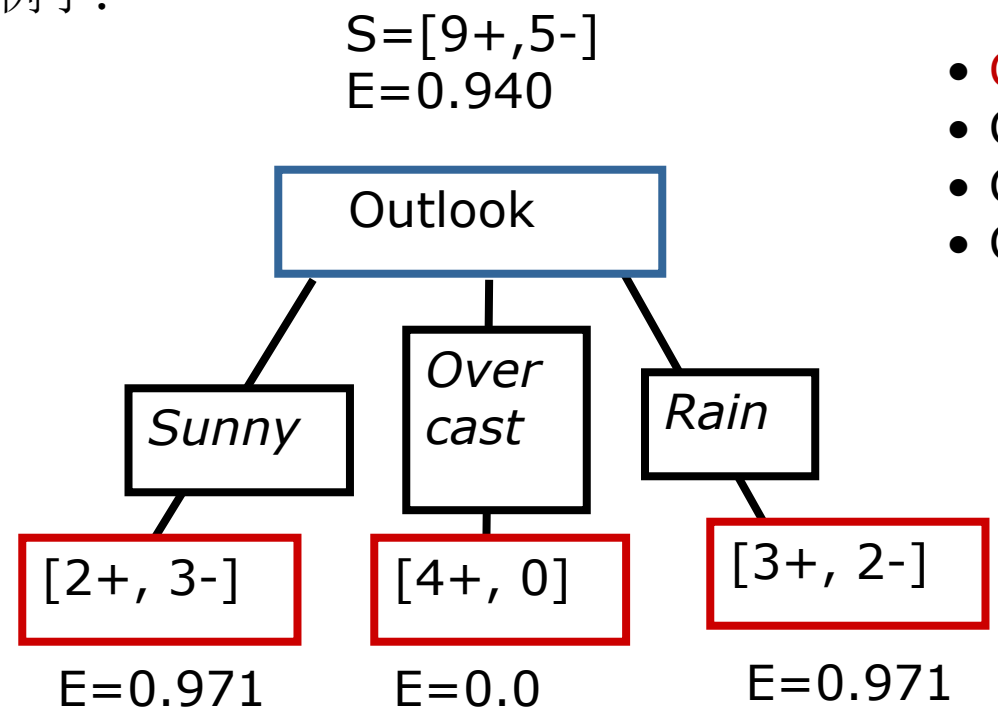


E=0.811

Gain(S,Wind)
=0.940-(8/14)*0.811
– (6/14)*1.0
=0.048

决策树

例子：



- Gain(S,Outlook) =0.247
- Gain(S,Humidity) =0.151
- Gain(S,Wind) =0.048
- Gain(S,Temperature) =0.029

