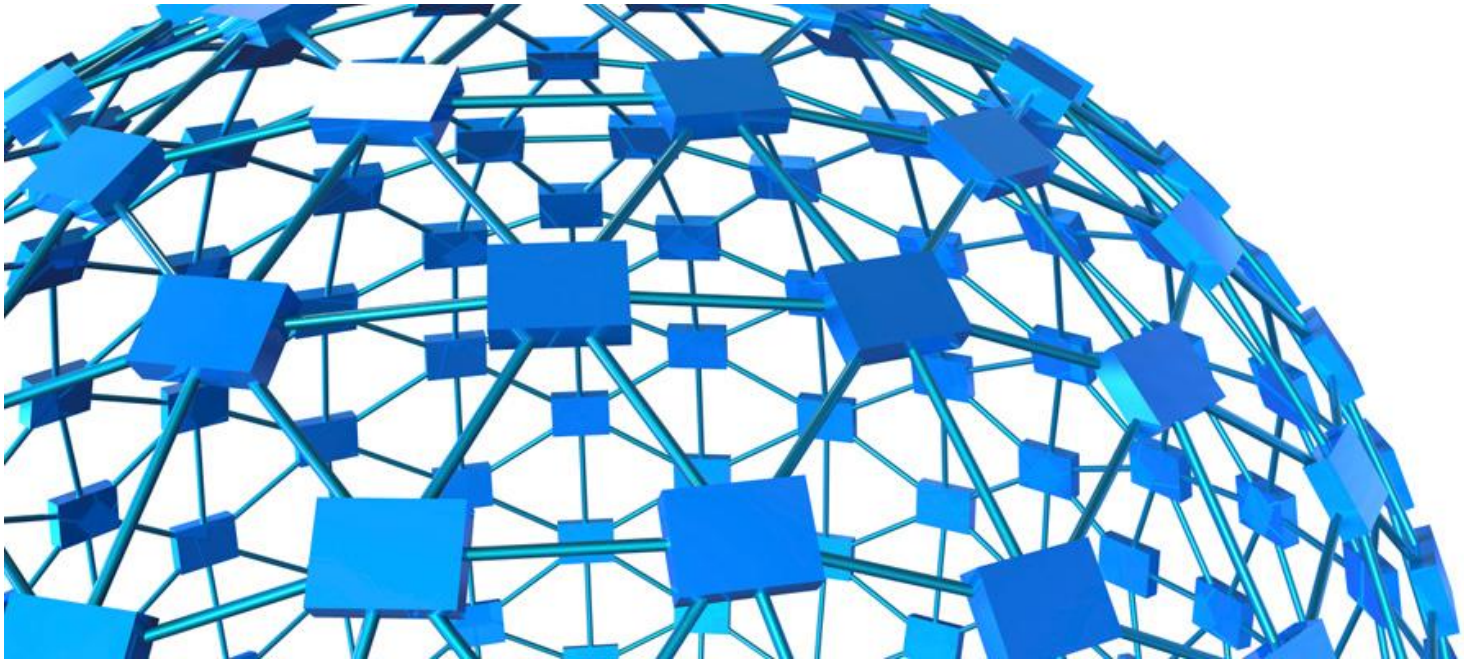
法律声明

- 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。

- 课程详情请咨询
  - 微信公众号：大数据分析挖掘
  - 新浪微博：ChinaHadoop

第九讲



项目实战

--梁斌

目录

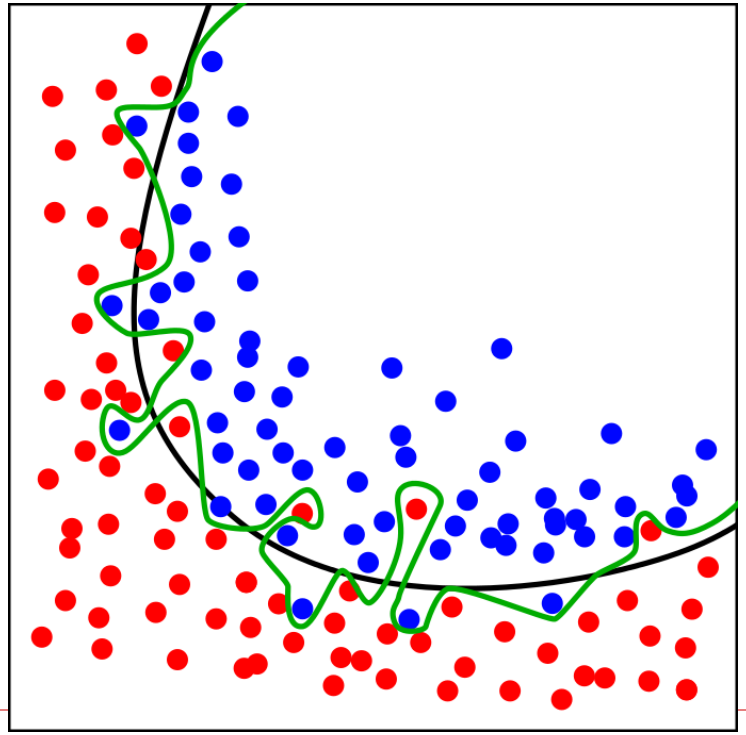
- 过拟合与欠拟合
- 交叉验证及参数调整
- 特征选择
- 评价指标补充
- 项目实战：通过移动设备行为数据预测性别年龄
- 课程总结

