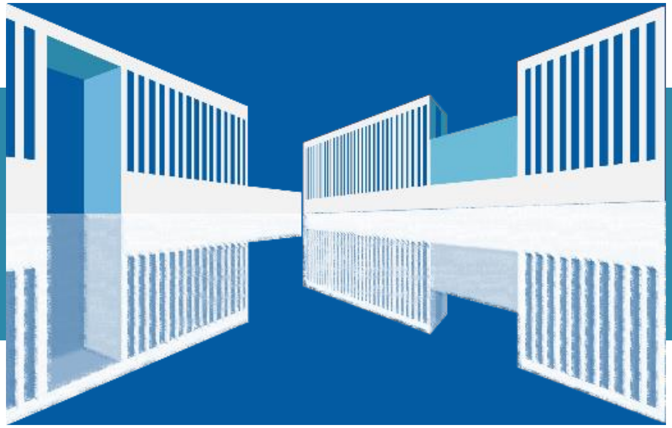
北京航空航天大学
BEIHANG UNIVERSITY

人工智能及其航空应用

第十二章 异常检测——火箭发动机异常检测

申晓斌



目录

0 介绍

- 在日常生活中，我们通常需要注意一些与大多数现象不同的异常数据。
- 异常数据
  - 与预期模式不匹配
  - 与事件或观测的正常值偏差过大

问题	应用案例
银行欺诈	识别使用地址检测信用卡诈骗
结构缺陷	机械加工领域识别未达标的产品
网站维护	网络通信领域识别异常信息流

异常检测算法---半监督学习方法

目录

1 火箭发动机异常检测背景

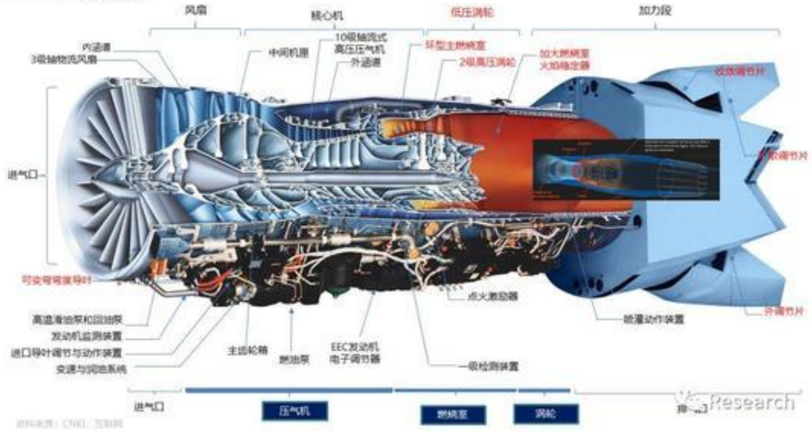
喷气式发动机

仅依靠从发动机内部向飞行器后高速喷出物质流（主要是气流）而获得反作用推力

空气喷气发动机

- 利用大气作为主要物质之一来产生喷射气流因而产生推力
- 工作范围：大气层内

图：F119-PW涡扇发动机



火箭发动机

- 不利用周围介质，只利用飞行器自身携带的物质生成工作物质（简称工质）产生推力
- 工作范围：大气层内、外

图片全来自互联网

1 火箭发动机异常检测背景

伴随着航天事业的飞速发展和航天发射任务的日益频繁，各种安全事故不可避免地接踵而来。这其中又以火箭发动机故障的影响和危害性最大，其复杂的工作环境和几近极致的工作条件也常使它成为整个航天运输系统中故障敏感多发的部位。发动机故障不仅会带来巨大的经济损失，而且会导致灾难性的事故，产生难以估量的影响。

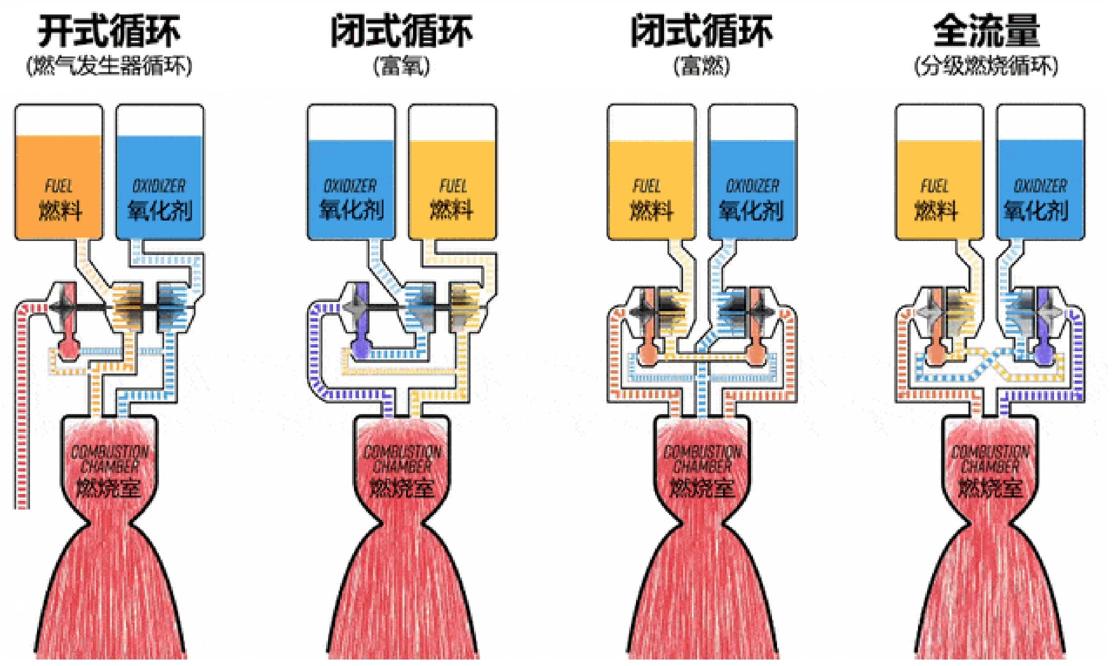
1 火箭发动机异常检测背景

液体火箭具有发动机燃料比冲高、推重比大、推力可调，关机和启动灵活、燃料成本低，回收再使用优势大等优点；

液体火箭发动机结构复杂，包括了燃料输送管、高速涡轮泵、燃烧室和喷管、燃料冷却槽道等多个重要部分，制造难度极大；

液体火箭发动机的检查与诊断方法及技术：

- 基于信号分析的方法
- 基于模型的方法
- 基于人工智能的方法



目录

2 异常检测理论

2.1 异常检测概念

2.2 高斯分布与异常检测

2.3 基于高斯分布的异常检测算法

2.4 异常检测开发与调试

2.5 异常检测与监督学习的对比

2.6 异常检测特征的选择

2.1 异常检测的概念

- **异常检测(Anomaly detection)，是机器学习算法的一个常见应用。**
- **它虽然主要用于非监督学习问题，但从某些角度看，它又类似于一些监督学习问题。**

半监督学习方法

2 异常检测理论

2.1 异常检测的概念

- **异常检测就是发现与大部分对象不同的对象（发现离群点）**
- **异常对象的属性值明显偏离期望的属性值，异常检测也称偏差检测；**
- **异常在某种意义上是一种例外，也被称为例外挖掘；异常对象是相对罕见的。**

异常分类

1）**点异常**（Point Anomalies）：是指单个数据对象相对于其他数据对象异常。点异常是最简单，也是研究得最多的异常类型。
2）**上下文异常**（Contextual Anomalies）：是指一个数据对象在特定的上下文中的异常，也称为条件异常（Conditional Anomaly）。数据集的内部结构定义了上下文，而且成为异常问题定义的一部分，它包含上下文属性和行为属性两部分。
3）**集合异常**（Collective Anomalies）：是指一批相关的数据对象相对于整个数据集是异常的。集合异常中的各个数据对象可能自身不是异常，但它们作为一个集合整体出现时，则是异常。

2 异常检测理论

2.1 异常检测的概念

异常检测主要使用数理统计和数据挖掘技术，算法主要有：

基于模型、**基于邻近度**、**基于密度**和**基于聚类**。

- 基于模型的技术：建立一个数据模型，异常是那些同模型不能完美拟合的对象。例如，数据分布的模型可以通过估计概率分布的参数来创建。如果一个对象不服从该分布，则认为它是一个异常。

优点

- ✓ 具有坚实的基础，即建立在标准的统计学基础之上；
- ✓ 当存在充分的数据和有效的先验知识时，表现得非常好；
- ✓ 该方法简单，无须训练，可以用在小数据集上。

缺点

- ➢ 对于多元数据，可用的分布选择太少；
- ➢ 对于高维数据，基本不可能拟合出数据分布；
- ➢ 离群点对模型参数影响很大。

2.1 异常检测的概念

异常检测主要使用数理统计和数据挖掘技术，算法主要有：

基于模型、**基于邻近度**、**基于密度**和**基于聚类**。

- 基于邻近度的技术：在对象之间定义邻近性度量，异常对象是那些远离大部分其他对象的对象。当数据能够以二维或者三维散布图呈现时，可以从视觉上检测出基于距离的离群点。
- 基于密度的技术：对象的密度估计可以相对直接计算，特别是当对象之间存在邻近性度量。低密度区域中的对象相对远离近邻，可能被看做为异常。

优点

- ✓ 原理简单，无须训练，可用在任何数据集；
- ✓ 定量度量，结果相对准确。

缺点

- ➢ 阈值确定困难；
- ➢ 时间复杂度为O(n^2)。

2 异常检测理论

2.1 异常检测的概念

异常检测主要使用数理统计和数据挖掘技术，算法主要有：

基于模型、**基于邻近度**、**基于密度**和**基于聚类**。

- 基于聚类的技术：聚类和异常检测的目标是估计分布的参数，以最大化数据的总似然（概率）。聚类分析用于发现强相关的对象组，异常检测是发现与其他对象弱相关的对象，因此，聚类可以用于异常检测。

优点

- ✓ 对于时间和空间复杂度是线性或接近线性的聚类，离群点检测技术是高度有效的
- ✓ 可同时发现簇和离群点
- ✓ 在样本充足的情况下准确度会相对较高

缺点

- ➢ 非常依赖所用的簇的个数和数据总离群点的存在性
- ➢ 产生簇的质量对算法产生的离群点的质量影响非常大
- ➢ 每种聚类算法只适合特定的数据类型，需要谨慎地选择聚类算法

