清华大学数据库技术与应用

实用机器学习 I

授课教师：计算机系王健楠

授课学期：2026年（春季）

清華大學 Tsinghua University

课程大纲

清華大學 Tsinghua University

01
异常检测

02
自动机器学习

03
可解释机器学习

课程大纲

清華大學 Tsinghua University

01 异常检测

- 异常检测基本概念与分类
- 应用案例：网络入侵检测

02 自动机器学习

03 可解释机器学习

什么是异常检测?

异常检测（Anomaly Detection）字典定义

a·nom·a·ly
/əˈnäməlē/

noun

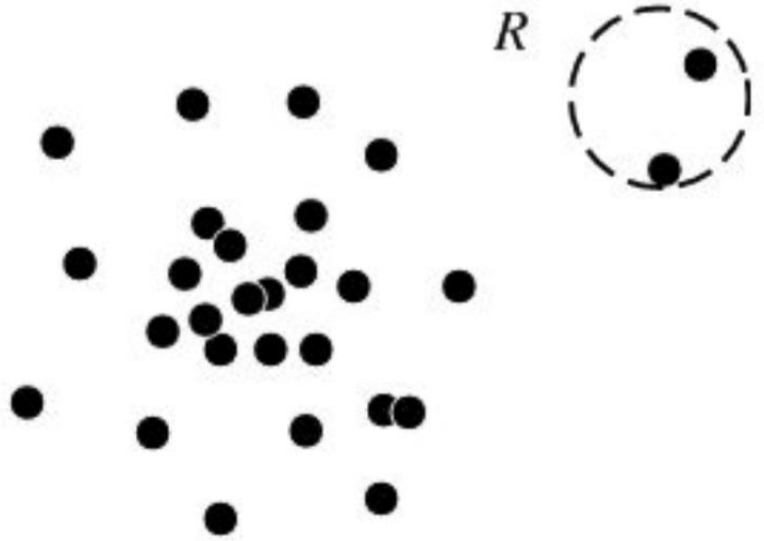
1. something that deviates from what is standard, normal, or expected.

也称为离群点检测（Outlier Detection）

异常类别（一）

全局异常

- 一个数据点相对于**整体数据**是异常的
- 例：某人的年龄是110岁

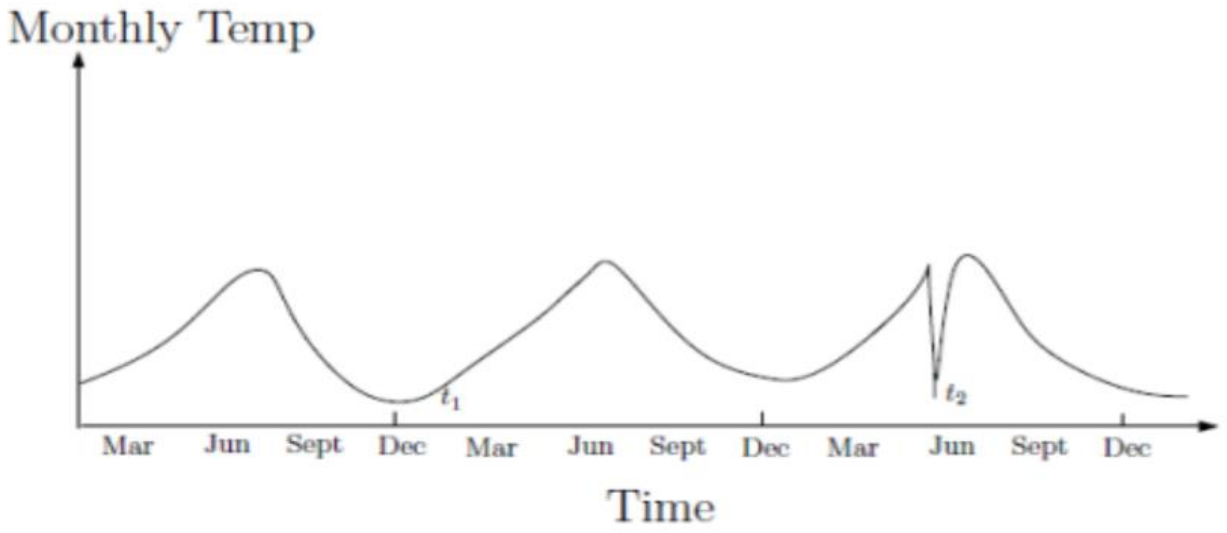


The objects in region R are outliers.