目录

过拟合与欠拟合

过拟合（Overfitting）

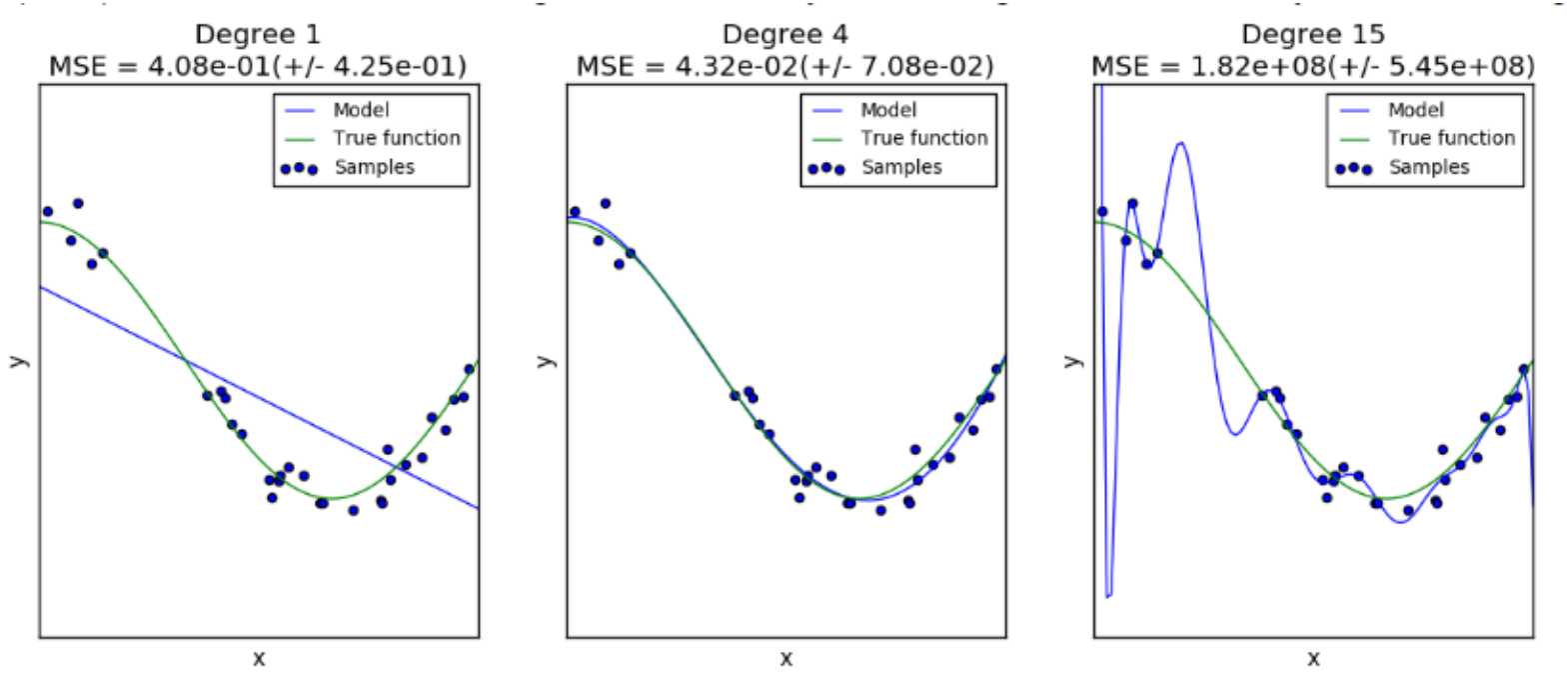
- 是指在调适一个统计模型时，使用过多参数。模型对于训练数据拟合程度过当，以致太适应训练数据而非一般情况。
- 在训练数据上表现非常好，但是在测试数据或验证数据上表现很差。