2.2 高斯分布与异常检测

介绍一种基于模型的异常检测技术，即基于**高斯分布**的异常检测算法

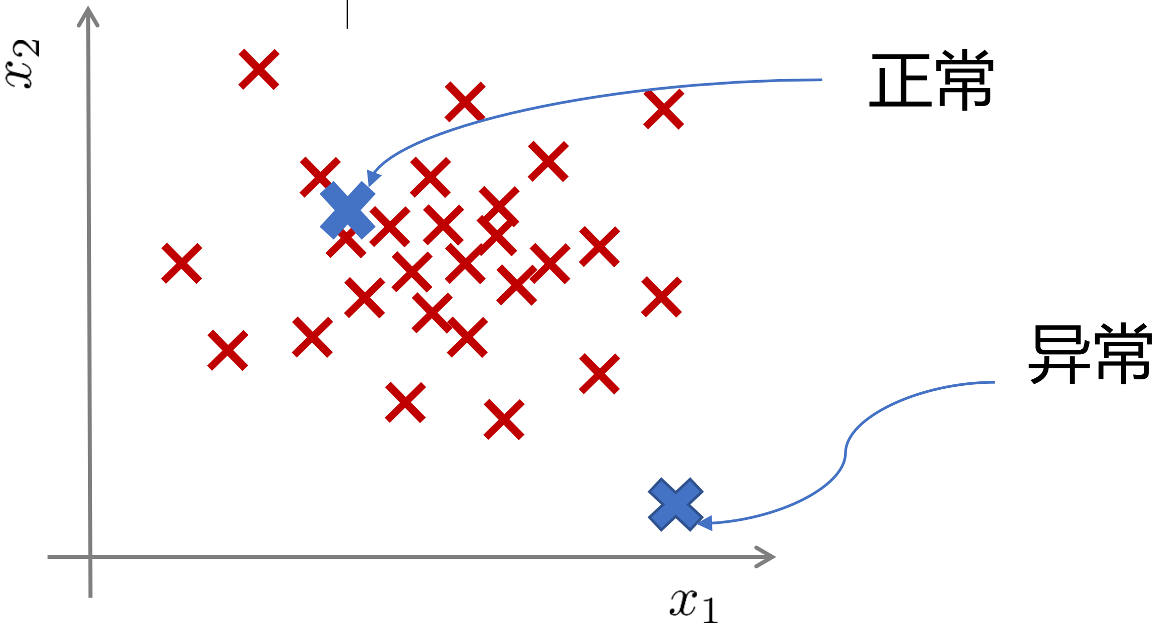
样本数据特征：

抽象形式

x_1 x_2

数据集：$\{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$

新数据：x_{test}



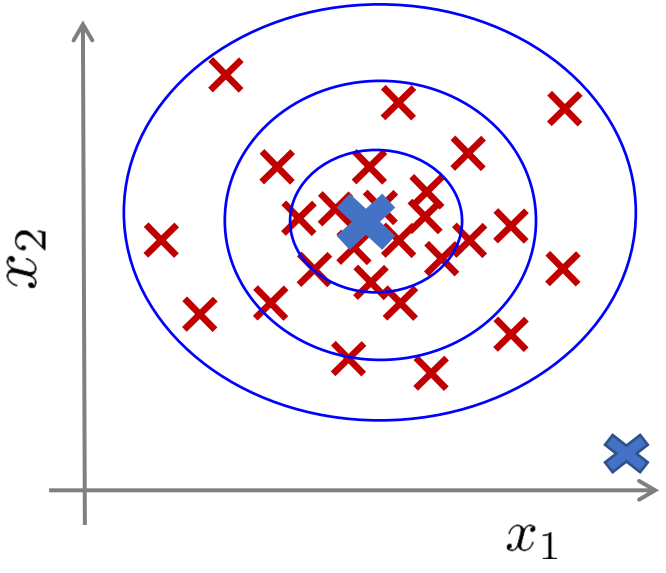
2.2 高斯分布与异常检测

概率估计

抽象形式

数据集: $\{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$

x_{test} 是否异常? 此测试数据属于该组数据的几率 $p(x)$



$p(x_{test}) < \varepsilon$ 记为异常

$p(x_{test}) \geq \varepsilon$ 正常

2.2 高斯分布与异常检测

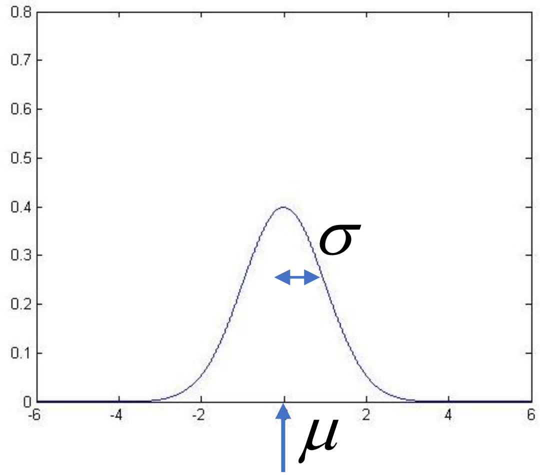
高斯分布（正态分布）

设$x \in \mathbb{R}$，若x符合高斯分布，即 $x \sim N(\mu, \sigma^2)$

则其概率密度函数为： $p(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{(x-\mu)^2}{2\sigma^2})$

平均值 $\mu = \frac{1}{m}\sum_{i=1}^{m} x^{(i)}$

方差 $\sigma^2 = \frac{1}{m}\sum_{i=1}^{m}(x^{(i)} - \mu)^2$ 标准差 σ

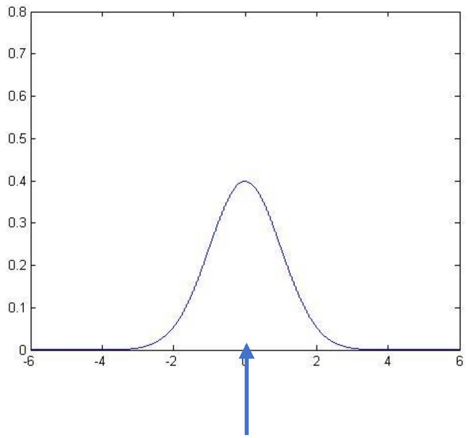


注：机器学习中对于方差我们通常只除以m而非统计学中的$(m-1)$。

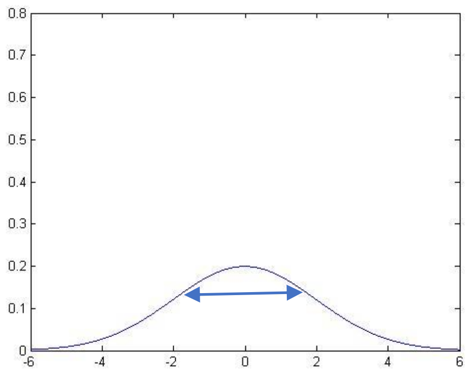
2.2 高斯分布与异常检测

高斯分布参数影响

$\mu = 0, \sigma = 1$



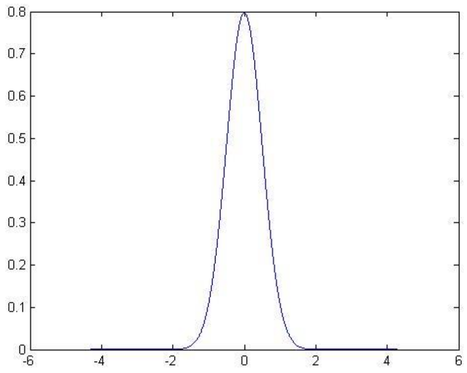
$\mu = 0, \sigma = 2$



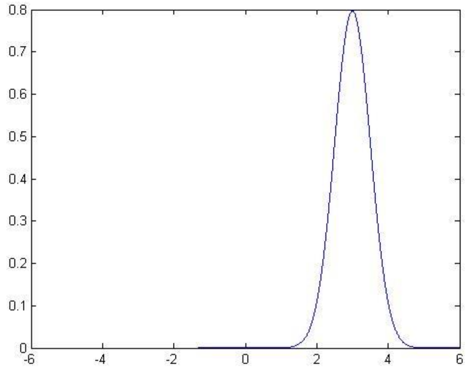
钟形围成的面积为1

$$\sigma^2 = \sigma \cdot \sigma$$

$\mu = 0, \sigma = 0.5$



$\mu = 3, \sigma = 0.5$

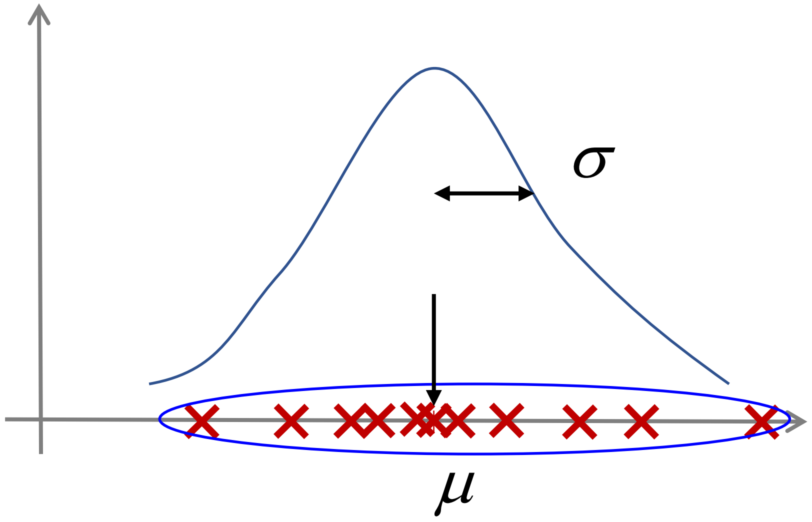


2.2 高斯分布与异常检测

参数估计

数据集：$\{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$　　$x^{(i)} \in \mathbb{R}$

$x^{(i)} \sim N(\mu, \sigma^2)$　　对高斯分布的参数进行估计



$$\mu = \frac{1}{m}\sum_{i=1}^{m} x^{(i)}$$

$$\sigma^2 = \frac{1}{m}\sum_{i=1}^{m}(x^{(i)} - \mu)^2$$

2.3基于高斯分布的异常检测算法

异常检测概率估计

训练集： $\{x^{(1)},\ldots,x^{(m)}\}$

每个样本都有 $x \in \mathbb{R}^n$

特征向量的每个特征都符合高斯分布

$$x_1 \sim N(\mu_1, \sigma_1^2)$$
$$x_2 \sim N(\mu_2, \sigma_2^2)$$
$$\cdots$$
$$x_n \sim N(\mu_n, \sigma_n^2)$$