特征为连续值？

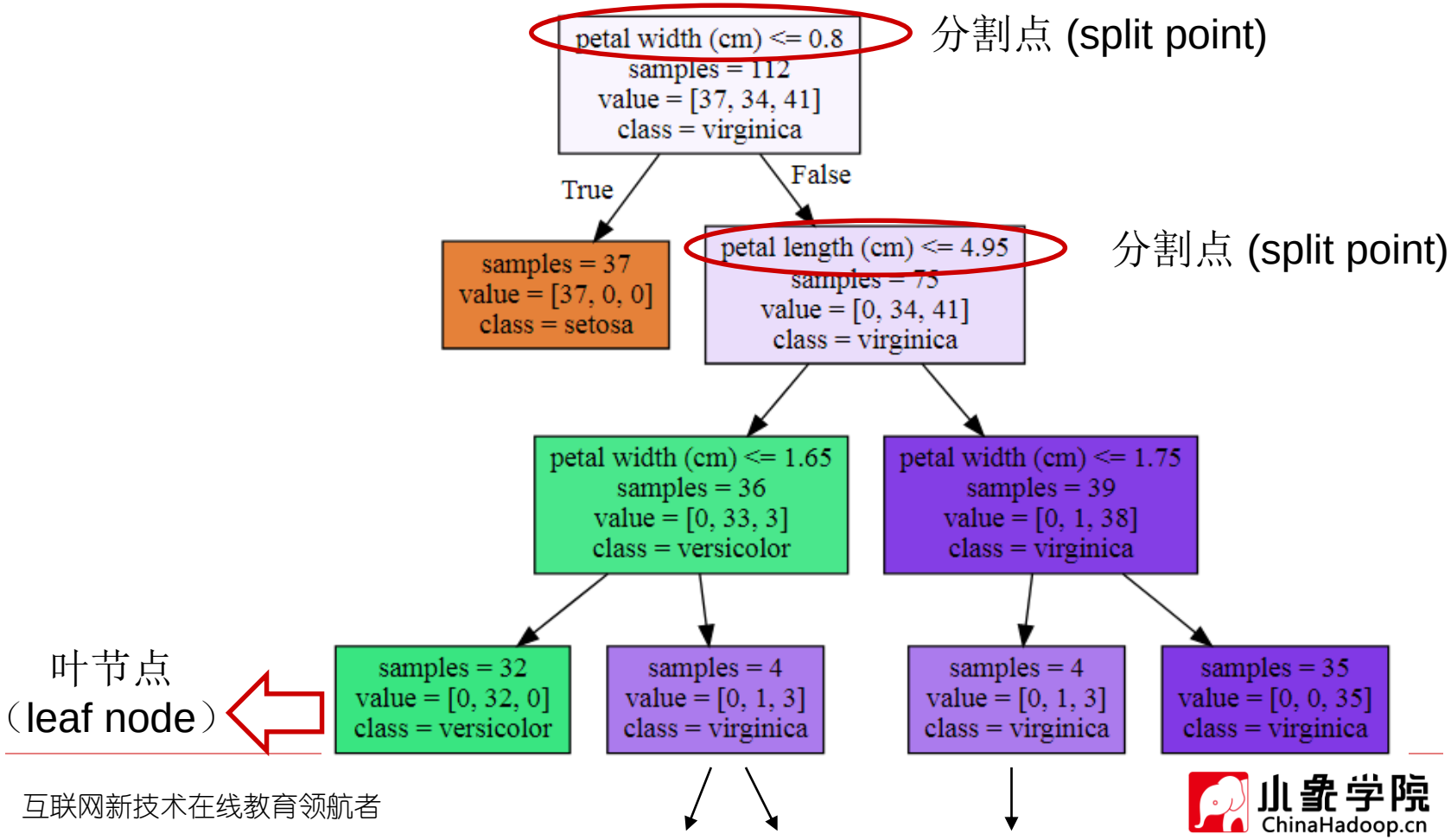
- 离散化连续值
  - 二分法

Gain(S,Outlook)
=0.940-(5/14)*0.971
-(4/14)*0.0 – (5/14)*0.0971
=0.247

决策树

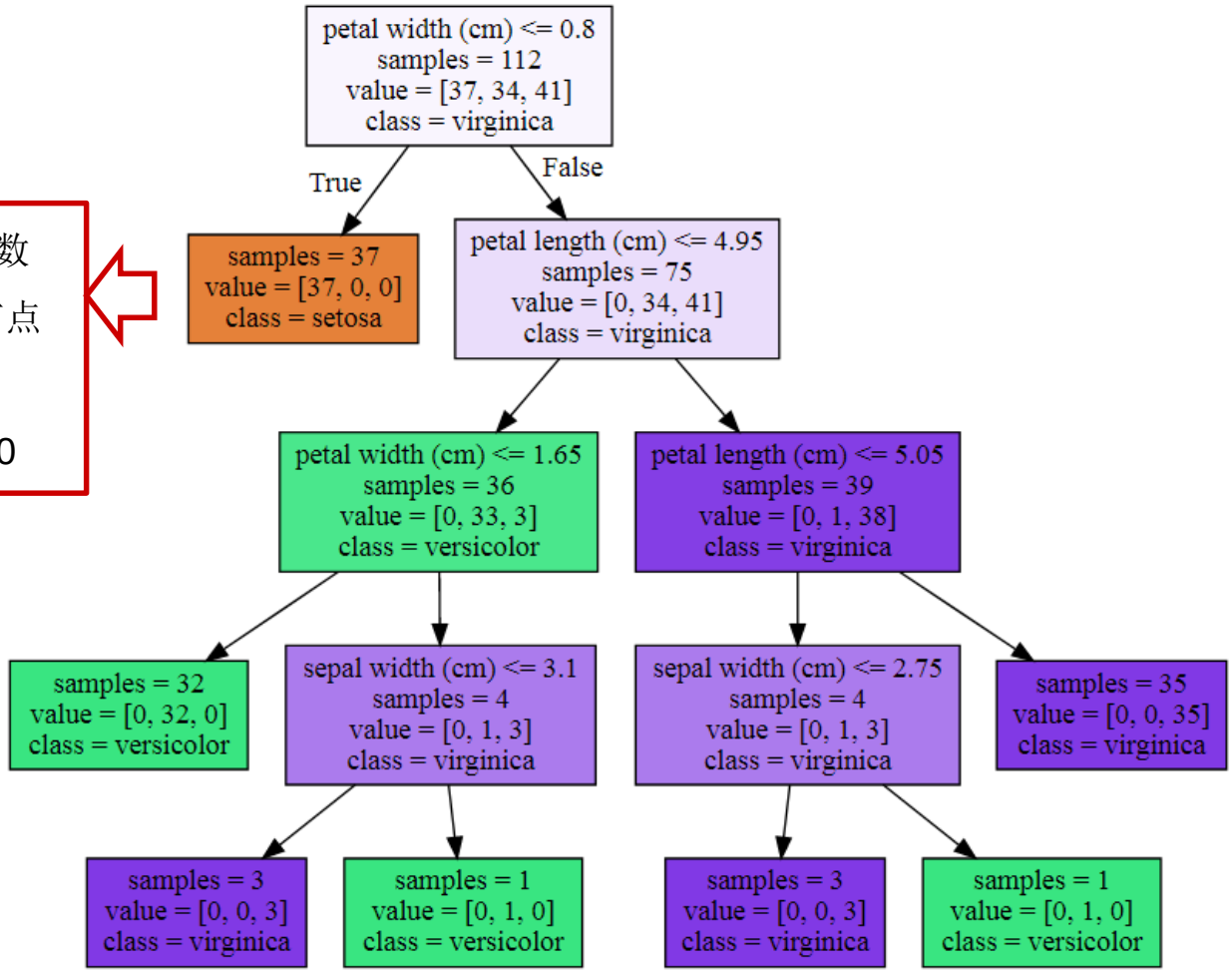
- 在iris数据集上使用决策树

lect04_eg01.ipynb



决策树

- 每个叶节点包含不同分类的样本个数
- 如：values=[37, 0, 0]表示在该叶节点中，属于setosa的样本个数为37，versicolor, viginica的样本个数均为0



决策树

- 构建树的过程：

  ID3　　ID4.5　　CART

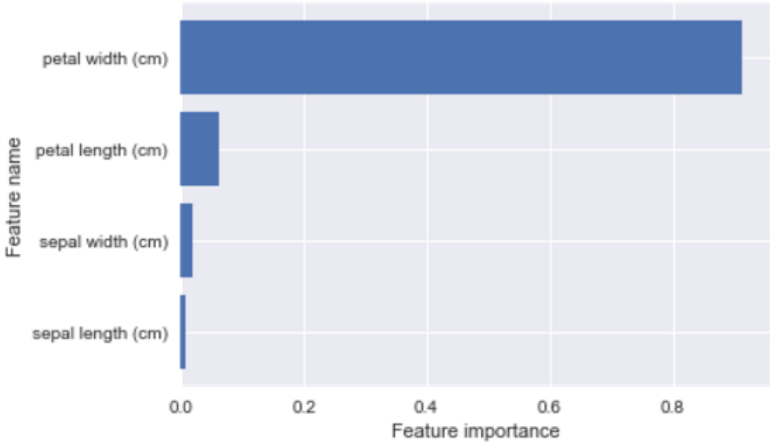
  1. 从根节点开始，计算所有特征值的信息增益（信息增益比、基尼指数），选择计算结果最大的特征作为根节点
  2. 根据算出的特征建立子节点，执行第1步，直到所有特征的信息增益（信息增益比）很小或没有特征可以选择为止
- 直接按照以上步骤构建树容易产生过拟合
- 防止过拟合：减少模型的复杂度。简化决策树->剪枝（pruning）
  - 预剪枝（pre-pruning），构造树的同时进行剪枝
  - 后剪枝（post-pruning），决策树构建完后再进行剪枝
  - 关于“剪枝” 的详细资料可参考：