异常类别（二）

上下文异常（Context Anomaly）

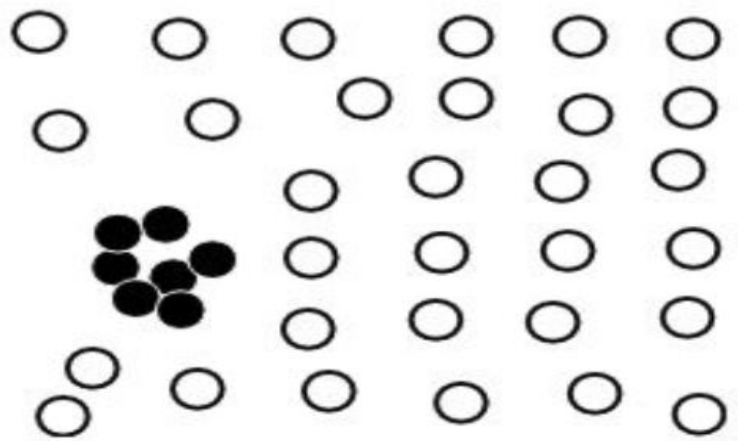
- 一个数据点在**特定上下文**中是异常的
- 例：班级中有一位10岁的学生



异常类别（三）

集体异常（Collective Anomaly）

- 数据子集整体显著偏离整个数据集
- 例：一个订单可能有延迟。但如果1000个订单都出现延迟呢?



The black objects form a collective outlier.

