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

第八讲

终极项目：“闪电约会”配对预测

--梁斌

目录

- 过拟合与欠拟合
- 交叉验证补充
- 评价指标补充
- 终极项目：“闪电约会”配对预测
- 课程总结

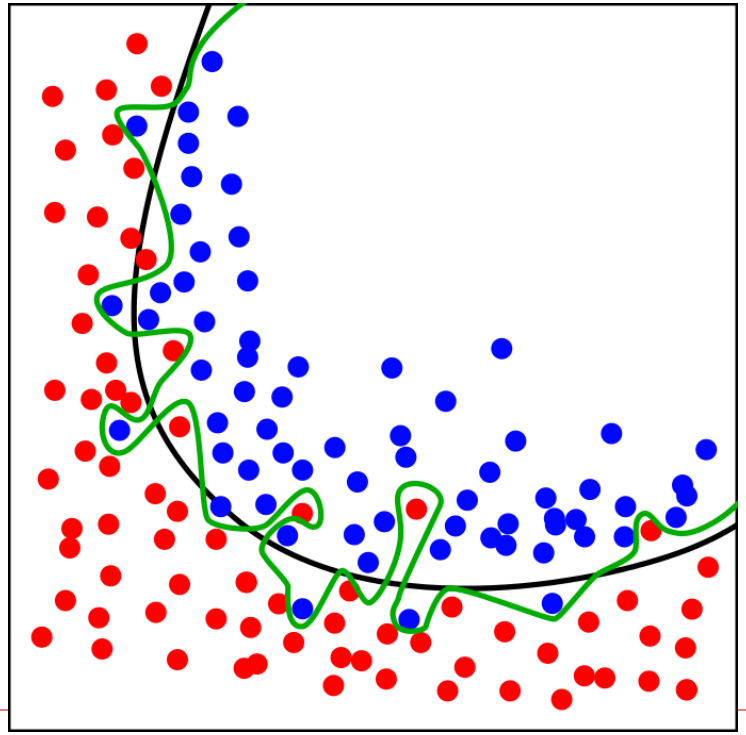
目录

- 过拟合与欠拟合
- 交叉验证补充
- 评价指标补充
- 终极项目：“闪电约会”配对预测
- 课程总结

过拟合与欠拟合

过拟合（Overfitting）

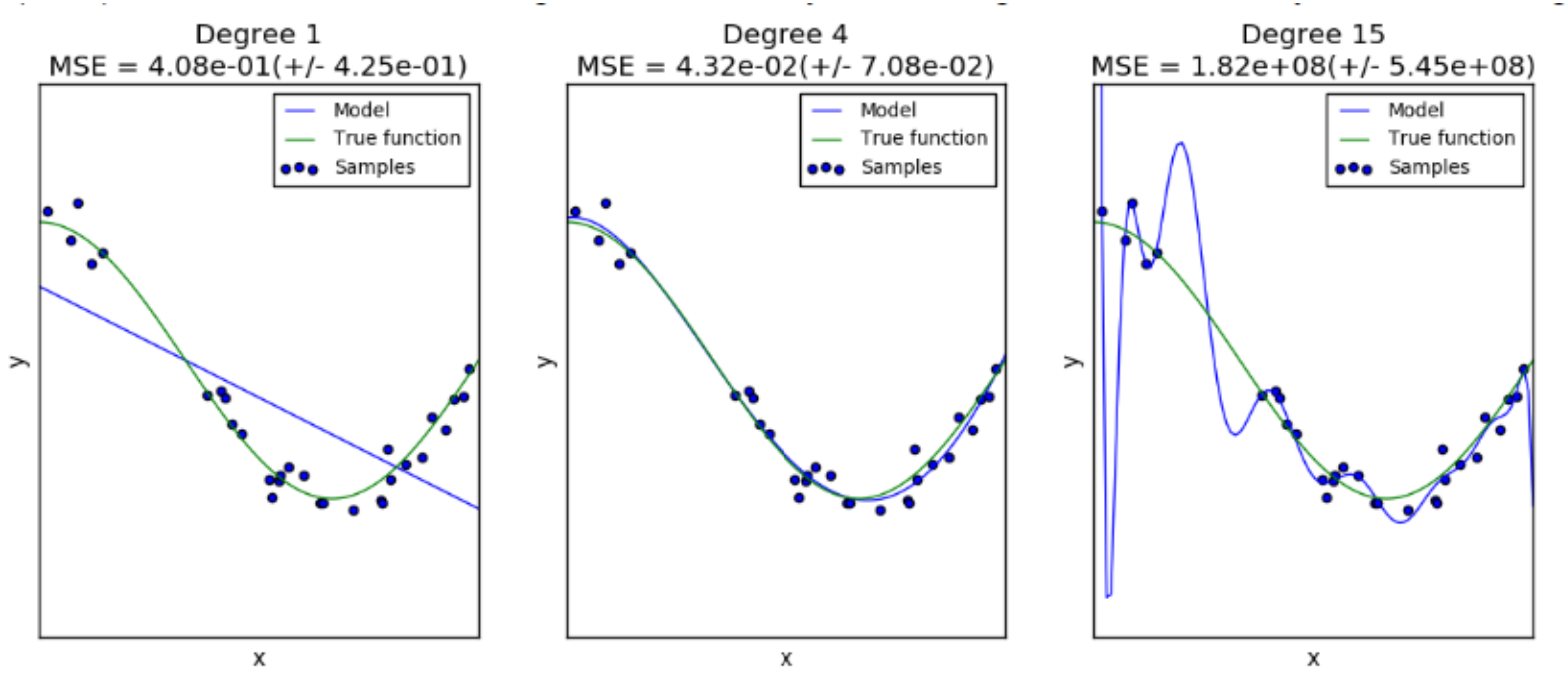
- 是指在调适一个统计模型时，使用过多参数。模型对于训练数据拟合程度过当，以致太适应训练数据而非一般情况。
- 在训练数据上表现非常好，但是在测试数据或验证数据上表现很差。