    http://www.saedsayad.com/decision_tree_overfitting.htm

决策树

- sklean中决策树重要的参数
  - max_depth：树的最大深度（分割点的个数），最常用的用于减少模型复杂度防止过拟合的参数
  - min_samples_leaf：每个叶子拥有的最少的样本个数
  - max_leaf_nodes：树中叶子的最大个数
- 实际应用中，通常只需要调整max_depth就已足够防止决策树模型的过拟合

- feature importance：
  - 得分范围：0-1
  - 得分为0：特征在预测时没有作用
  - 得分为1：单独使用该特征即可完成预测
  - 每个特征的得分总和为1



决策树

- 决策树的优缺点

优点	缺点
• 容易可视化，容易解释 • 无需对特征进行归一化处理 • 可适用于混合特征类型的数据集（连续性特征，类别型特征等）	• 即使剪枝后也很难避免过拟合 • 通常需要进行ensemble才能达到较好的效果（如：随机森林）

目录

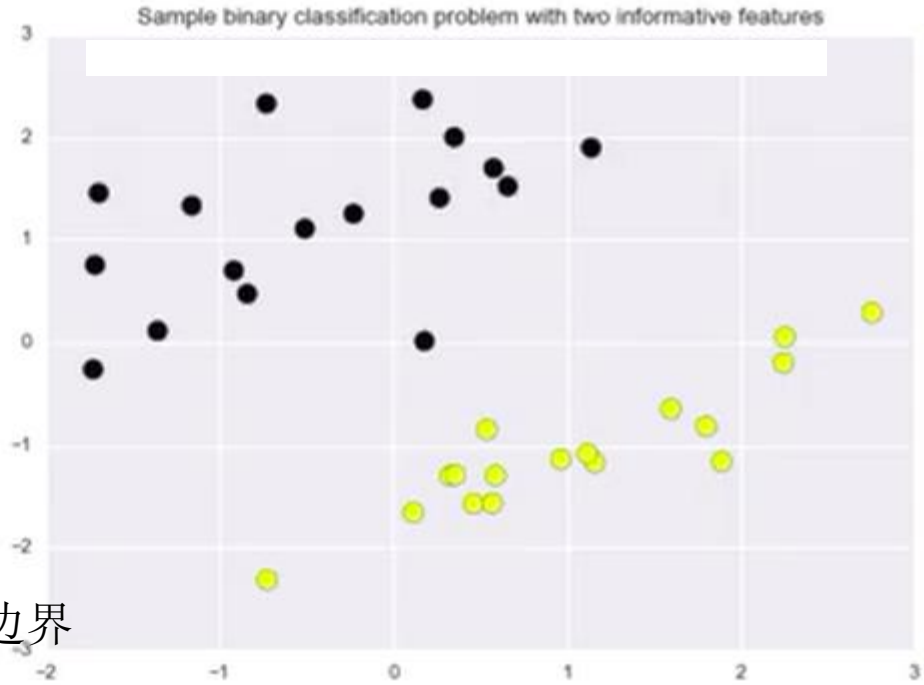
SVM

Feature vector

Class value

$$f(x,w,b) = sign(w \circ x + b)$$