异常检测的实际应用

- 欺诈检测
- 医疗健康
- 公共安全
- 网络入侵检测

异常检测的挑战

清華大學 Tsinghua University

1. 建模正常对象和异常的有效性：

- 难以穷举所有可能的正常行为
- 正常对象和异常之间的边界可能是灰色地带

2. 应用特定的异常检测：

- 难以开发通用目的的异常检测工具

3. 可理解性：

- 不仅要检测异常，还要理解为什么它们是异常的

课程大纲

清華大學 Tsinghua University

01

异常检测

- 异常检测基本概念与分类
- ✓ **应用案例：网络入侵检测**

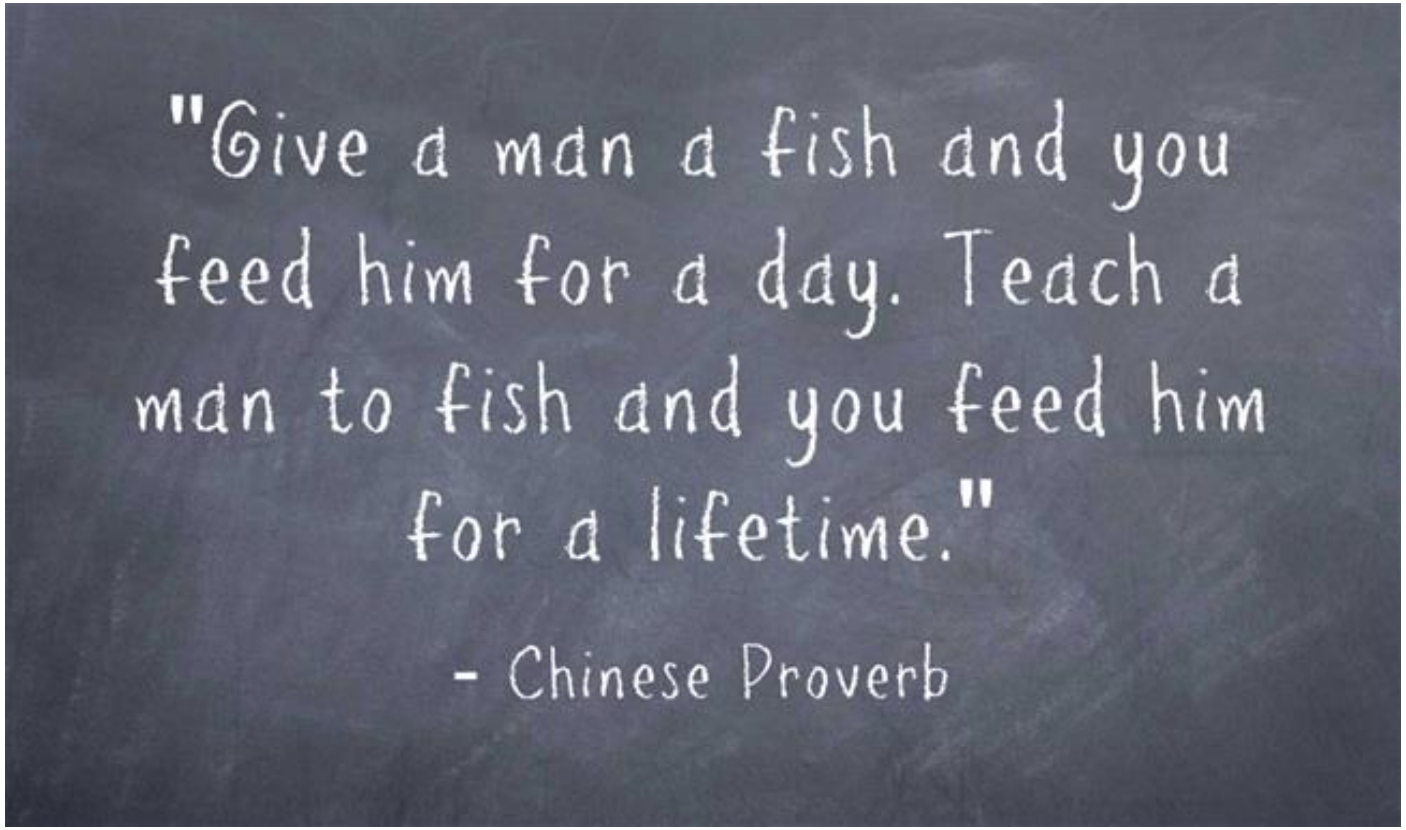
02

描述性统计

03

推断性统计

应用案例：网络入侵检测



教你一个现成的网络入侵解决方案 **vs.** **教你如何一步步想出这个解决方案**

（授人以鱼）　　**（授人以渔）**

网络入侵检测

"我们的Web服务器昨天遭到了攻击。
我不想让它再次发生。
请构建一个系统来解决这个问题！"

TODO 列表：

1. 找到相关数据集（如 /var/log/apache2/access.log）
2. 找出如何检测攻击（异常）←—— **关键问题**
3. 检测到攻击时触发警报（如发送邮件）

如何设计解决方案?

1. 调研相关工作

异常检测方法

综述论文

Anomaly detection: A survey

V Chandola, A Banerjee, V Kumar - ACM computing surveys (CSUR), 2009 - dl.acm.org

Abstract Anomaly detection is an important problem that has been researched within diverse research areas and application domains. Many anomaly detection techniques have been specifically developed for certain application domains, while others are more generic. This

☆ 99 Cited by 4705 Related articles All 36 versions

Intrusion **detection**: A **survey**

F Sabahi, A Movaghar - Systems and Networks …, 2008 - ieeexplore.ieee.org

… presents a taxonomy of intrusion **detection** systems that is then used to **survey** and classify … This method works by using the definition "**anomalies** are not normal … There are many **anomaly detection** that proposed algorithms with differences in the information used for analysis and …

☆ 99 Cited by 117 Related articles All 4 versions

1. 监督学习（例：情感分析）
2. 无监督学习（例：发现Twitter上的热门话题）

为什么无监督学习更常用?

无需标注数据

- 标注是一个繁琐且昂贵的过程

能够识别"未知的未知"

- 不仅检测已知攻击模式（如红苹果）
- 还能检测未知攻击模式（如西瓜）

如何设计解决方案?

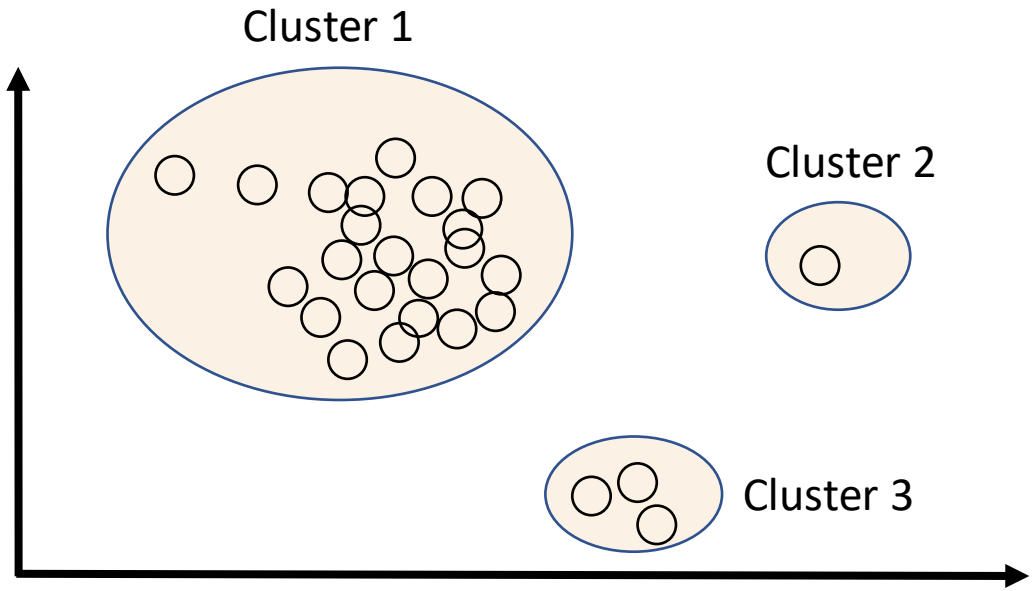
1.调研相关工作

2.选择无监督学习方法

基于聚类的方法

基本思想：

- 将数据点聚类为不同的组
- 判定哪些点是异常的：
  - 小Cluster中的数据点
  - 与最近聚类中心距离远的数据点



如何设计解决方案?

1.调研相关工作

2.选择无监督学习方法

3.选择聚类算法

K-Means 聚类算法

迭代算法

算法概述：

1. 随机选择K个点作为聚类中心
2. 将每个数据点分配到最近的中心
3. 相应更新聚类中心
4. 重复步骤2和3，直到满足终止条件

如何设计解决方案?