每个特征独立，则训练集的概率模型为：

$$p(x) = p(x_1)p(x_2)\ldots p(x_n)$$

$$p(x) = p(x_1;\mu_1,\sigma_1^2)p(x_2;\mu_2,\sigma_2^2)\ldots p(x_n;\mu_n,\sigma_n^2)$$

即： $p(x) = \prod_{j=1}^{n} p(x_j;\mu_j,\sigma_j^2)$

2.3基于高斯分布的异常检测算法

异常检测算法的建立过程

1. 找到可能反映异常的特征x_i，建立数据集 $\{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$
2. 估计每个特征的高斯分布参数

$$\mu_1, \ldots, \mu_n, \sigma_1^2, \ldots, \sigma_n^2$$

$$\mu_j = \frac{1}{m}\sum_{i=1}^{m} x_j^{(i)}$$

也可以采用向量的形式来表示

$$\sigma_j^2 = \frac{1}{m}\sum_{i=1}^{m}(x_j^{(i)} - \mu_j)^2$$

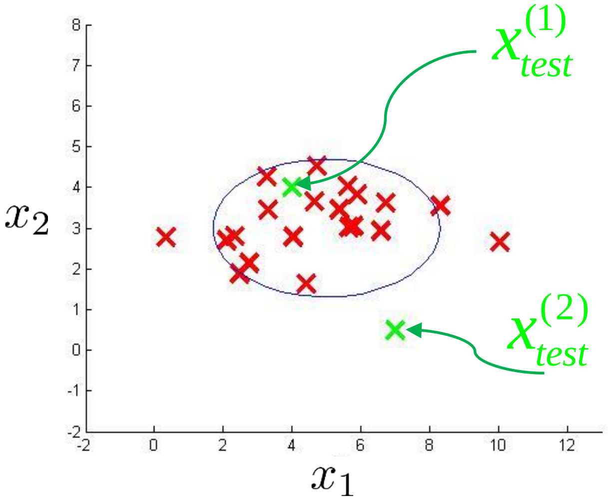
3. 对于新的待测数据x，计算概率 $p(x)$：

$$p(x) = \prod_{j=1}^{n} p(x_j; \mu_j, \sigma_j^2) = \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_j} \exp\left(-\frac{(x_j - \mu_j)^2}{2\sigma_j^2}\right)$$

若 $p(x) < \varepsilon$，则新待测数据为异常

2.3基于高斯分布的异常检测算法

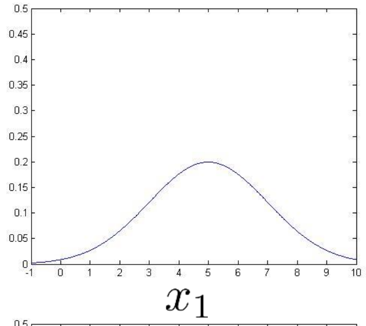
参数估计与异常检测



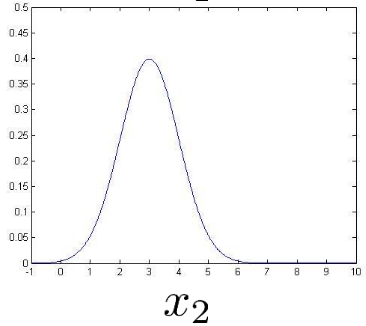
高斯参数：

$$\mu_1 = 5, \sigma_1 = 2$$

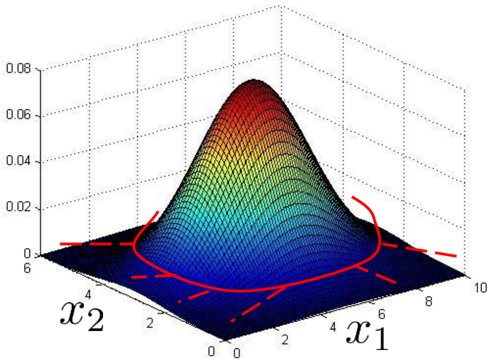
$$\mu_2 = 3, \sigma_2 = 1$$



$p(x_1; \mu_1, \sigma_1^2)$



$p(x_2; \mu_2, \sigma_2^2)$



$$p(x) = p(x_1; \mu_1, \sigma_1^2) \cdot p(x_2; \mu_2, \sigma_2^2)$$

设定： $\varepsilon = 0.02$

$$p(x_{test}^{(1)}) = 0.0426 \quad > \varepsilon$$

$$p(x_{test}^{(2)}) = 0.0021 \quad \leq \varepsilon$$

2.4 异常检测开发与调试

当进行一个学习算法的开发时（选择特征等），如果有一种此算法的评价手段，将更容易决定算法特征与参数。

异常检测算法具有非监督学习的特性，开发时无标签。意味着无法根据结果变量y 的值来确定数据是否真的是异常的。

需要已知带标签（异常或正常）的数据进行算法评价，类似于监督学习

半监督学习方法

异常检测的开发步骤：数据准备、数据分组、异常评估、异常输出

2.4 异常检测开发与调试

1）数据准备

- 了解输入数据的**特征**
- 除了正常的样本进行训练之外，也需要用异常的样本进行测试

  (正常数据：$y=0$　　异常数据：$y=1$)

- 异常样本通常**非常稀少**，一般只用标签为0的样本来进行训练

选取特征，准备：	10000 台	好的发动机 (正常数据)
	20 台	有缺陷的发动机(异常数据)

2.4 异常检测开发与调试

2）数据分组

- 在训练中可以看成一种无监督学习算法
- 根据变量判断数据是否异常，需要另一种方法来检测算法是否有效

选择训练集： $x^{(1)}, x^{(2)}, \ldots, x^{(m)}$ (全都为正常数据)

交叉验证集： $(x_{cv}^{(1)}, y_{cv}^{(1)}), \ldots, (x_{cv}^{(m_{cv})}, y_{cv}^{(m_{cv})})$ 含有 $y=1$

测试集: $(x_{test}^{(1)}, y_{test}^{(1)}), \ldots, (x_{test}^{(m_{test})}, y_{test}^{(m_{test})})$

训练集： 6000台 好的发动机数据($y=0$)

交叉验证集：2000台好的($y=0$)， 10台缺陷的($y=1$)

测试集：2000台好的($y=0$)， 10台缺陷的($y=1$)

2.4 异常检测开发与调试

3）异常评估

基于训练集$\{x^{(1)},\ldots,x^{(m)}\}$构建模型 $p(x)$

查准率(precision)：指被分类器判定正例中的正样本的比重 $\frac{\text{TP}}{\text{TP+FP}}$

查全率 (recall)：指的是被预测为正例的占总的正例的比重 $\frac{\text{TP}}{\text{TP+FN}}$

对于验证集或测试集，一般用的评价指标:

- 正确肯定(True positive, TP), 错误肯定(false positive, FP), 错误否定(false negative, FN),正确否定 (true negative, TN)
- 查准率（P），查全率 （R）
- F_1值（F1 Score，查准率和查全率的调和平均数, $2\frac{\text{PR}}{\text{P+R}}$）

对交叉检验集，尝试使用不同的ε值作为阈值，并预测数据是否异常，根据评价指标选择ε值

2.4 异常检测开发与调试

4）异常输出

任何异常检测技术最终都需要输出检测到的异常。通常，由异常检测算法产生的输出有以下两类。

1）评分（score）：评分技术通过异常判定函数为测试集中的每个数据实例赋予一个异常得分，这个得分代表了异常程度。因此输出是异常的排序列表，使用阈值来判定异常。

2）标签（label）：标签分类技术直接对每个测试实例贴上正常或异常的标签。

2.5 异常检测与监督学习的对比

- **异常检测算法用于非监督学习问题**
- **但在评价异常检测系统时也使用了带标签的数据，与监督学习有些相似**
- **通过对比有助于选择采用监督学习还是异常检测**

$y = 0$

2 异常检测理论

2.5 异常检测与监督学习的对比

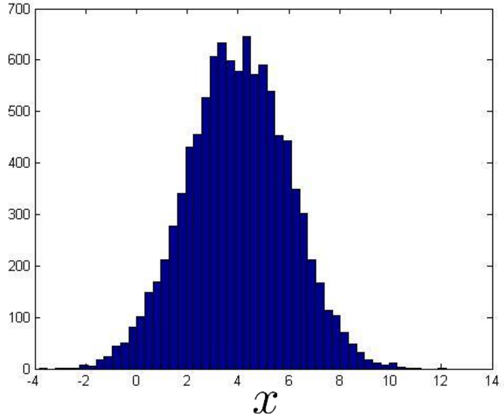
异常检测	监督学习
非常少量的正向类（异常数据 $y=1$），大量的负向类（$y=0$） (0-20)	同时有大量的正向类和负向类
许多不同种类的异常 非常难 根据非常 少量的正向类数据来训练算法。 $p(x)$	有足够多的正向类实例，足够用于训练 算法，未来遇到的正向类实例可能与训练集中的非常近似。
未来遇到的异常可能与已掌握的异常 非常的不同。	
例如：欺诈行为检测 生产（例如飞机引擎）检测数据中心的计算机运行状况	例如：邮件过滤器 天气预报 肿瘤分类

2.6 异常检测特征的选择

- 对于异常检测算法，使用的**特征**至关重要
- 异常检测中，假设特征都符合**高斯分布**
- 如果特征的分布不是高斯分布，异常检测算法**也能**够工作，但是最好还是将数据转换成高斯分布
- 一般使用的函数包括：**对数函数，指数函数**

2.6 异常检测特征的选择

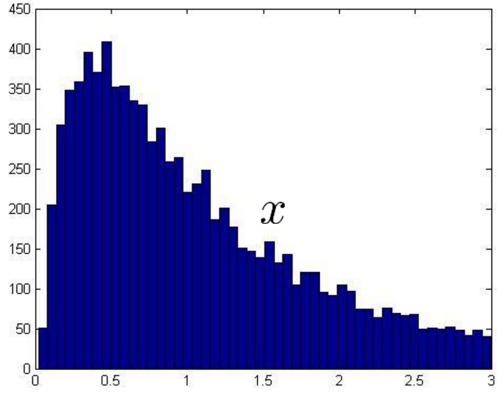
非高斯分布特征的转换



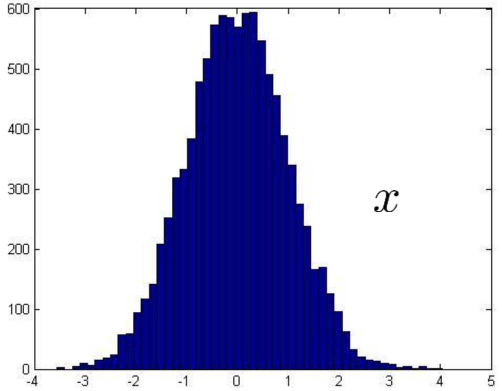
直方图

$$p(x_i;\mu_i,\sigma_i^2)$$

$$x_1 = \log(x_1)$$
$$x_2 = \log(x_2 + c)$$
$$x_3 = \sqrt{x_3} = x_3^{1/2}$$
$$x_4 = x_4^{1/3}$$