$$= sign\left(\sum w[i]x[i] + b\right)$$



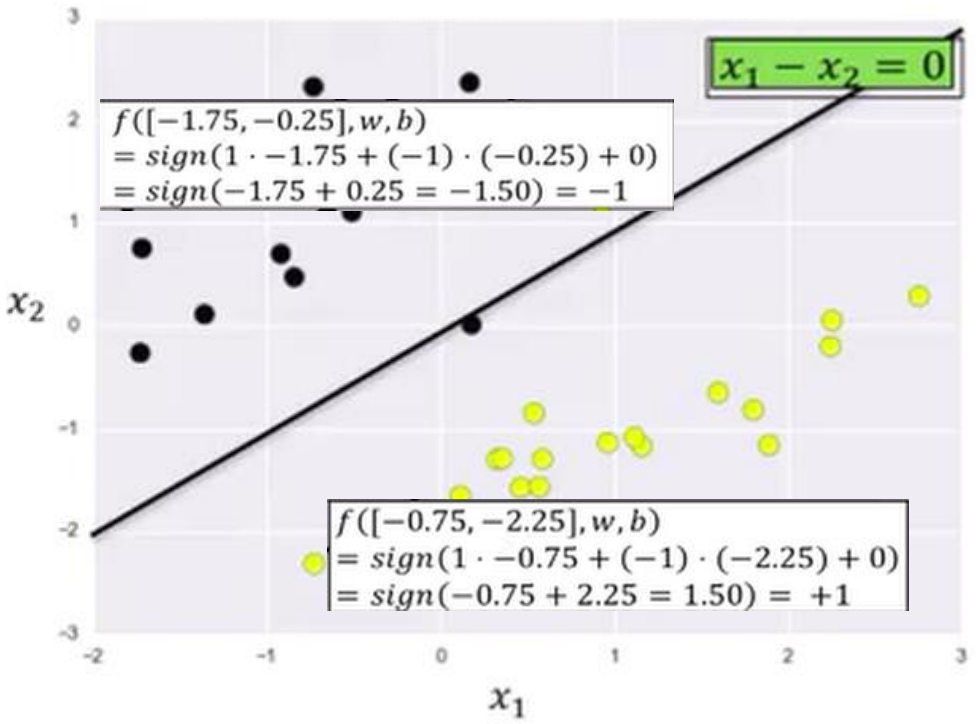
目标：找到一个可以分割这两个类别的边界

SVM

- 例子，假设这个边界已知

$$f(x, w, b) = sign(w \circ x + b)$$

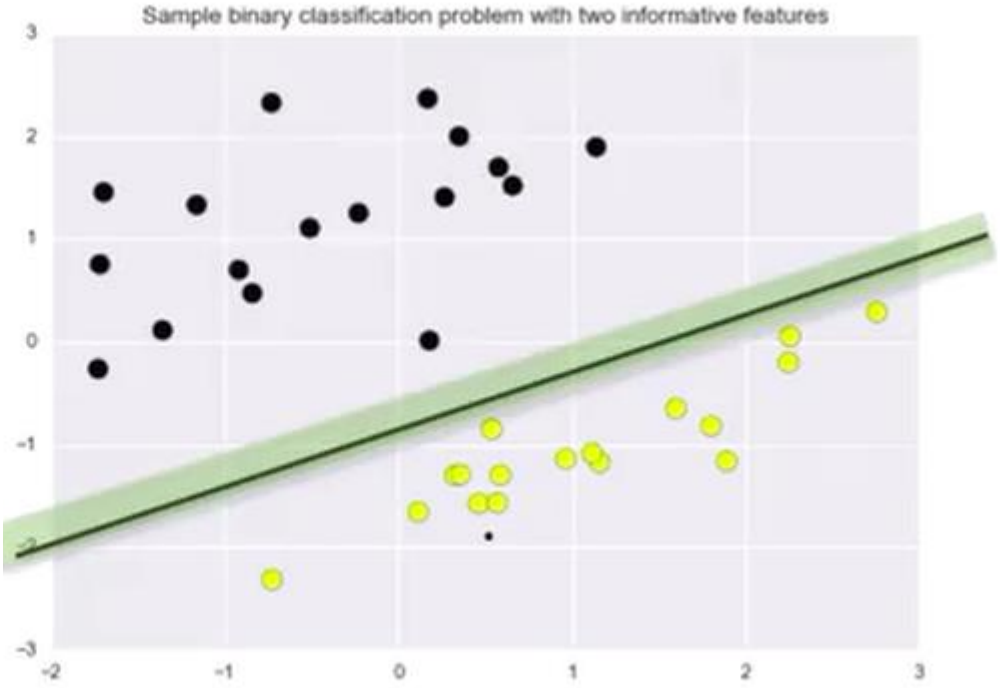
$$x_1 - x_2 = 0$$
$$w = [1, -1]$$
$$b = 0$$



SVM

- 间隔（ Margin ）

$$f(x, w, b) = sign(w \circ x + b)$$

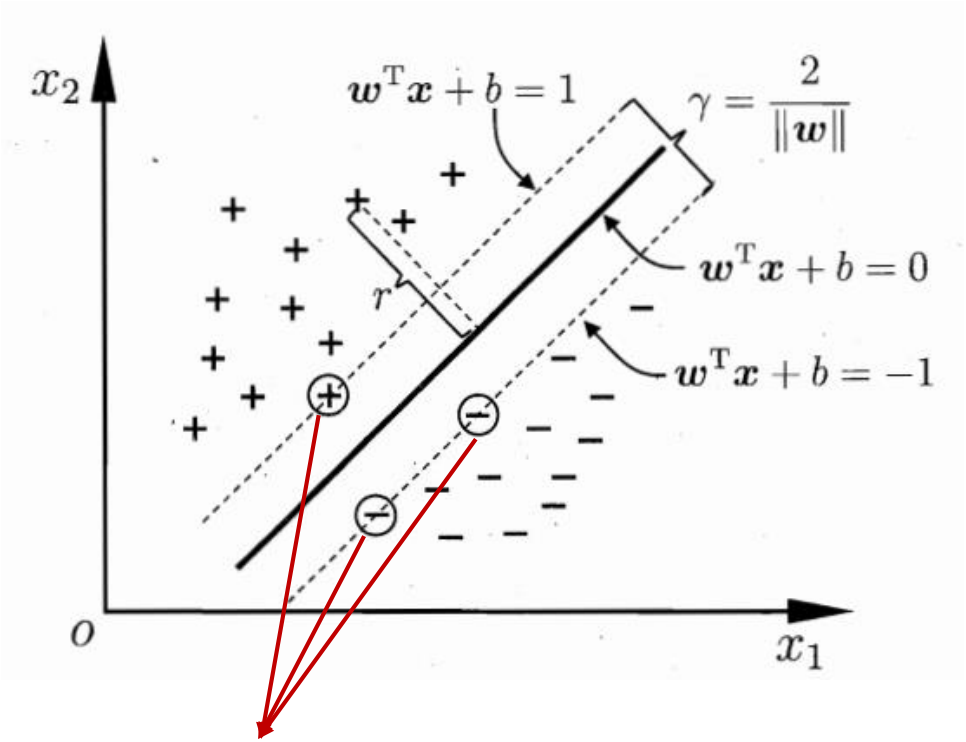


分类器间隔：

- 分类边界可以扩展到样本点的最大宽度。
- 例子中分类器的间隔就是绿色区域

SVM

- 间隔（ Margin ）



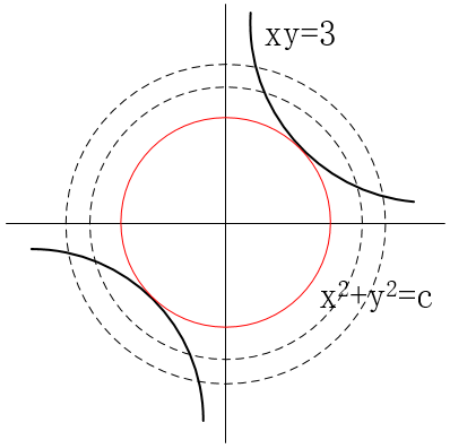
支持向量

- 带不等式约束的优化问题

$$\max_{w,b} \ \frac{2}{\|w\|}$$
$$\text{s.t. } y_i(w^T x_i + b) \geqslant 1, \quad i = 1, 2, \ldots, m.$$