1.调研相关工作

2.选择无监督学习方法

3.选择聚类算法

4.选择和转换特征

特征提取

原始数据

```
1 in24.inetnebr.com - - [01/Aug/1995:00:00:01 -0400] "GET /shuttle/missions/sts-68/news/sts-68-mcc-05
2 uplherc.upl.com - - [01/Aug/1995:00:00:07 -0400] "GET / HTTP/1.0" 304 0
3 uplherc.upl.com - - [01/Aug/1995:00:00:08 -0400] "GET /images/ksclogo-medium.gif HTTP/1.0" 304 0
4 uplherc.upl.com - - [01/Aug/1995:00:00:08 -0400] "GET /images/MOSAIC-logosmall.gif HTTP/1.0" 304 0
```

原始数据 → 连接数据

- 一个连接是一系列HTTP请求，有明确的起止时间

连接数据 → 特征向量

- 需要大量领域知识
- 思考如何区分攻击与正常连接 （例：失败登录次数、连接持续时间等）

特征缩放（Feature Scaling）

特征向量

- 例：[http, BC, 0, 105, 146, 0, … , 0.00, 0.00]

这个特征向量能直接用于KMeans吗?

- "http" 和 "ftp" 之间的距离是多少?
- 数值范围大的特征会主导距离计算（如第4个特征）

分类特征 → 数值特征

简单方案：

- http → 0
- ftp → 1
- ssh → 2

One-Hot 编码：

- http → [1,0,0]
- ftp → [0,1,0]
- ssh → [0,0,1]

简单方案的缺点?

Distance("http", "ssh") > Distance("http", "ftp")

隐含了错误的距离关系

数值特征缩放

1. 归一化

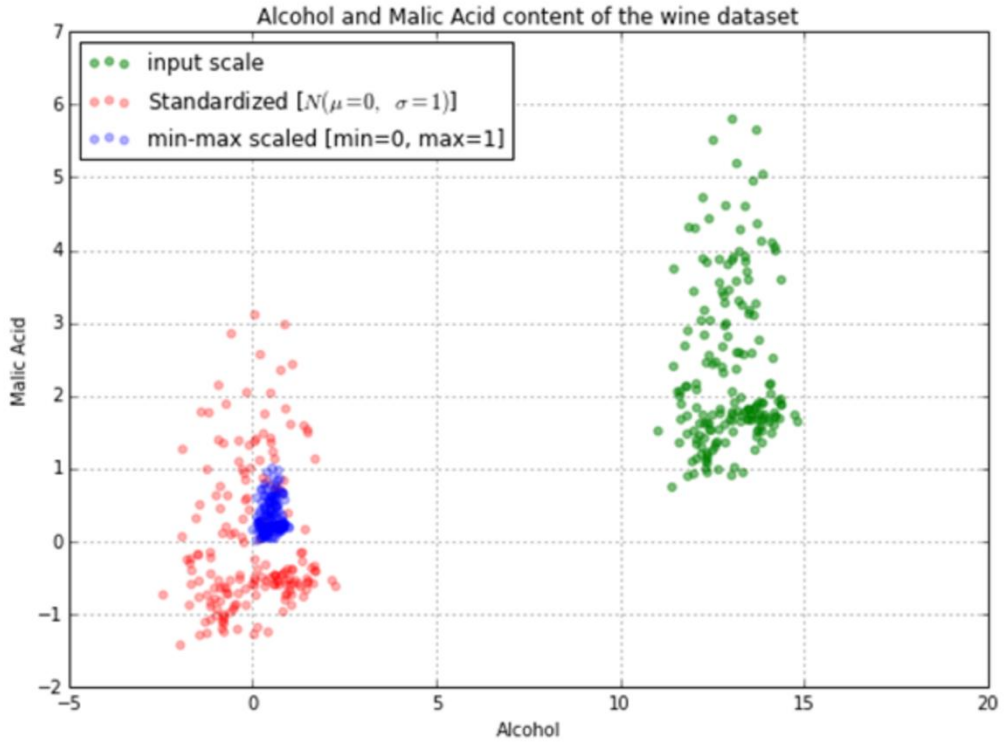
$$x' = \frac{x - min(x)}{max(x) - min(x)}$$