$\log(x)$



2.6 异常检测特征的选择

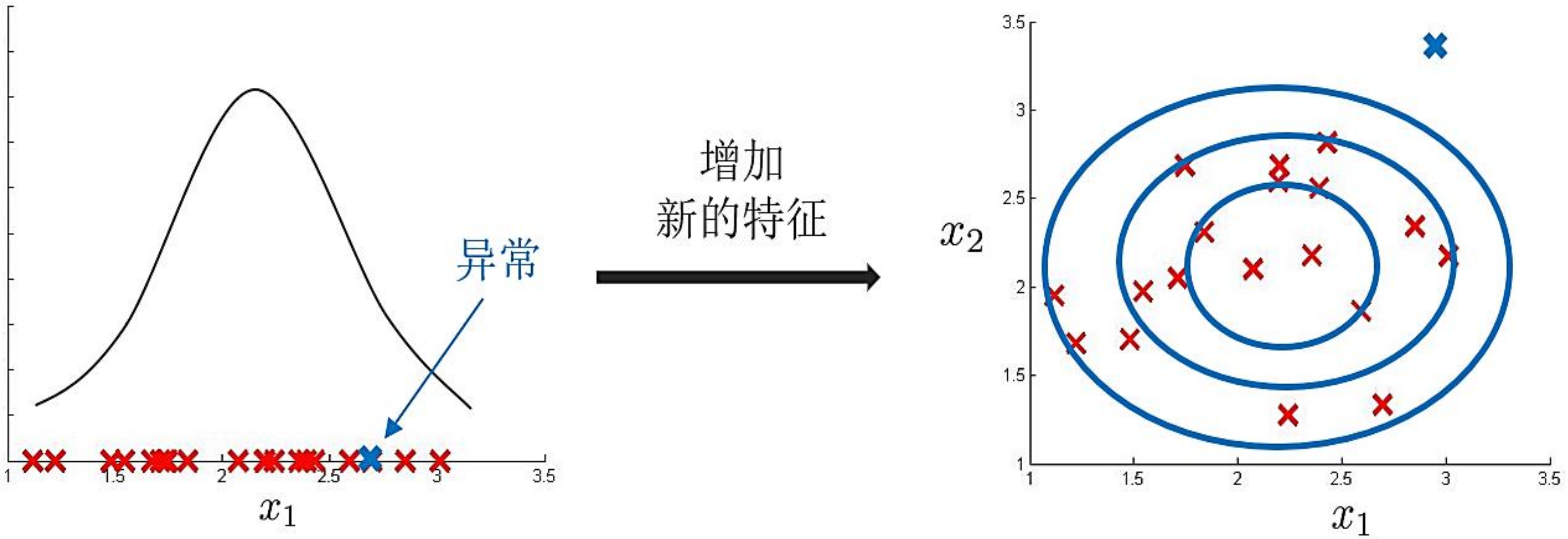
异常检测的误差分析

异常检测希望：x为正常数据时，$p(x)$较大。

x为异常数据时，$p(x)$较小。

但常见的问题是：

正常数据与异常数据预测得到的 $p(x)$结果相当。 （可能都很大）



2.6 异常检测特征的选择

数据中心的计算机检测

选择计算机发生异常时，以出现非正常数值对应的特征来进行异常检测系统的建立

x_1 = 内存使用

x_2 = 被访问的硬盘数量

x_3 = CPU负载

x_4 = 网络通讯量

$$x_5 = \frac{x_3}{x_4} = \frac{\text{CPU负载}}{\text{网络通讯量}}$$

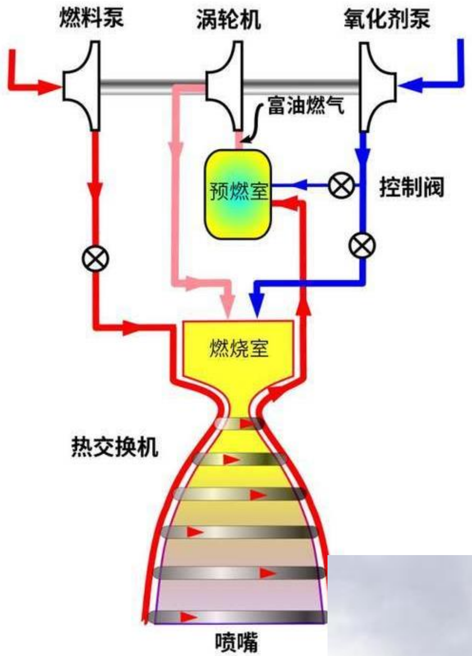
$$x_6 = \frac{x_3^2}{x_4} = \frac{\text{CPU负载}^2}{\text{网络通讯量}}$$

在检测数据中心的计算机状况时，可以用**CPU**负载与网络通信量的比例作为一个新的特征，如果该值异常大，便有可能意味着计算机是陷入了一些问题中

目录

3 液体火箭发动机异常检测案例

异常检测案例分析



为了保证运载火箭飞行的成功，在火箭发动机设计制造后，一般需要进行**地面试车**，对其工作状态及工作参数进行检测与分析。液体火箭具有发动机燃料比冲高、推重比大、推力可调，关机和启动灵活、燃料成本低，回收再使用优势大等优点，在航天发射任务中有重大的贡献。我国的**长征五号火箭**、**美国SpaceX的重型猎鹰火箭**等都采用了液体火箭发动机。

3 液体火箭发动机异常检测案例

异常检测案例分析

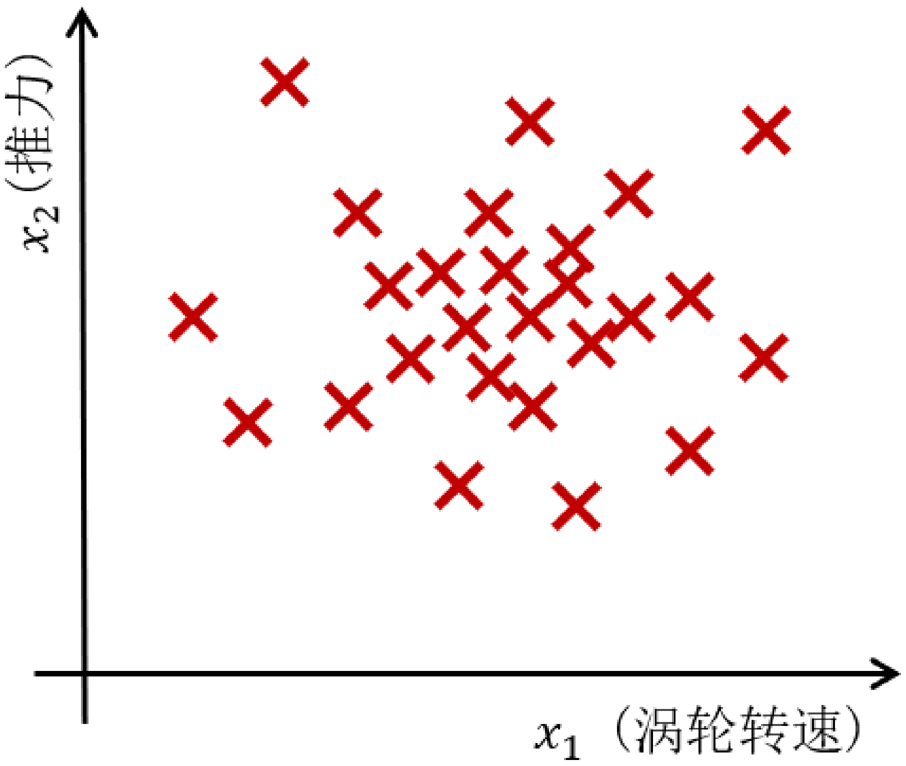
试车还可以为火箭批量生产的质量抽检和飞行试验的故障分析提供手段。液体火箭发动机在试车中的测量参数包括了**推进剂消耗量、结构的动应力和变形、振动、冲击、噪声、温度和压力**等，而大型火箭在一次试车中，需要测量数百个以至**数千个参数**，比飞行试验测量的参数更多。这些测量参数包括了发动机的一些**特征变量**，而特征变量则共同构成了液体火箭发动机的**特征向量**。

3 液体火箭发动机异常检测案例

异常检测案例分析

只考虑液体火箭发动机的涡轮转速（x_1）和推力(x_2)的异常

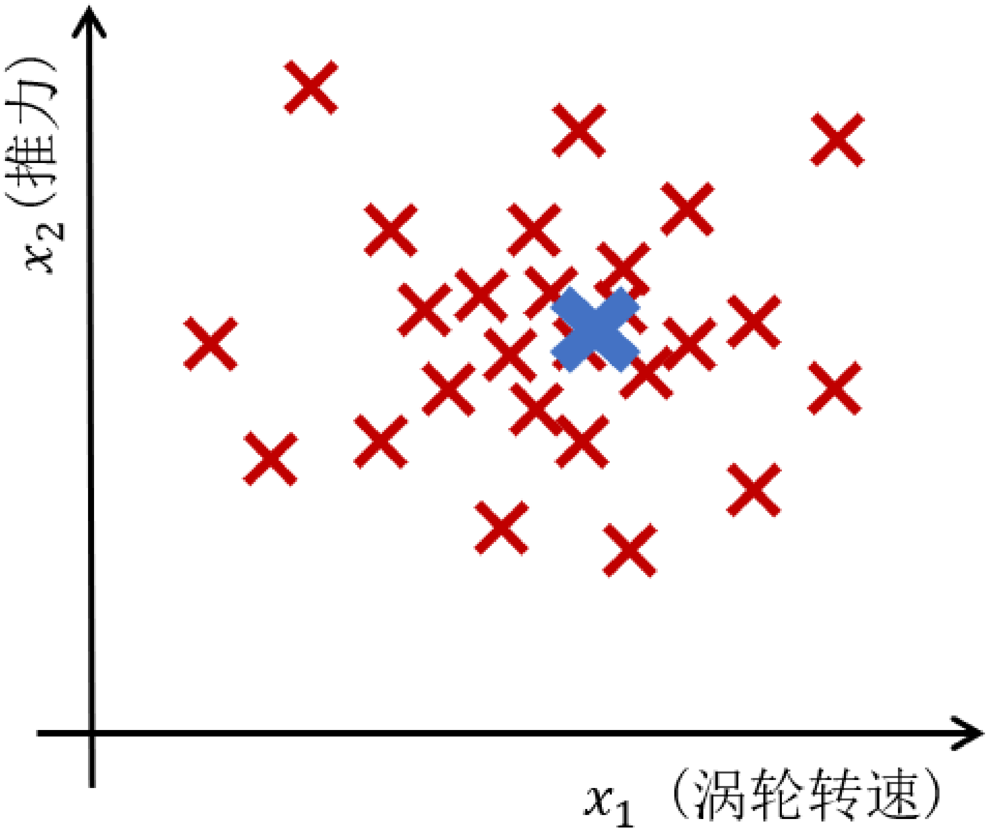
m台发动机测量获得的特征向量组成了一个数据集 $\left\{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\right\}$