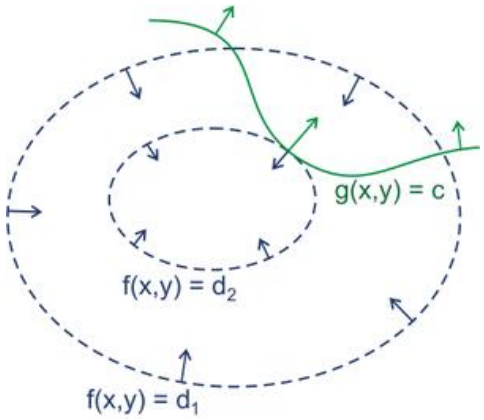
$$\min_{w,b} \ \frac{1}{2}\|w\|^2$$
$$\text{s.t. } y_i(w^T x_i + b) \geqslant 1, \quad i = 1, 2, \ldots, m.$$

SVM



预备知识 -- 拉格朗日乘子法

- 带等式约束的优化问题
- 例子：min $f(x,y)=x^2+y^2$, s.t. $xy=3$.
- 拉格朗日乘子法的基本形式：
- 求函数 $z=f(x,y)$ 在满足 $\varphi(x,y)=0$ 下的条件极值，可以转化为函数 $F(x,y,\lambda)=f(x,y)+\lambda\varphi(x,y)$ 的无条件极值问题。
- 解法：令F对x, y, lambda的偏导为零



SVM

预备知识 --拉格朗日乘子法

- 带不等式约束的优化问题

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & f(\boldsymbol{x}) \\ \text{s.t.} \quad & h_i(\boldsymbol{x}) = 0 \quad (i = 1, \ldots, m) \ , \\ & g_j(\boldsymbol{x}) \leqslant 0 \quad (j = 1, \ldots, n) \ . \end{aligned}$$

$$L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = f(\boldsymbol{x}) + \sum_{i=1}^{m} \lambda_i h_i(\boldsymbol{x}) + \sum_{j=1}^{n} \mu_j g_j(\boldsymbol{x})$$

- KKT条件：在满足一些有规则的条件下，一个非线性规划问题能有最优化解法的一个必要和充分条件

$$\begin{cases} g_j(\boldsymbol{x}) \leqslant 0; \\ \mu_j \geqslant 0 \ ; \\ \mu_j g_j(\boldsymbol{x}) = 0 \ . \end{cases}$$

例子：min $f(x)=x^2$, s.t. $x>=b$

SVM

- 拉格朗日乘子法

$$\min_{\boldsymbol{w},b} \ \frac{1}{2}\|\boldsymbol{w}\|^2 \quad \text{s.t. } y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i+b)\geqslant 1, \quad i=1,2,\ldots,m.$$

$$\Rightarrow \quad L(\boldsymbol{w},b,\boldsymbol{\alpha})=\frac{1}{2}\|\boldsymbol{w}\|^2+\sum_{i=1}^{m}\alpha_i\left(1-y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i+b)\right)$$

KKT条件

$$\begin{cases}\alpha_i\geqslant 0\,;\\ y_i f(\boldsymbol{x}_i)-1\geqslant 0\,;\\ \alpha_i\left(y_i f(\boldsymbol{x}_i)-1\right)=0\end{cases}$$

其中 $\boldsymbol{\alpha}=(\alpha_1;\alpha_2;\ldots;\alpha_m)$，令*L*对*w*和*b*的偏导为零可得：

$$\boldsymbol{w}=\sum_{i=1}^{m}\alpha_i y_i\boldsymbol{x}_i \qquad 0=\sum_{i=1}^{m}\alpha_i y_i$$

将*w*带入*L*中，可得

$$\max_{\boldsymbol{\alpha}} \ \sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j \qquad \text{s.t.} \ \sum_{i=1}^{m}\alpha_i y_i=0\,,$$

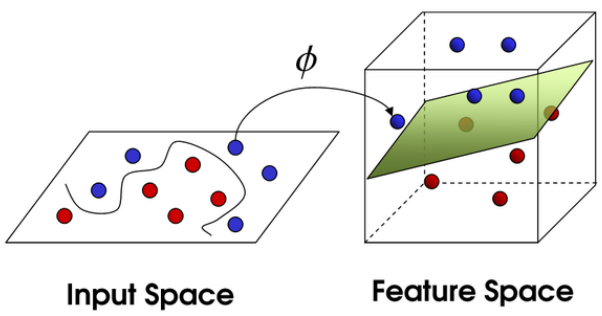
$$\alpha_i\geqslant 0\,,\quad i=1,2,\ldots,m\,.$$

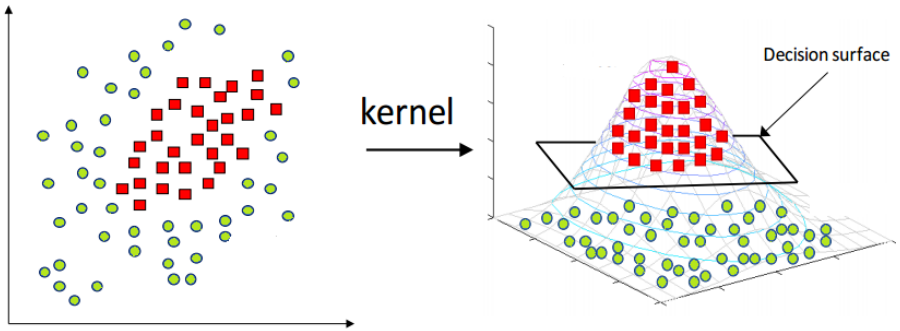
二次规划问题：可以使用SMO算法进行求解

SVM

核函数

- 如果数据线性不可分，怎么办？





- 将样本从原始空间映射到一个更高维的特征空间，从而使得样本在特征空间内线性可分

原始空间

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

$$\max_{\boldsymbol{\alpha}} \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j$$