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

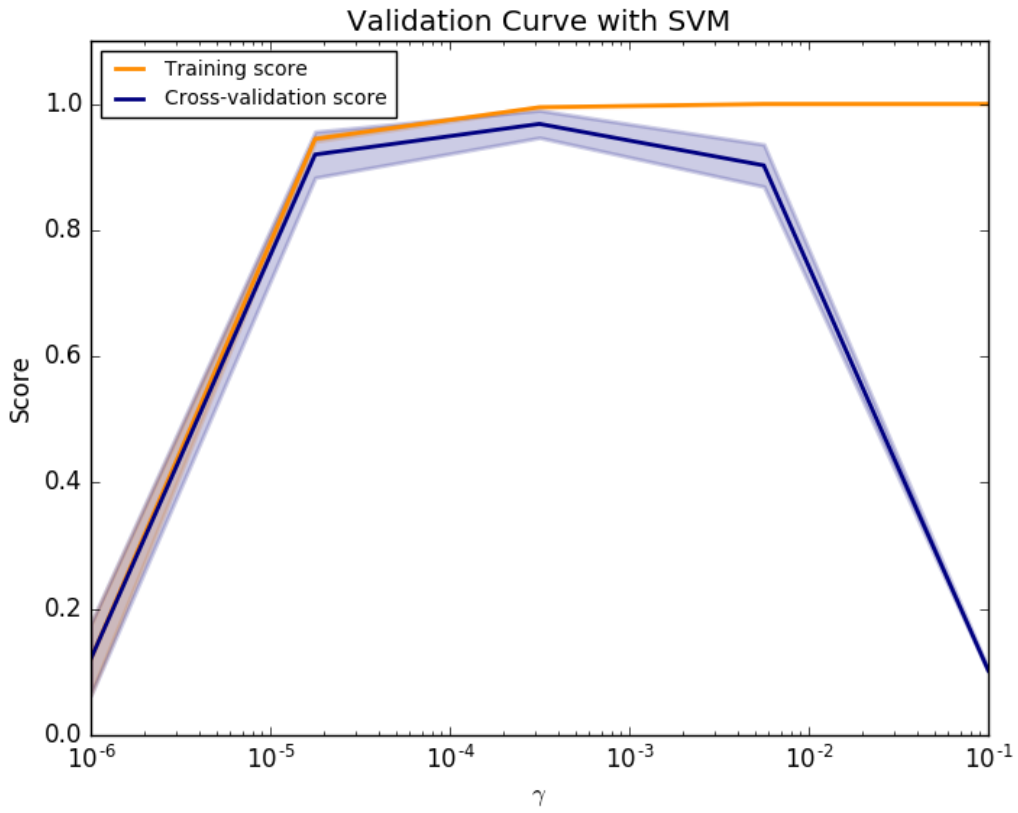
目录

- 过拟合与欠拟合
- 交叉验证补充
- 评价指标补充
- 终极项目：“闪电约会”配对预测
- 课程总结

交叉验证补充

验证曲线（validation curve）

- `sklearn.model_selection.validation_curve`



示例代码：02_val_curve.ipynb

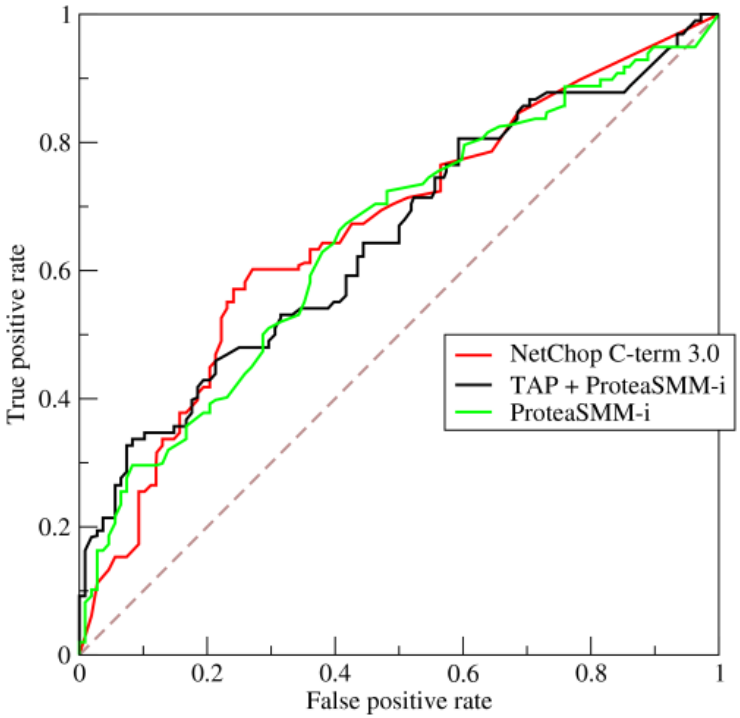
目录

- 过拟合与欠拟合
- 交叉验证补充
- 评价指标补充
- 终极项目：“闪电约会”配对预测
- 课程总结

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？



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

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

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

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

目录

- 过拟合与欠拟合
- 交叉验证补充
- 终极项目：“闪电约会”配对预测
- 课程总结

终极项目

项目介绍

- https://www.kaggle.com/annavictoria/speed-dating-experiment
- 数据采集自2002-2004“闪电约会”实验
  - 参与者有4分钟时间与异性交流
  - 4分钟后参与者回答是否愿意同该异性再次约会
  - 同时双方需要为对方的6个属性进行评分：
    1. 吸引力(Attractiveness),
    2. 忠诚实(Sincerity),
    3. 智慧(Intelligence),
    4. 幽默(Fun),
    5. 野心(Ambition),
    6. 共同爱好(Shared Interest)

示例代码：lecture09_proj.zip

终极项目

项目介绍（续）

- 该数据集也包括来自约会过程中不同时间点的问卷调查数据：
  1. 人口统计学信息(demographics)
  2. 约会习惯(dating habits)
  3. 自我认知(self-perception across key attributes)
  4. 信仰(beliefs on what others find valuable in a mate)
  5. 生活方式(lifestyle information)