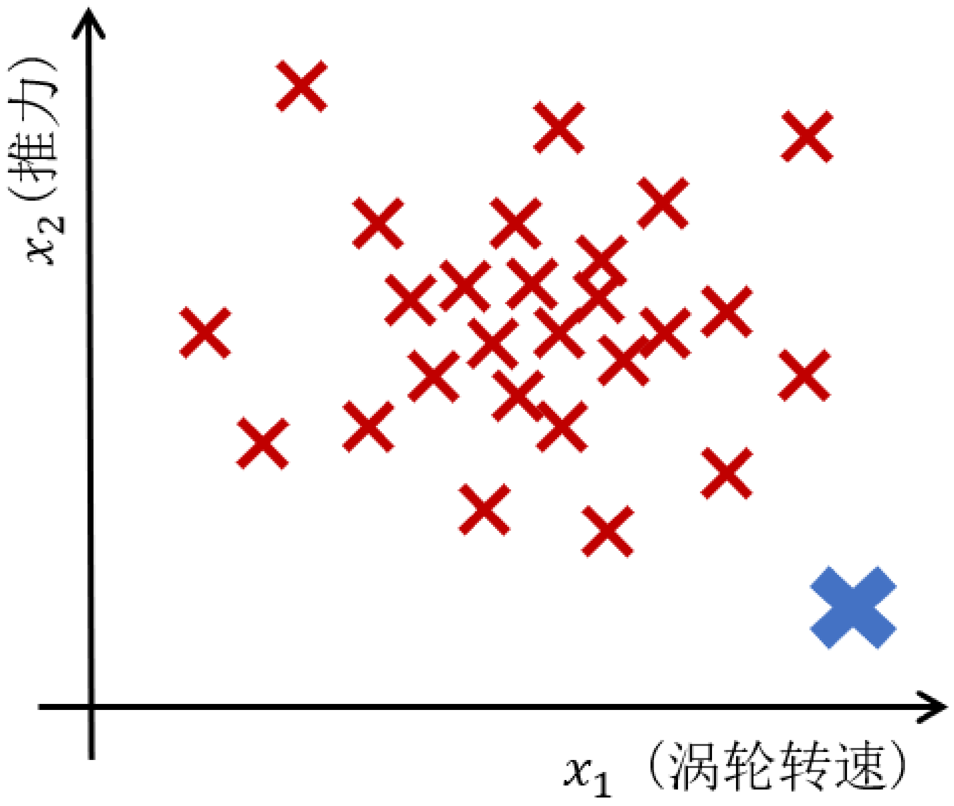
无标签数据

3 液体火箭发动机异常检测案例

异常检测案例分析



新的发动机特征点位置（正常）



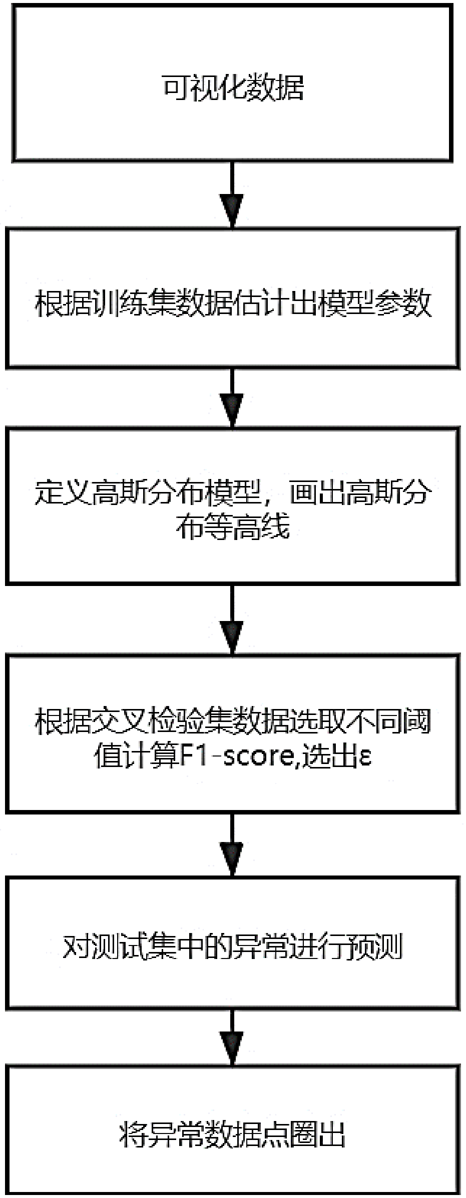
新的发动机特征点位置（异常）

异常检测程序流程

1. 根据训练集数据估计特征的平均值和方差并构建$p(x)$函数

2. 对交叉检验集，我们尝试使用不同的ε值作为阈值，并预测数据是否异常，根据 $F1$ 值或者查准率与查全率的比例来选择。选出ε后，针对测试集进行预测

3. 通过数据可视化以及 $F1$ 值判断模型检测能力



异常检测程序流程

先可视化数据

```python
data = sio.loadmat("data\\ex8data1.mat")
X = data['X']   # (307,2)
plt.scatter(X[..., 0], X[..., 1], marker='x', label='point')
plt.show()
```

根据数据估计出模型的参数

```python
def estimate_parameters_for_gaussian_distribution(X):
    """
    估计数据估计参数
    :param X: ndarray,数据
    :return: (ndarray,ndarray),均值和方差
    """
    mu = np.mean(X, axis=0)  # 计算方向因该是沿着0, 遍历每组数据
    sigma2 = np.var(X, axis=0)  # N-ddof为除数,ddof默认为0
    return mu, sigma2
```

异常检测程序流程

定义高斯概率模型

```python
def gaussian_distribution(X, mu, sigma2):
    """
    根据高斯模型参数，计算概率
    :param X: ndarray,数据
    :param mu: ndarray,均值
    :param sigma2: ndarray,方差
    :return: ndarray,概率
    """
    p = (1 / np.sqrt(2 * np.pi * sigma2)) * np.exp(-(X - mu) ** 2 / (2 * sigma2))
    return np.prod(p, axis=1)  # 横向累乘
```

根据估计出的参数，
画出高斯分布的等高线

```python
def visualize_contours(mu, sigma2):
    """
    画出高斯分布的等高线
    :param mu: ndarray,均值
    :param sigma2: ndarray,方差
    :return: None
    """
    x = np.linspace(5, 25, 100)
    y = np.linspace(5, 25, 100)
    xx, yy = np.meshgrid(x, y)
    X = np.concatenate((xx.reshape(-1, 1), yy.reshape(-1, 1)), axis=1)
    z = gaussian_distribution(X, mu, sigma2).reshape(xx.shape)
    cont_levels = [10 ** h for h in range(-20, 0, 3)]  # 当z为当前列表的
    plt.contour(xx, yy, z, cont_levels)
```

```python
    mu, sigma2 = estimate_parameters_for_gaussian_distribution(X)
    p = gaussian_distribution(X, mu, sigma2)
    visualize_contours(mu, sigma2)
```

3 液体火箭发动机异常检测案例

异常检测程序流程

模型计算出的结果，进行误差分析，计算F1-score

```python
def error_analysis(yp, yt):
    """
    计算误差分析值F1-score
    :param yp: ndarray,预测值
    :param yt: ndarray,实际值
    :return: float,F1-score
    """
    tp, fp, fn, tn = 0, 0, 0, 0
    for i in range(len(yp)):
        if yp[i] == yt[i]:
            if yp[i] == 1:
                tp += 1
            else:
                tn += 1
        else:
            if yp[i] == 1:
                fp += 1
            else:
                fn += 1
    precision = tp / (tp + fp) if tp + fp else 0  # 防止除以0
    recall = tp / (tp + fn) if tp + fn else 0
    f1 = 2 * precision * recall / (precision + recall) if precision + recall else 0
    return f1
```

3 液体火箭发动机异常检测案例

异常检测程序流程

封装阈值选择函数，阈值从预测值的最小到最大的范围中遍历，计算F1-score选择出最好的阈值选择

```python
def select_threshold(yval, pval):
    """
    根据预测值和真实值确定最好的阈值
    :param yval: ndarray,真实值（这里是0或1）
    :param pval: ndarray,预测值（这里是[0,1]的概率）
    :return: (float,float),阈值和F1-score
    """
    epsilons = np.linspace(min(pval), max(pval), 1000)
    l = np.zeros((1, 2))
    for e in epsilons:
        ypre = (pval < e).astype(float)
        f1 = error_analysis(ypre, yval)
        l = np.concatenate((l, np.array([[e, f1]])), axis=0)
    index = np.argmax(l[..., 1])
    return l[index, 0], l[index, 1]
```

测试

```python
Xval = data['Xval']  # (307,2)
yval = data['yval']  # (307,1)
e, f1 = select_threshold(yval.ravel(), gaussian_distribution(Xval, mu, sigma2))
print('best choice of epsilon is ', e, ',the F1 score is ', f1)
# best choice of epsilon is  8.999852631901393e-05 ,the F1 score is  0.875000000000
```

异常检测程序流程

利用选择出来的阈值完善模型，对异常进行预测

```
def detection(X, e, mu, sigma2):
    """
    根据高斯模型检测出异常数据
    :param X: ndarray,需要检查的数据
    :param e: float,阈值
    :param mu: ndarray,均值
    :param sigma2: ndarray,方差
    :return: ndarray,异常数据
    """
    p = gaussian_distribution(X, mu, sigma2)
    anomaly_points = np.array([X[i] for i in range(len(p)) if p[i] < e])
    return anomaly_points
```