特征空间

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

$$\max_{\boldsymbol{\alpha}} \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j \phi(\boldsymbol{x}_i)^{\mathrm{T}}\phi(\boldsymbol{x}_j)$$

SVM

特征空间

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

$$\max_{\boldsymbol{\alpha}} \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j \phi(\boldsymbol{x}_i)^{\mathrm{T}}\phi(\boldsymbol{x}_j)$$

$$\max_{\boldsymbol{\alpha}} \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j \kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

$$\begin{aligned} f(\boldsymbol{x}) &= \boldsymbol{w}^{\mathrm{T}}\phi(\boldsymbol{x}) + b \\ &= \sum_{i=1}^{m}\alpha_i y_i \phi(\boldsymbol{x}_i)^{\mathrm{T}}\phi(\boldsymbol{x}) + b \\ &= \sum_{i=1}^{m}\alpha_i y_i \kappa(\boldsymbol{x}, \boldsymbol{x}_i) + b\ . \end{aligned}$$

如果特征空间维数很高，直接计算会非常困难

核函数

$$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_j) \rangle = \phi(\boldsymbol{x}_i)^{\mathrm{T}}\phi(\boldsymbol{x}_j)$$

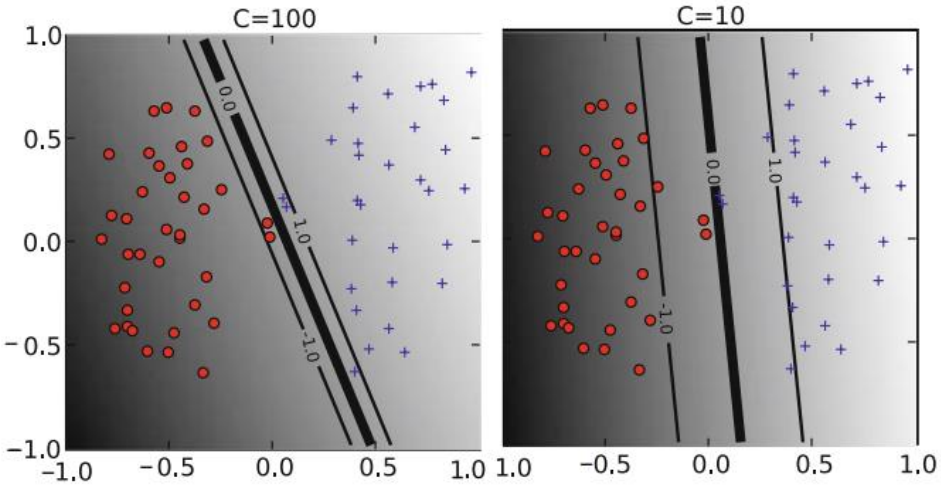
常用的核函数

名称	表达式	参数
线性核	$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j$	
多项式核	$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j)^d$	$d \geqslant 1$ 为多项式的次数
高斯核	$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{2\sigma^2}\right)$	$\sigma > 0$ 为高斯核的带宽(width)
拉普拉斯核	$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|}{\sigma}\right)$	$\sigma > 0$
Sigmoid 核	$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \tanh(\beta \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j + \theta)$	tanh 为双曲正切函数, $\beta > 0$, $\theta < 0$

SVM

$$\min_f \sum_{i=1}^{n} V(f(x_i), y_i) + \lambda R(f)$$

损失函数　　正则项



C=1/λ

- 正则项中的C值决定了正则化的强度
- C值越大，正则化越弱 -> narrow margin
- C值越小，正则化越强 ->large margin

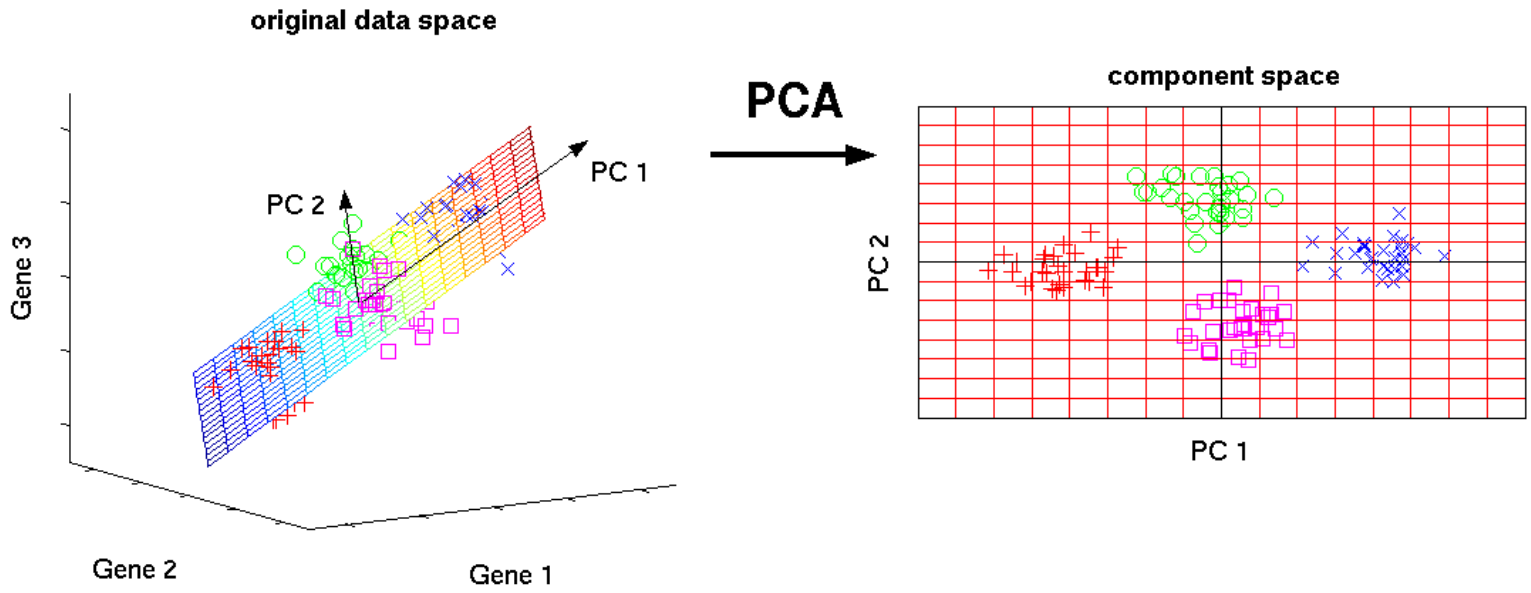
lect04_eg01.ipynb

目录

PCA

Principal components analysis (PCA)