2. 标准化

$$x' = \frac{x - \bar{x}}{\sigma}$$

3. 缩放到单位长度

$$x' = \frac{x}{||x||}$$



The impact of Standardization and Normalisation on the Wine dataset

用哪种方法？取决于需求：

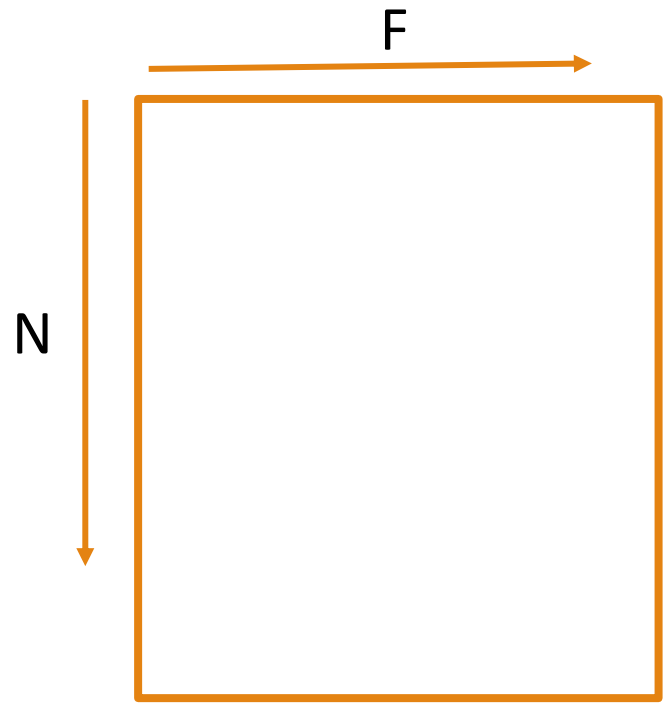
- 严格边界约束：(1), (3)
- 基于树的模型：不需要缩放

特征选择

特征选择：是什么？为什么？

数据通常以表格形式存在

- N: 训练样本数（如推文、图片）
- F: 特征数（如词袋、颜色直方图）



特征选择

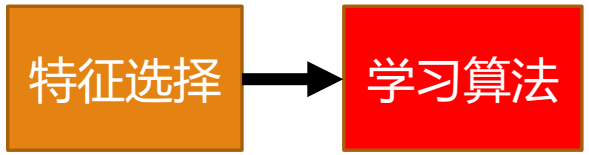
- 选择一个特征子集用于模型构建

"大F"有什么问题?

- 慢（训练时间长）
- 不准确（过拟合和维度灾难）
- 模型难以解释

特征选择方法

过滤法



包装法

嵌入法

过滤法

基本思想：

- 为每个特征打分
- 基于分数过滤掉无用特征

常用评分指标 [见 Yang and Pederson '97]：

- 分类任务：卡方检验、信息增益、文档频率
- 回归任务：相关性、互信息

包装法

基本思想：

- 评估特征子集
- 选择最佳子集

如何评估一个特征子集?

- 测试误差（通过交叉验证估计）

如何找到最佳子集?

- 贪心算法（如前向选择、后向消除）

嵌入法

基本思想：

- 修改学习算法，使其能自动惩罚无用特征

典型方法：Lasso 回归

- 通过L1正则化惩罚无用特征

$$\underset{\beta}{\operatorname{argmin}} \ \|y - X\beta\|_2^2 + \lambda\|\beta\|_1$$

惩罚无用的特征

三种方法的比较

过滤法

✓ 高效
✓ 不易过拟合
✗ 无法捕捉特征间的关系

包装法

✓ 能捕捉特征间的关系
✗ 效率低
✗ 可能过拟合

嵌入法

✓ 结合了上述两种方法的优点
✗ 依赖于特定的学习算法

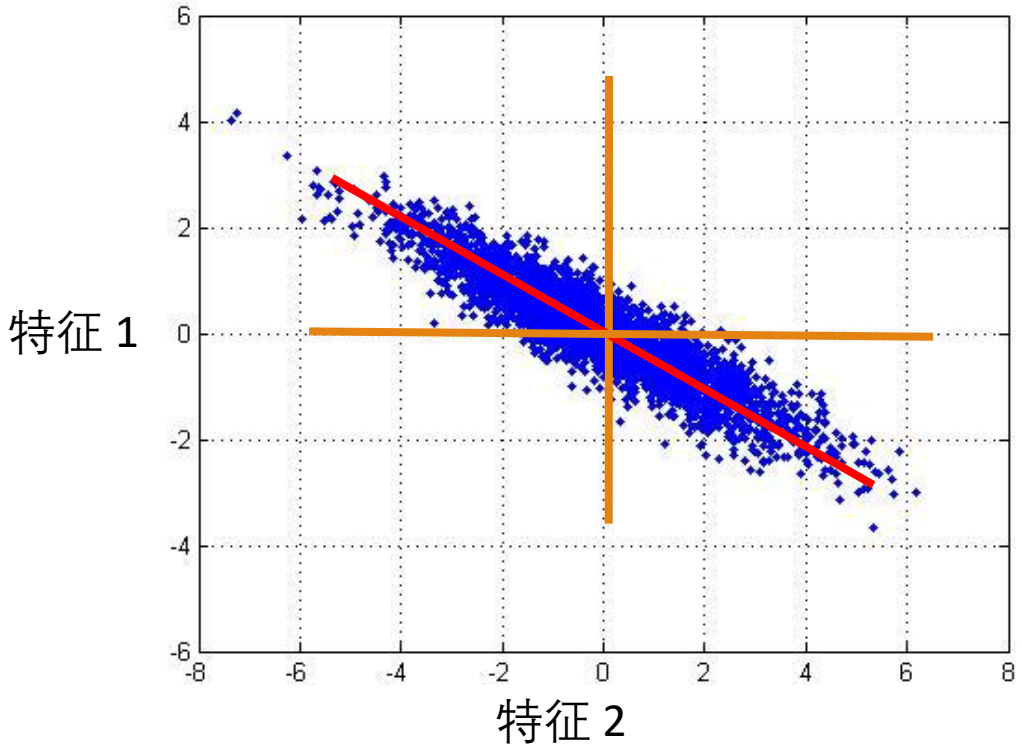
降维

特征选择

- 新特征必须是原始特征的子集

特征变换（如PCA）

- 新特征可以不是原始特征的子集
- 通过线性/非线性组合产生新的特征



如何设计解决方案?