过拟合与欠拟合

欠拟合（Underfitting）

- 模型在训练和预测时表现都不好的情况
- 欠拟合很容易被发现



欠拟合　　“刚刚好”　　过拟合

示例代码：01_fit_prob.ipynb

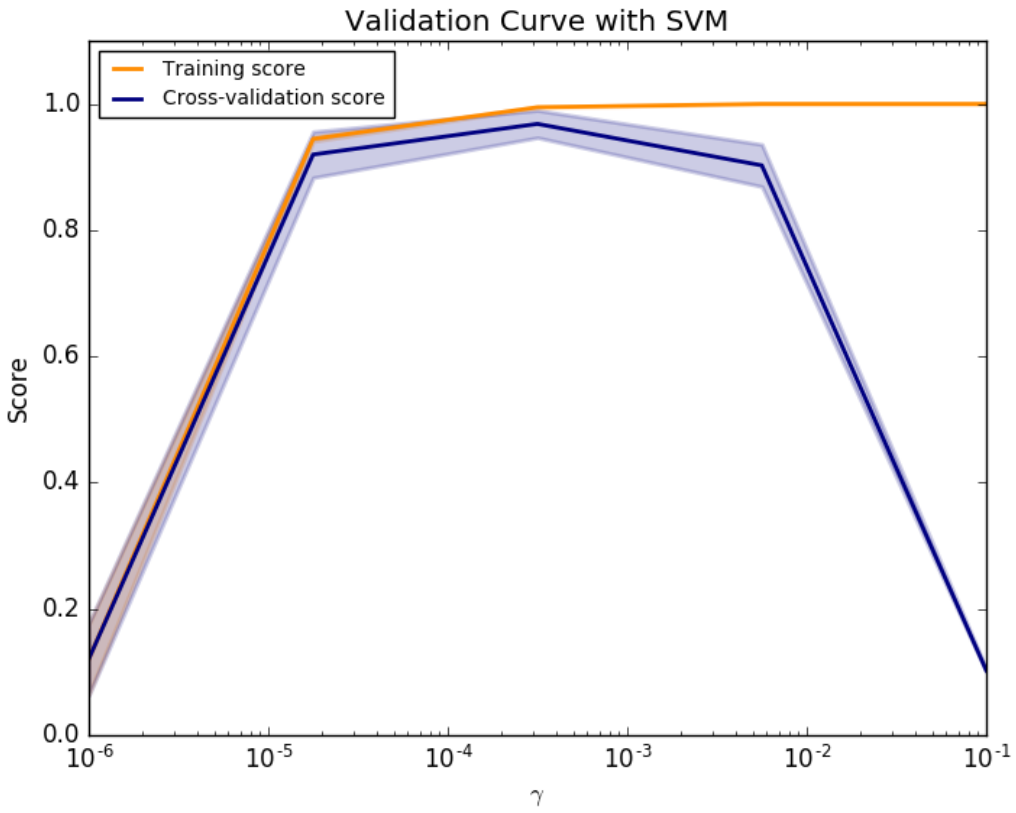
目录

- 过拟合与欠拟合
- 交叉验证及参数调整
- 特征选择
- 评价指标补充
- 项目实战：通过移动设备行为数据预测性别年龄
- 课程总结

交叉验证及参数调整

验证曲线（validation curve）

- `sklearn.model_selection.validation_curve`



示例代码：02_val_curve.ipynb

交叉验证及参数调整

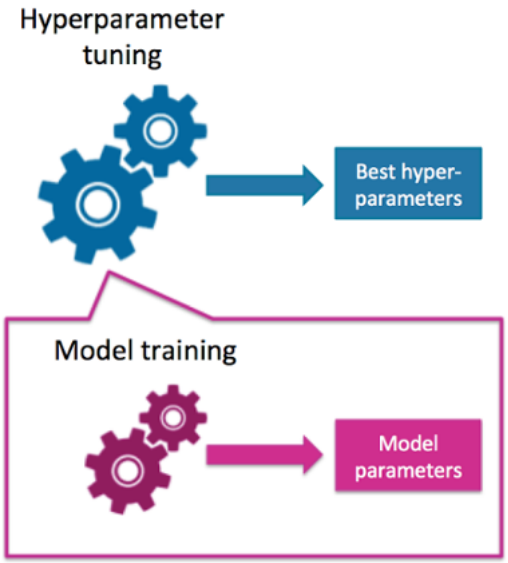
参数调整

- 模型参数包括两种
  1. 模型自身参数，通过样本学习得到的参数。如：逻辑回归及神经网络中的权重及偏置的学习等
  2. 超参数，模型框架的参数，如kmeans中的k，神经网络中的网络层数及每层的节点个数。通常由手工设定
- 如何调整参数
  1. 交叉验证