- 用于减少数据集的维度，同时保持数据集中的对方差贡献最大的特征
- 保留低阶主成分，忽略高阶成分，这样的低阶成分往往能够保留住数据的最重要方面



PCA

方差与协方差

- 用于衡量一系列点在它们的重心或均值附近的分散程度
- 方差：衡量这些点在一个维度的偏差
- 协方差：衡量一个维度是否会对另一个维度有所影响，从而查看这两个维度之间是否有关系
  - 某个维度和自身之间的协方差就是其方差

$$COV(X,Y) = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{n-1}$$

协方差矩阵

- 如果数据集是d维的，(x1，x2，…，xd)，则可计算出(x1，x2)，(x1，x3)，…，(x1，xd)，(x2，x3)，…(x2，xd)，…(xd-1，xd)之间的协方差。由于协方差的对称性，再加上各维度自身的协方差，可以构成协方差矩阵

PCA

$$\begin{bmatrix} \mathrm{cov}(x1,x1) & \mathrm{cov}(x1,x2) & \mathrm{cov}(x1,x3) \\ \mathrm{cov}(x2,x1) & \mathrm{cov}(x2,x2) & \mathrm{cov}(x2,x3) \\ \mathrm{cov}(x3,x1) & \mathrm{cov}(x3,x2) & \mathrm{cov}(x3,x3) \end{bmatrix}$$

协方差矩阵 (续)

- 其中对角线上的是方差
- 协方差为正,代表两个变量变化趋势相同；反之亦然

PCA

- 通过线型变换将原数据映射到新的坐标系统中，使映射后的第一个坐标上的方差最大（即第一个主成分）,第二个坐标上的方差第二大（即第二个主成分）…

PCA

PCA步骤：

1. 数据集 $\mathbf{X} \in R^{m \times n}$ ，其中每个样本 $\mathbf{x}^{(\mathrm{i})} = [\mathbf{x}_1^{(i)}, \mathbf{x}_2^{(i)}, ..., \mathbf{x}_n^{(i)}]$

   计算每个维度的均值

$$\bar{\mathbf{x}} = \frac{1}{m} \sum_{i=1}^{m} [\mathbf{x}_1^{(i)}, \mathbf{x}_2^{(i)}, ..., \mathbf{x}_n^{(i)}] \in R^n$$

   每个维度减去这个均值，得到一个矩阵

   相当于将坐标系进行了平移

$$\mathbf{Y} = \begin{bmatrix} \mathbf{x}^{(1)} - \bar{\mathbf{x}} \\ \mathbf{x}^{(2)} - \bar{\mathbf{x}} \\ \cdots \\ \mathbf{x}^{(\mathrm{m})} - \bar{\mathbf{x}} \end{bmatrix}$$

PCA

PCA步骤：

2. 构建协方差矩阵 $\mathbf{Q} = \mathbf{Y}^T\mathbf{Y} = \begin{bmatrix} \mathbf{x}^{(1)} - \bar{\mathbf{x}} & \mathbf{x}^{(2)} - \bar{\mathbf{x}} & \dots & \mathbf{x}^{(m)} - \bar{\mathbf{x}} \end{bmatrix} \begin{bmatrix} \mathbf{x}^{(1)} - \bar{\mathbf{x}} \\ \mathbf{x}^{(2)} - \bar{\mathbf{x}} \\ \dots \\ \mathbf{x}^{(m)} - \bar{\mathbf{x}} \end{bmatrix}$

3. 矩阵分解 (如SVD)，得到特征值(eigenvalues)及特征向量 (eigenvectors)

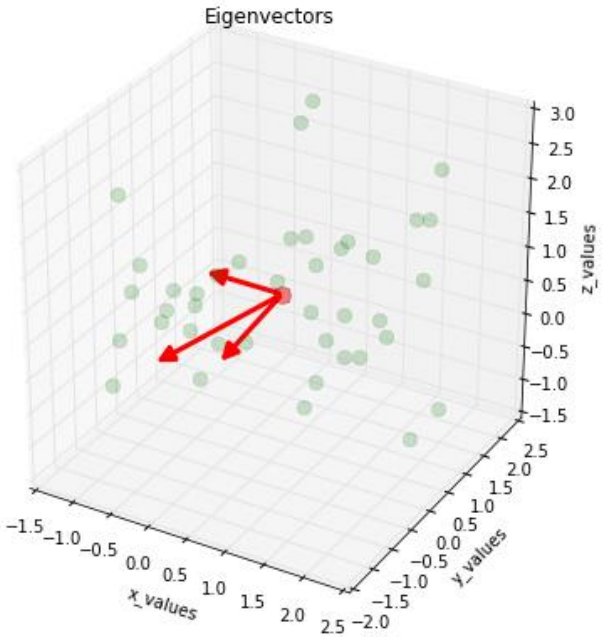
4. 将特征值从大到小排序，对应的特征向量就是第一个主成分，第二个主成分…

如何选择主成分个数？

- 交叉验证
- 根据主成分的累计贡献率(t)