1.调研相关工作

2.选择无监督学习方法

3.选择聚类算法

4.选择和转换特征

5.参数调优与评估

参数调优与评估

评估

- 真实标签?
- 评估指标?

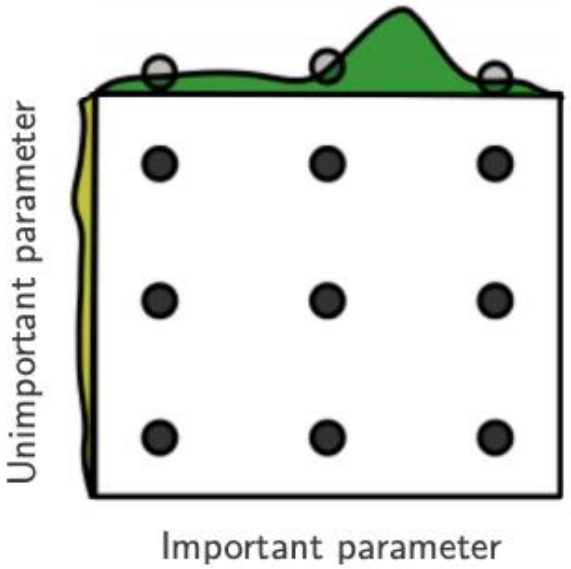
参数调优

- 网格搜索
- 随机搜索
- 贝叶斯优化

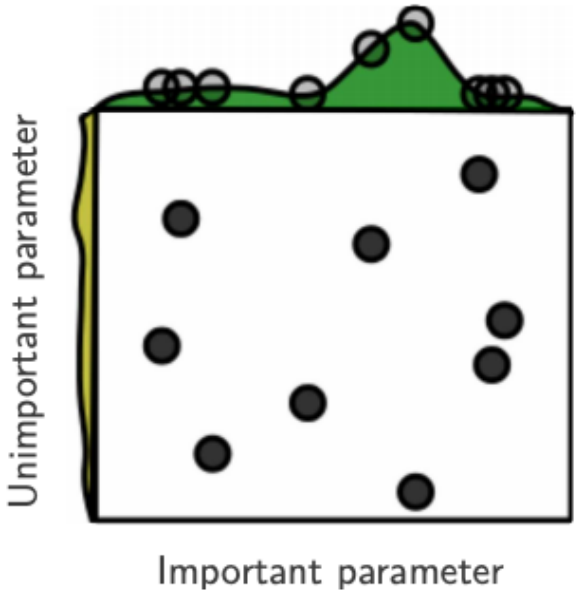
网格搜索与随机搜索

x: 工作时间 (1, 2, ..., 12)
y: 睡眠时间 (1, 2, ..., 12)

Income(x, y) = Work(x) + Sleep(y)