项目任务

- 配对预测
- 掌握交叉验证
- 掌握ROC曲线的绘制

示例代码：lecture09_proj.zip

终极项目

分析步骤

1．查看数据集

2．明确分析目标

3．处理缺失数据

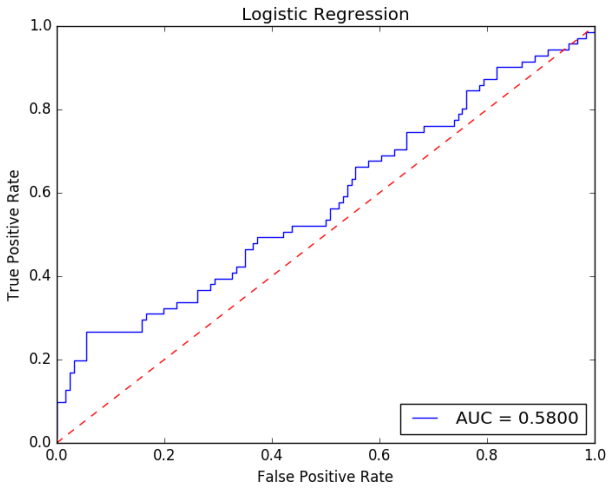
4．数据处理、重构

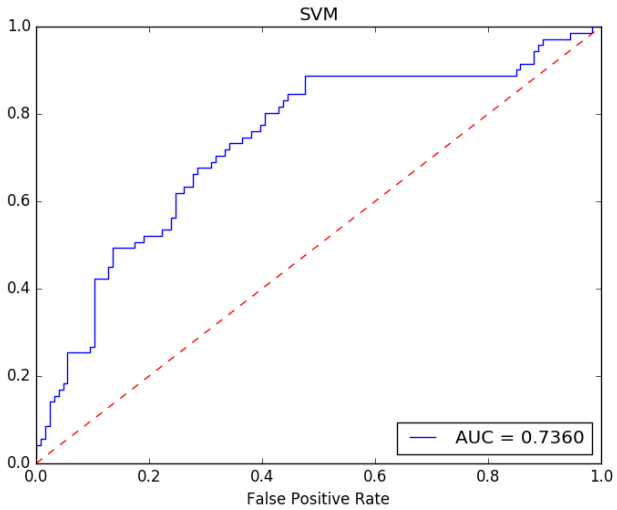
5．选择模型、特征

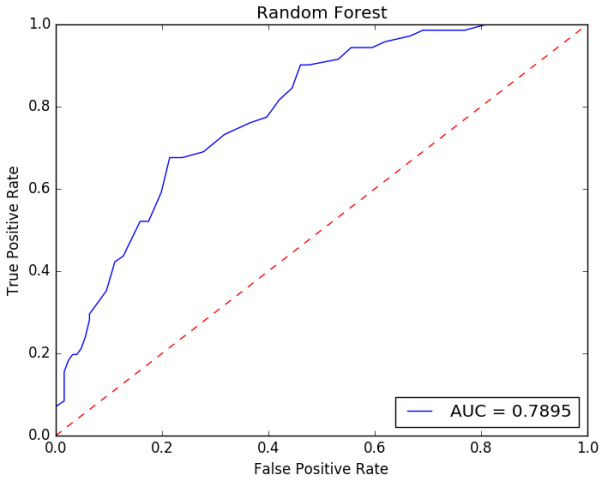
- 训练模型
- 交叉验证

6．保存分析结果

- 评价模型（ROC）





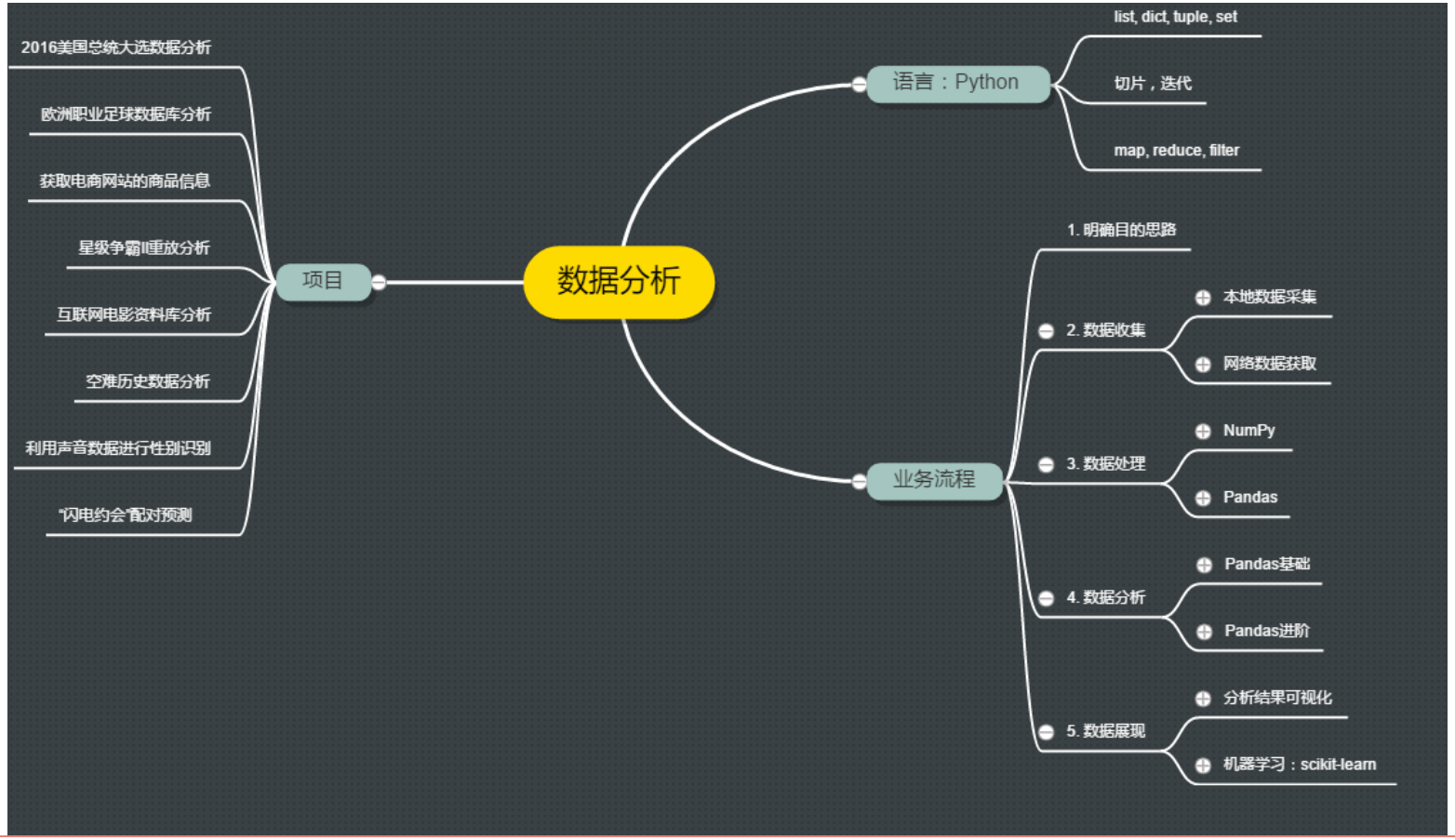


示例代码：lecture09_proj.zip

目录

课程总结

http://naotu.baidu.com/file/f61e8b7e404540b503a403229d11f7e0?token=b219e14f2d92ef79



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

- 项目参考论文

  Fisman, Raymond, et al. "*Gender differences in mate selection: Evidence from a speed dating experiment*." The Quarterly Journal of Economics (2006): 673-697.

- 利用scikit-learn绘制roc曲线

  http://scikit-learn.org/stable/auto_examples/model_selection/plot_roc.html

- 项目参考自

  https://www.kaggle.com/samshipengs/d/annavictoria/speed-dating-experiment/predict-match-between-two-person-v1/notebook

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