$$\frac{\sum_{i=1}^{d'} \lambda_i}{\sum_{i=1}^{d} \lambda_i} \geqslant t$$

应用: 特征提取、数据降维



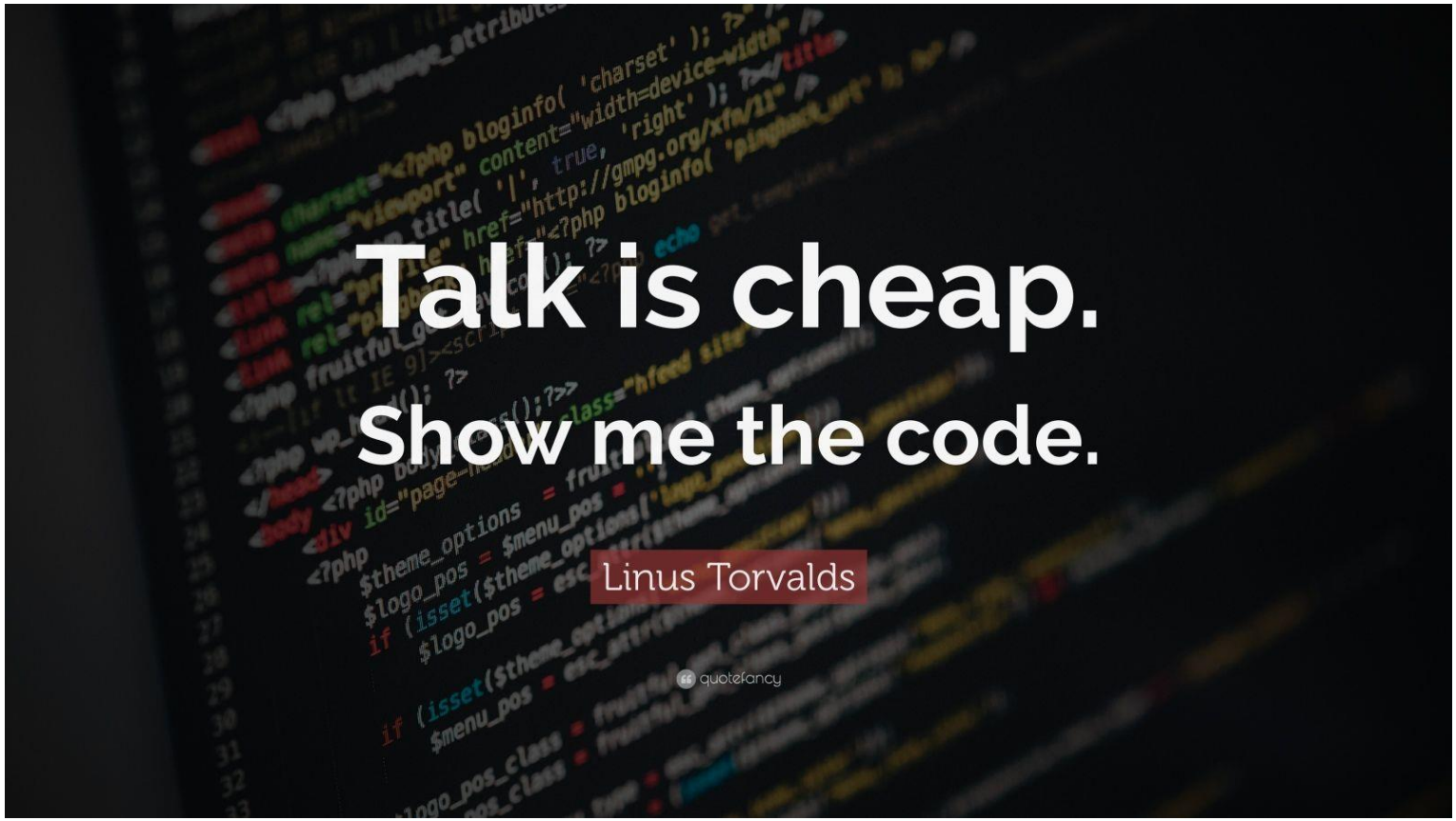
示例代码： 04_eg02.ipynb

目录

实战案例 3-2

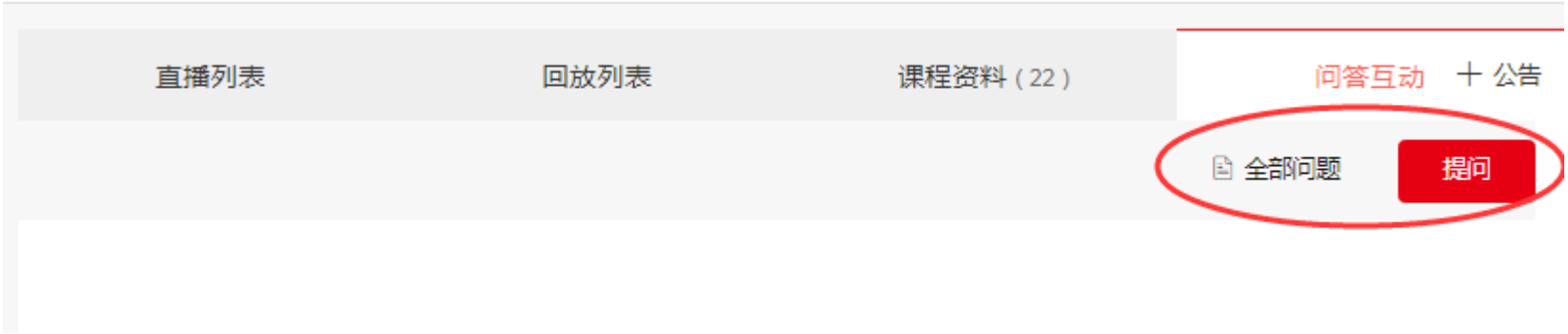
项目名称：贷款违约预测（2）

- 请参考相应的配套代码及案例讲解文档



问答互动

在所报课的课程页面，
1、点击“全部问题”显示本课程所有学员提问的问题。
2、点击“提问”即可向该课程的老师和助教提问问题。



联系我们

小象学院：互联网新技术在线教育领航者

- 微信公众号：**小象学院**