网格搜索

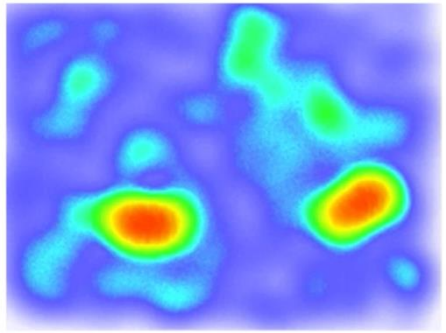


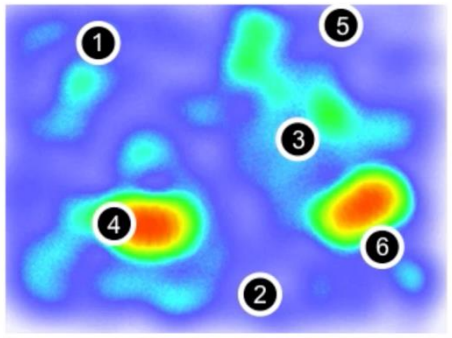
随机搜索

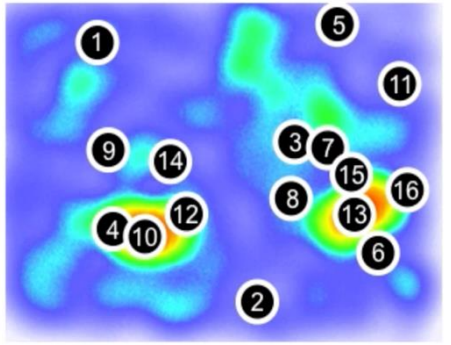
贝叶斯优化

直观理解：

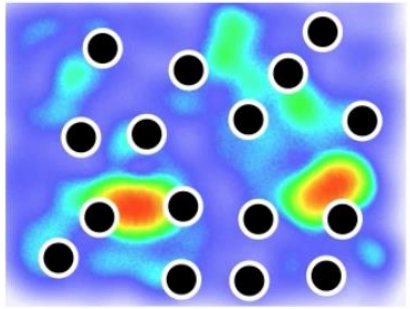
- 目标： 想要找到目标函数的峰值点（例如：寻找能使模型准确率达到最高的参数组合）。
- 方法： 为已观测到的点拟合一个统计模型，并根据该模型挑选我们认为最可能是最大值的下一个点。
- 决策： 下一个点的选择由采集函数 决定，该函数会在探索 (Exploration) 和利用 (Exploitation) 之间进行权衡。