将异常的数据点圈出来

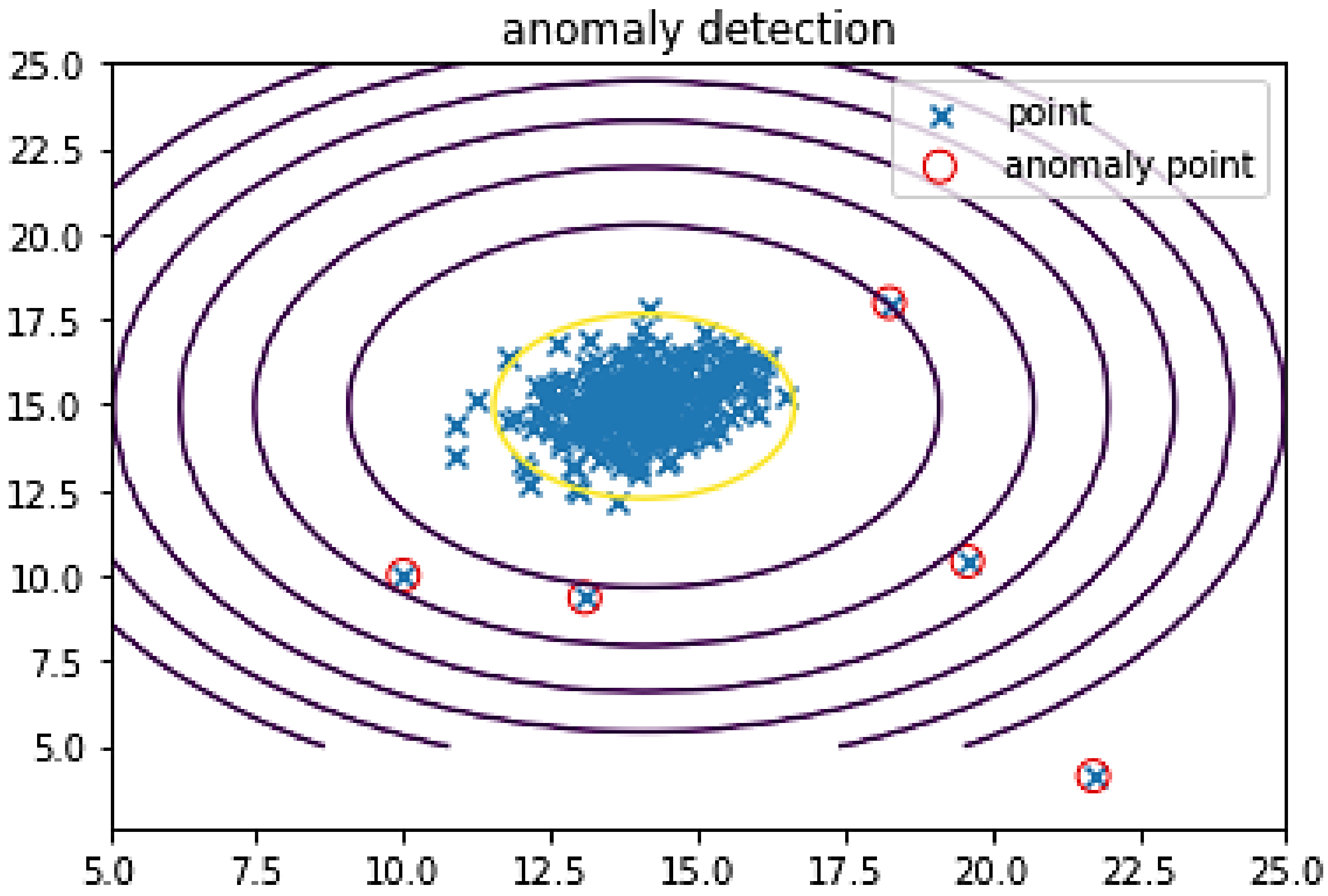
```
def visualize_dataset(X):
    plt.scatter(X[..., 0], X[..., 1], marker='x', label='point')

def circle_anomaly_points(X):
    plt.scatter(X[..., 0], X[..., 1], s=80, facecolors='none', edgecolors='r',
```

```
    anomaly_points = detection(X, e, mu, sigma2)
    circle_anomaly_points(anomaly_points)
    plt.title('anomaly detection')
    plt.legend()
    plt.show()
```

3 液体火箭发动机异常检测案例

异常检测程序流程



液体火箭发动机异常检测案例结果示意图

目录

总结

- 异常检测
  - 异常检测背景
  - 异常检测理论
    - 异常检测概念
      - 半监督学习方法
    - 异常检测方法
      - 基于模型
      - 基于邻近度
      - 基于密度
      - 基于聚类
    - 异常检测系统
      - 基于高斯分布的异常检测算法
    - 异常检测步骤
      - 数据准备
      - 数据分组
      - 异常评估
      - 异常输出
    - 异常检测与监督学习对比
    - 异常检测特征的选择
  - 异常检测知识扩展
    - 多元高斯分布

作业

（1）简要描述航空发动机和火箭发动机的区别，并对火箭发动机进行分类描述。

（2）简要描述异常检测算法原理。

（3）参照12.3节算法流程图以及关键代码，运行液体火箭发动机异常检测程序。

自行修改“ex8data1.mat”数据文件中某个异常点的位置或新增随机点，重新运行程序，输出最佳阈值 ε和其对应的$F1$值，并在图像中标出异常点。

https://blog.csdn.net/weixin_44027820/article/details/104616231?utm_medium=distribute.pc_relevant.none-task-blog-BlogCommendFromMachineLearnPai2-1.channel_param&depth_1-utm_source=distribute.pc_relevant.none-task-blog-BlogCommendFromMachineLearnPai2-1.channel_param

请将第1、2简述题回答，第3题程序代码和程序的运行结果一并整理成Word文档并上交，将该文件于限定时间内发送至指定邮箱sxb762@163.com。

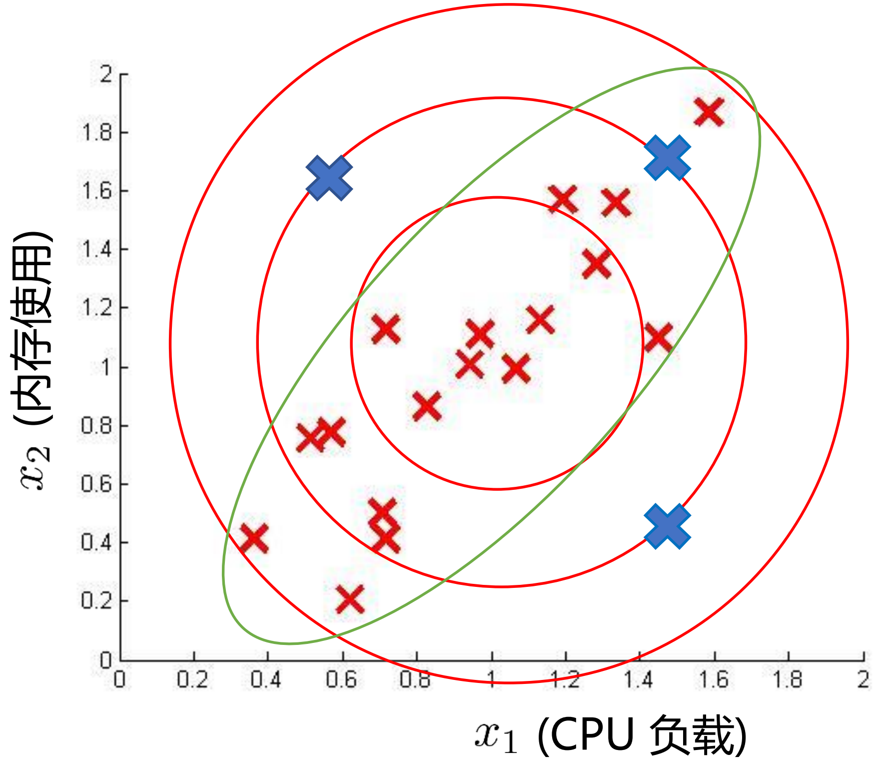
作业要求

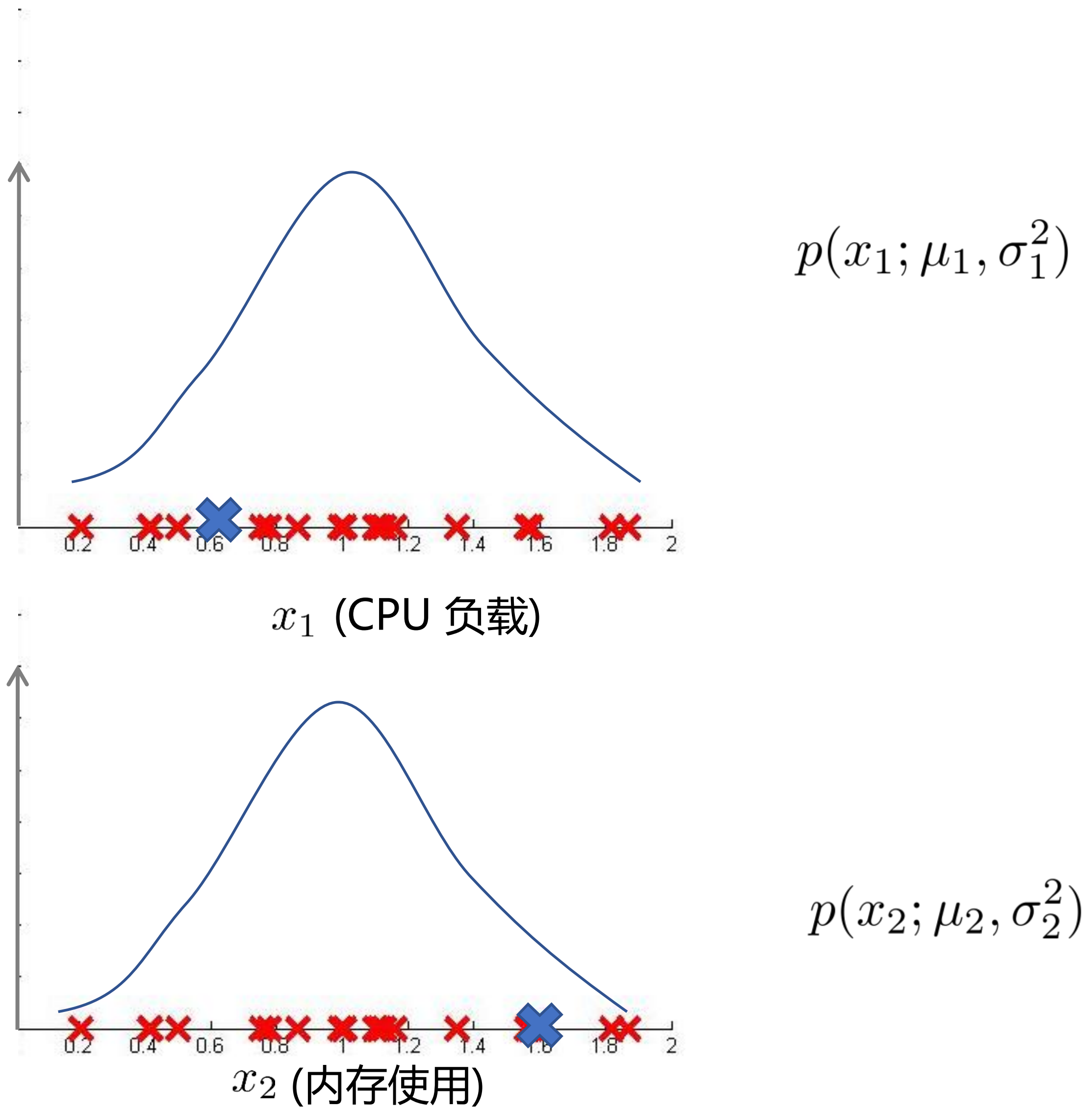
- 学号：#####　　姓名：### 班级：#####
- 题目与要求
- 运行结果
- 过程要点记录、分析与体会等。