     sklearn.model_selection.cross_val_score

     参考第8课中的 02_scikit_tutorial.ipynb
  2. 网格搜索(Grid Search)

     sklearn.model_selection.GridSearchCV



示例代码：03_param_tuning.ipynb

目录

- 过拟合与欠拟合
- 交叉验证及参数调整
- 特征选择
- 评价指标补充
- 项目实战：通过移动设备行为数据预测性别年龄
- 课程总结

特征选择

Sklearn中的特征选择方法

- 去除方差小的特征，VarianceThreshold

  极端情况，如果所有样本在某个维度上的特征全都相同，即0方差，说明该特征描述或代表样本的能力很弱

- 基于单变量统计特征选择

  根据单变量统计测试选取特征，SelectKBest

- 基于模型的特征选择

  如：随机森林等

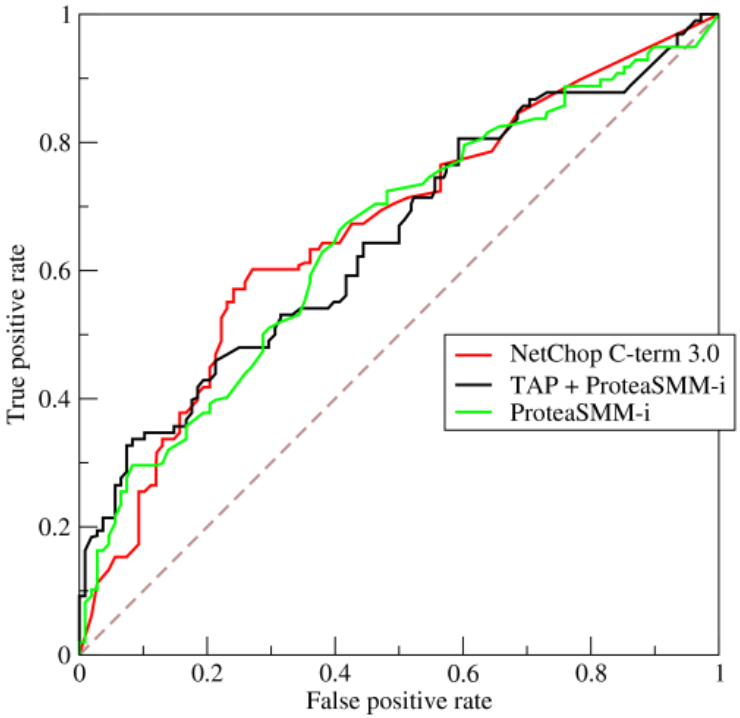
示例代码：04_feat_selection.ipynb

目录

评价指标补充

准确率越高，模型越好？

- 准确率99%的模型是优秀的模型么？
- 在100个样本中，99个负样本，1个正样本，
  如果全部预测负样本，就可以得到准确率99%!
- 但是，这样的模型是你想要的么？



1. 曲线下面积（Area Under Curve，AUC）

- 二分类模型的评价指标
- 曲线：接收者操作特征曲线
  （receiver operating characteristic curve，ROC曲线）
- AUC的值就是ROC曲线下的面积

评价指标补充

1. 曲线下面积（Area Under Curve，AUC）（续）

- 真阳性(TP)，预测值是1，真实值是1
- 伪阳性(FP)，预测值是1，但真实值是0
- 真阴性(TN)，预测值是0，真实值是0
- 伪阴性(FN)，预测值是0，但真实值是1

		Prediction		
		Positive	Negative	
Ground truth	Positive	**True positive (TP)**	**False negative (FN)**	**True positive rate** $\frac{\#TP}{\#TP + \#FN}$
	Negative	**False positive (FP)**	**True negative (TN)**	**False positive rate** $\frac{\#FP}{\#FP + \#TN}$

评价指标补充

1. 曲线下面积（Area Under Curve，AUC）（续）

- TPR：在所有实际值是1的样本中，被正确地预测为1的比率

  TPR = TP / (TP + FN)