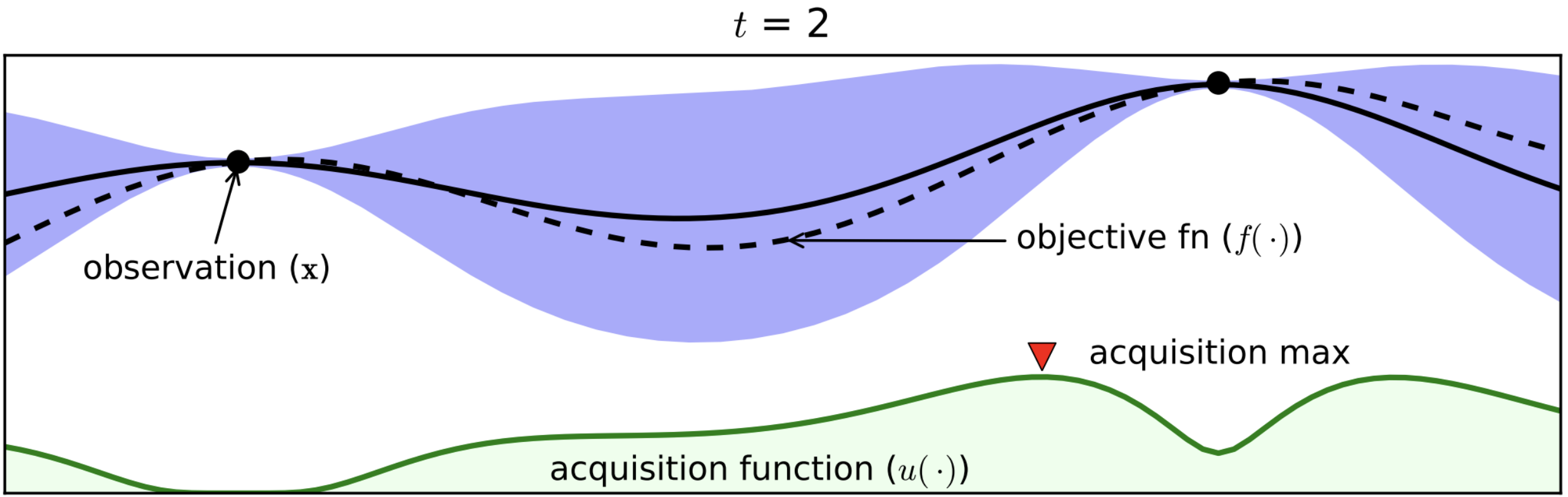




Random

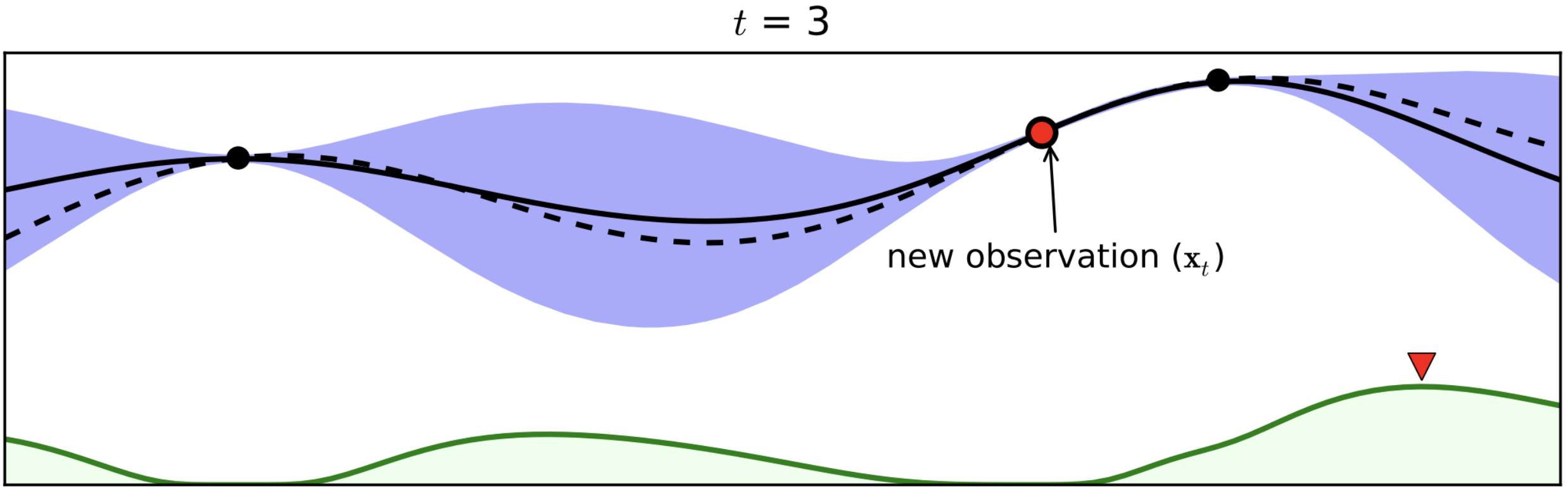


贝叶斯优化示例（一）



找到使采集函数最大化的下一个采样点

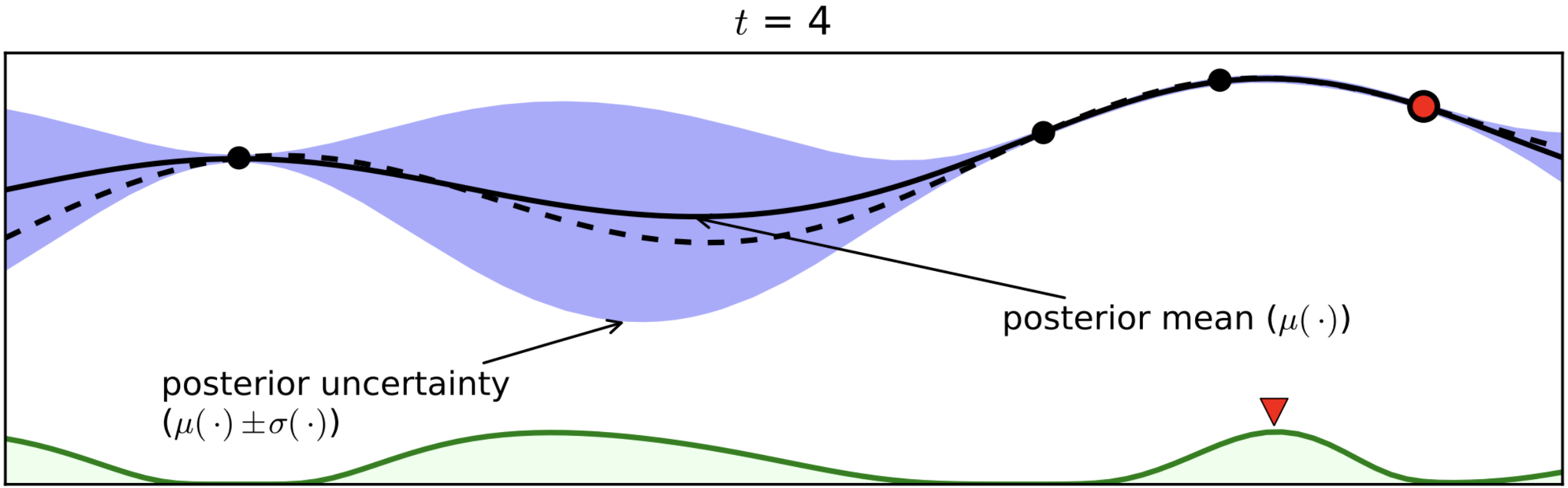
贝叶斯优化示例（二）



在新的观测点 x_t 处进行评估（计算真实值），并更新后验概率。

根据新的后验概率更新采集函数，并寻找下一个最佳点。

贝叶斯优化示例（三）


$t = 4$
posterior mean ($\mu(\cdot)$)
posterior uncertainty
($\mu(\cdot) \pm \sigma(\cdot)$)

如何设计解决方案?

1.调研相关工作

2.选择无监督学习方法

3.选择聚类算法

4.选择和转换特征

5.参数调优与评估

6.不满意？回到前面的步骤

如何设计解决方案?

1. 调研相关工作
2. 选择无监督学习方法
3. 选择聚类算法
4. 选择和转换特征
5. 参数调优与评估
6. 不满意？回到前面的步骤
7. **将模型部署到生产环境**

模型部署（Model Serving）

清華大學 Tsinghua University

非常重要的话题！

模型需要反映最新的数据更新

- KMeans → Streaming Kmeans

预测需要实时进行

- Amazon SageMaker
- TensorFlow Serving
- Azure Machine Learning
- Kubeflow 中的 ML 模型部署
- MLflow Model Serving on Databricks

第一部分小结

如何设计解决方案?

1. 调研相关工作
2. 选择无监督学习方法
3. 选择聚类算法

理论算法

4. 选择和转换特征
5. 参数调优与评估
6. 不满意？回到前面的步骤

实际应用

模型开发

7. 将模型部署到生产环境

模型部署