目录

多元高斯分布

数据中心的计算机检测





多元高斯分布

多元高斯（正态）分布

对于训练集：$x \in \mathbb{R}^n$

$$p(x_1), p(x_2), \ldots,$$

$$p(x)$$

所用参数为：所有特征的平均值与协方差矩阵

$$\mu \in \mathbb{R}^n, \Sigma \in \mathbb{R}^{n \times n}$$

$$\Sigma = \frac{1}{m}\sum_{i=1}^{m}(x^{(i)} - \mu)(x^{(i)} - \mu)^{\mathrm{T}}$$

$$p(x; \mu, \Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}}\exp\left(-\frac{1}{2}(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu)\right)$$

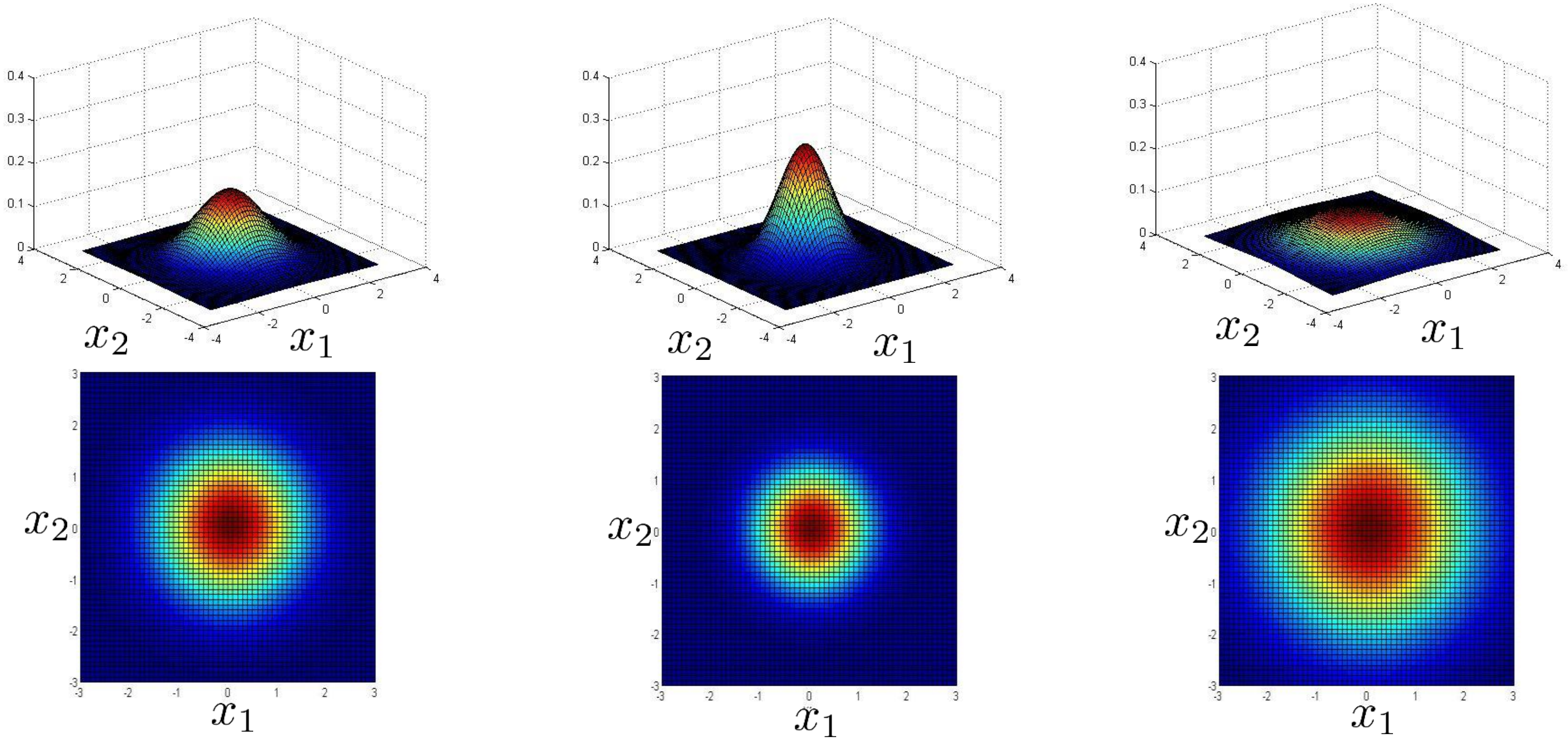
$|\Sigma|$ 矩阵的行列式

5 知识扩展

多元高斯分布

多元高斯（正态）分布参数影响

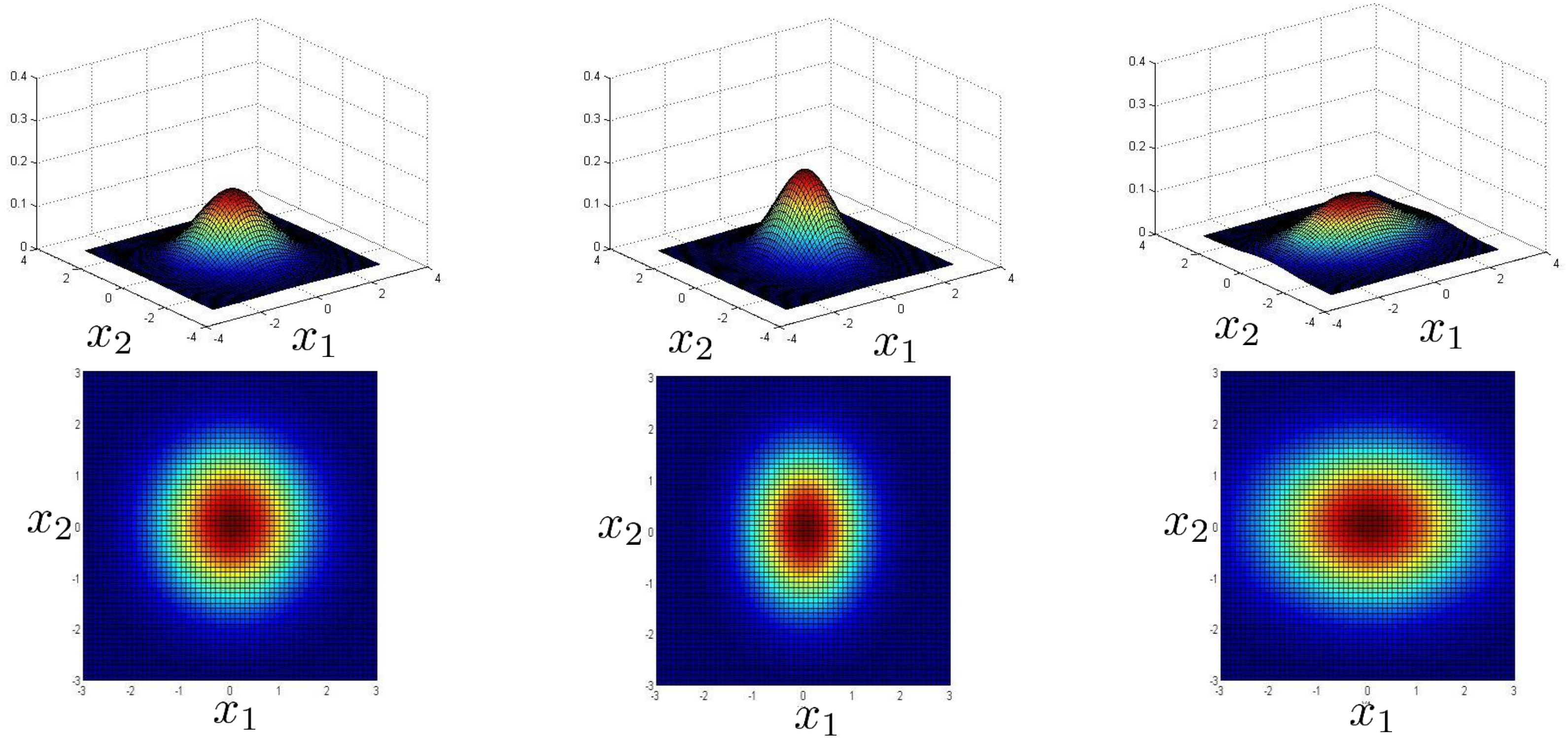
$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad \mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 0.6 & 0 \\ 0 & 0.6 \end{bmatrix} \qquad \mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$