- FPR：在所有实际值是0的样本中，被错误地预测为1的比率

  FPR = FP / (FP + TN)

- ROC空间将FPR定义为x轴，TPR定义为y轴
- 根据预测概率和设定的阈值将样本划到相应类别中

  如：某样本被预测为0的概率是0.7，被预测为1的概率是0.3

  如果设定阈值是0.2，该样本被划分到1

  如果设定阈值是0.4，该样本被划分到0
- 选取0~1每个点为阈值，根据所划分的类别分别计算TPR和FPR，描绘在ROC空间内，连接这些坐标点就得到了ROC曲线

评价指标补充

1. 曲线下面积（Area Under Curve，AUC）（续）

- AUC在0~1之间
- 0.5 < AUC < 1，优于随机猜测。这个分类器（模型）妥善设定阈值的话，能有预测价值。
- AUC = 0.5，跟随机猜测一样（例：丢铜板），模型没有预测价值。
- AUC < 0.5，比随机猜测还差；但只要总是反预测而行，就优于随机猜测。
- 详细讲解请参考：

  https://zh.wikipedia.org/wiki/ROC%E6%9B%B2%E7%BA%BF

评价指标补充

2．对数损失（logloss）

- 对于每个样本来说，预测结果会将其归到某一类中；但有时模型的输出结果是一组概率，比如3分类问题，输出结果可能是 [0.1, 0.8, 0.1]，那么这个样本被预测为第二个类
- 对于这类模型的输出可以用logloss来评价预测结果，公式如下：

$$logloss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{i,j}\log(p_{i,j})$$

其中，N是样本数量，M是类别个数

如果第i个样本属于第j个类，y_{ij}为1，否则为0

p_{ij}是第i个样本被预测为第j个类的概率

- sklearn.metrics.log_loss

目录

项目实战

项目介绍

- https://www.kaggle.com/c/talkingdata-mobile-user-demographics
- 通过行为习惯对移动用户人口属性（年龄+性别）进行预测
- 数据及包含~20万用户数据，分成12组，同时提供了用户行为属性，如：手机品牌、型号、APP的类型等
- 评价指标：logloss
- 步骤：
  1. 解读数据
  2. 特征工程
  3. 模型调参
- [注意]由于数据量相对较大，机器内存配置最好16G，或者考虑使用部分的数据进行实验

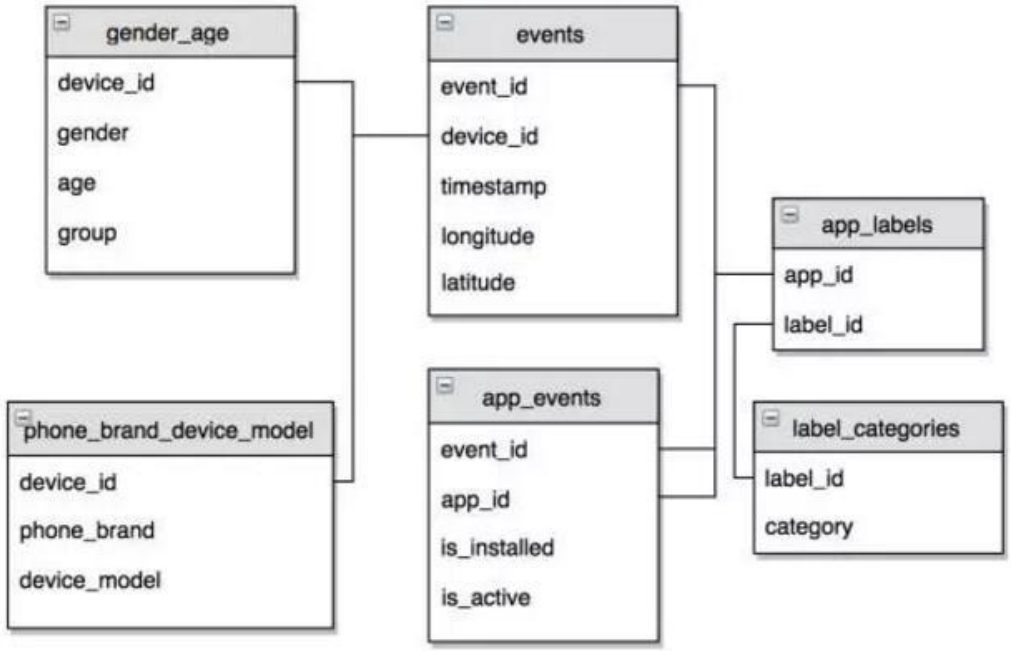
示例代码：lecture09_proj

项目实战

项目介绍

- 数据集结构
- 数据集说明：

  每个用户用一个ID表示，一个用户的行为是在一系列的Events里面，每个Event里面的信息包括该ID行为发生的时间、地理坐标信息，安装的APP类型、手机型号类别等



- 涉及知识点：

  1．pandas多表连接，数据处理； 2．OneHot编码

  3．特征选择；4．交叉验证选择参数

示例代码：lecture09_proj

目录

课程总结

明确思路

数据收集

数据处理

数据分析

数据展现

Python数据分析 (升级版) 课程安排					
	标题	内容	模型	项目案例	上课时间
第一课	工作环境准备及数据分析建模理论基础	1. 课程介绍 2. Python语言基础及Python 3.x新特性 3. 使用NumPy和SciPy进行科学计算 4. 数据分析建模理论基础 a. 数据分析建模过程 b. 常用的数据分析建模工具		科技工作者心理健康数据分析 (Mental Health in Tech Survey)	2017/02/18 15:00-17:00
第二课	数据采集与操作	1. 本地数据的采集与操作 a. 常用格式的本地数据读写 b. Python的数据库基本操作 2. 网络数据的获取与表示 a. BeautifulSoup解析网页 b. 爬虫框架Scrapy基础	回归分析-- Logistic回归	获取国内城市空气质量指数数据	2017/02/19 15:00-17:00
第三课	数据分析工具Pandas	1. Pandas的数据结构 2. Pandas的数据操作 a. 数据的导入、导出 b. 数据的过滤筛选 c. 索引及多重索引 3. Pandas统计计算和描述 4. 数据的分组与聚合 5. 数据清洗、合并、转化和重构	聚类模型 -- K-Means	全球食品数据分析 (World Food Facts)	2017/02/25 15:00-17:00
第四课	数据可视化	1. Matplotlib绘图 2. Pandas绘图 3. Seaborn绘图 4. 交互式数据可视化 -- Bokeh绘图		世界高峰数据可视化 (World's Highest Mountains)	2017/02/26 15:00-17:00
第五课	时间序列数据分析	1. Python的日期和时间处理及操作 2. Pandas的时间序列数据处理及操作 3. 时间数据重采样 4. 时间序列数据统计 -- 滑动窗口	时序模型 -- ARIMA	股票数据分析	2017/03/04 15:00-17:00
第六课	文本数据分析	1. Python文本分析工具NLTK 2. 分词 3. 情感分析 4. 文本分类	分类与预测模型 -- 朴素贝叶斯	微博情感分析	2017/03/05 15:00-17:00
第七课	图像数据处理及分析	1. 基本的图像操作和处理 2. 常用的图像特征描述	分类与预测模型 -- 人工神经网络	电影口碑与海报图像的相关性分析	2017/03/11 15:00-17:00
第八课	机器学习基础及机器学习库scikit-learn	1. 机器学习基础 2. Python机器学习库scikit-learn 3. 特征降维 -- 主成分分析		识别Twitter用户性别 (Twitter User Gender Classification)	2017/03/12 15:00-17:00
第九课	项目实战	1. 交叉验证及参数调整 2. 特征选择 3. 项目实操 4. 课程总结		通过移动设备行为数据预测使用者的性别和年龄	2017/03/18 15:00-17:00

参考

- scikit-learn中过拟合与欠拟合的例子
  http://scikit-learn.org/stable/auto_examples/model_selection/plot_underfitting_overfitting.html
- sklearn中的学习曲线
  http://scikit-learn.org/stable/modules/learning_curve.html
- 利用sklearn选择模型和参数
  http://scikit-learn.org/stable/tutorial/statistical_inference/model_selection.html
- 利用sklearn选择特征
  http://scikit-learn.org/stable/modules/feature_selection.html

参考

- 利用scikit-learn绘制roc曲线
  http://scikit-learn.org/stable/auto_examples/model_selection/plot_roc.html
- 什么是logloss
  https://www.kaggle.com/wiki/LogLoss
- sklearn中的logloss计算
  http://scikit-learn.org/stable/modules/generated/sklearn.metrics.log_loss.html

疑问

- 问题答疑： http://www.xxwenda.com/
  - 可邀请老师或者其他人回答问题

小象问答 @Robin_TY

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
- 新浪微博：ChinaHadoop