多元高斯分布

多元高斯（正态）分布参数影响

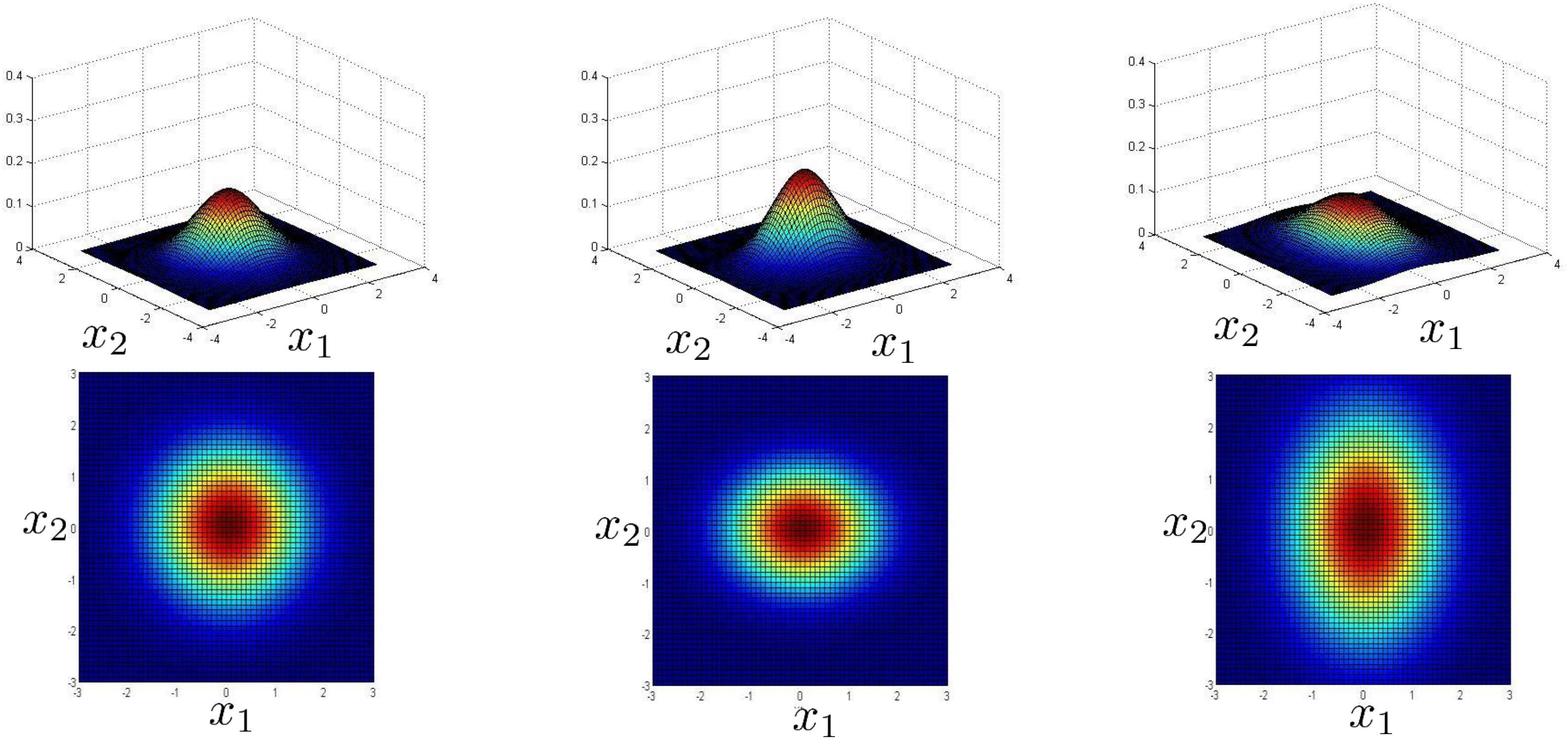
$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad \mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 0.6 & 0 \\ 0 & 1 \end{bmatrix} \qquad \mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$$



多元高斯分布

多元高斯（正态）分布参数影响

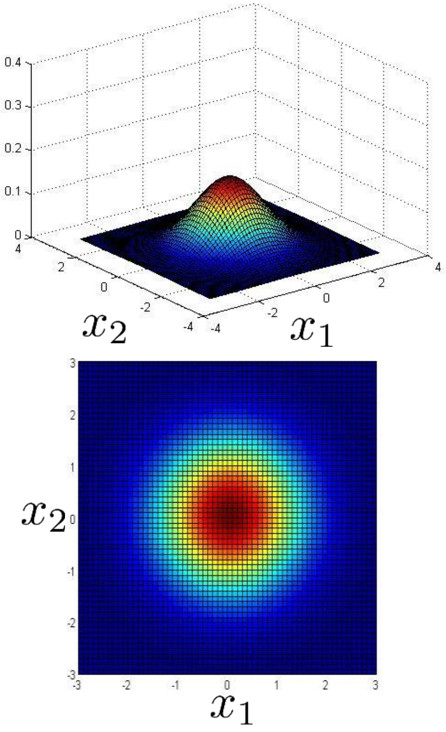
$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ $\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 0.6 \end{bmatrix}$ $\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$



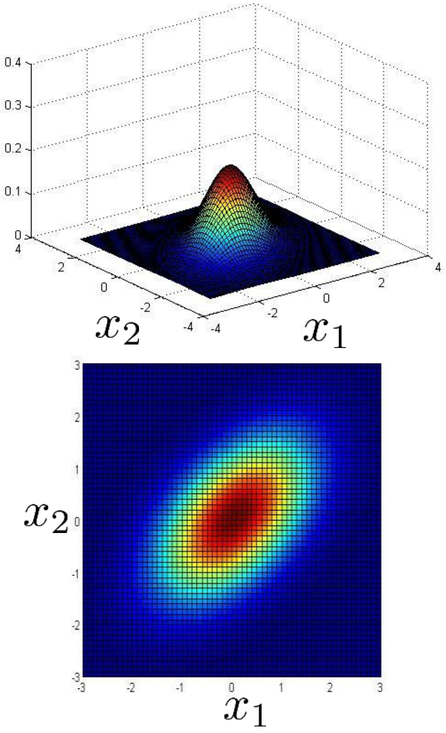
多元高斯分布

多元高斯（正态）分布参数影响

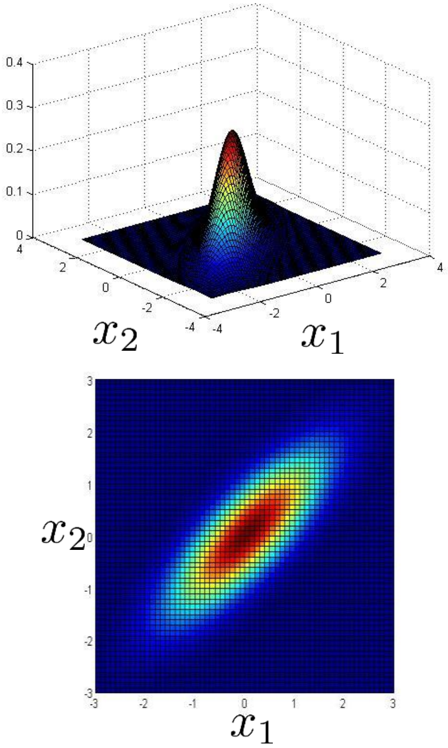
$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$



$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$



$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}$$



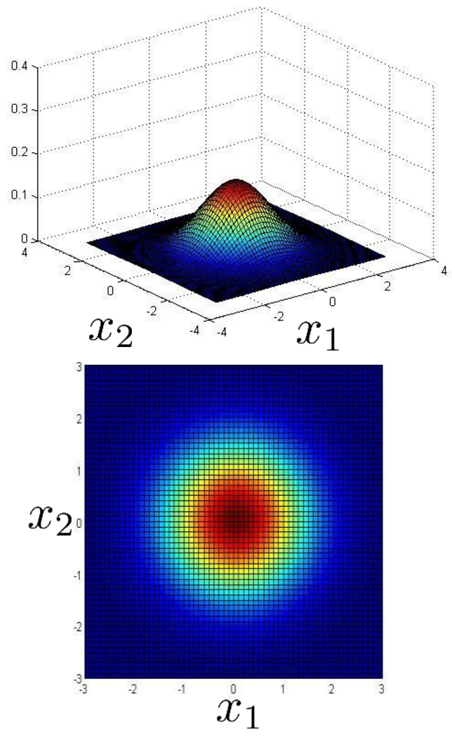
多元高斯分布

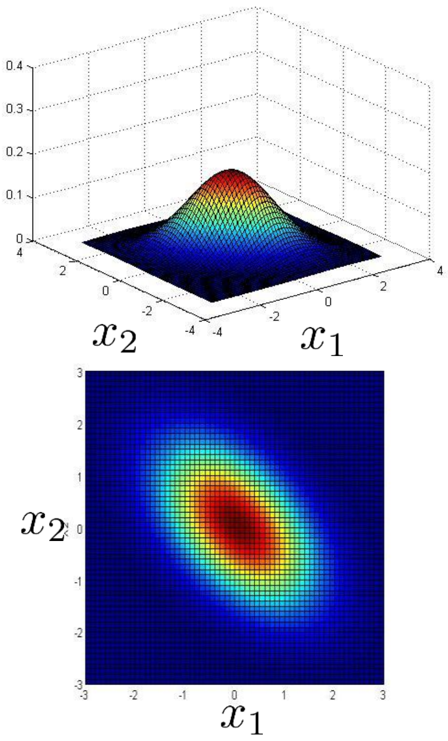
多元高斯（正态）分布参数影响

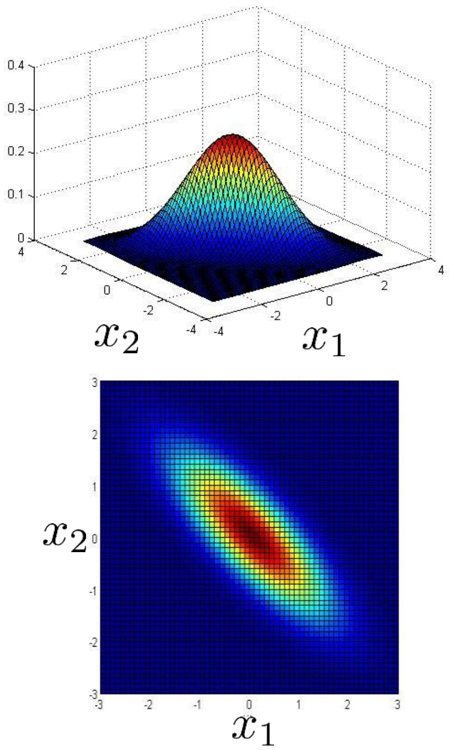
$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$$

$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & -0.8 \\ -0.8 & 1 \end{bmatrix}$$



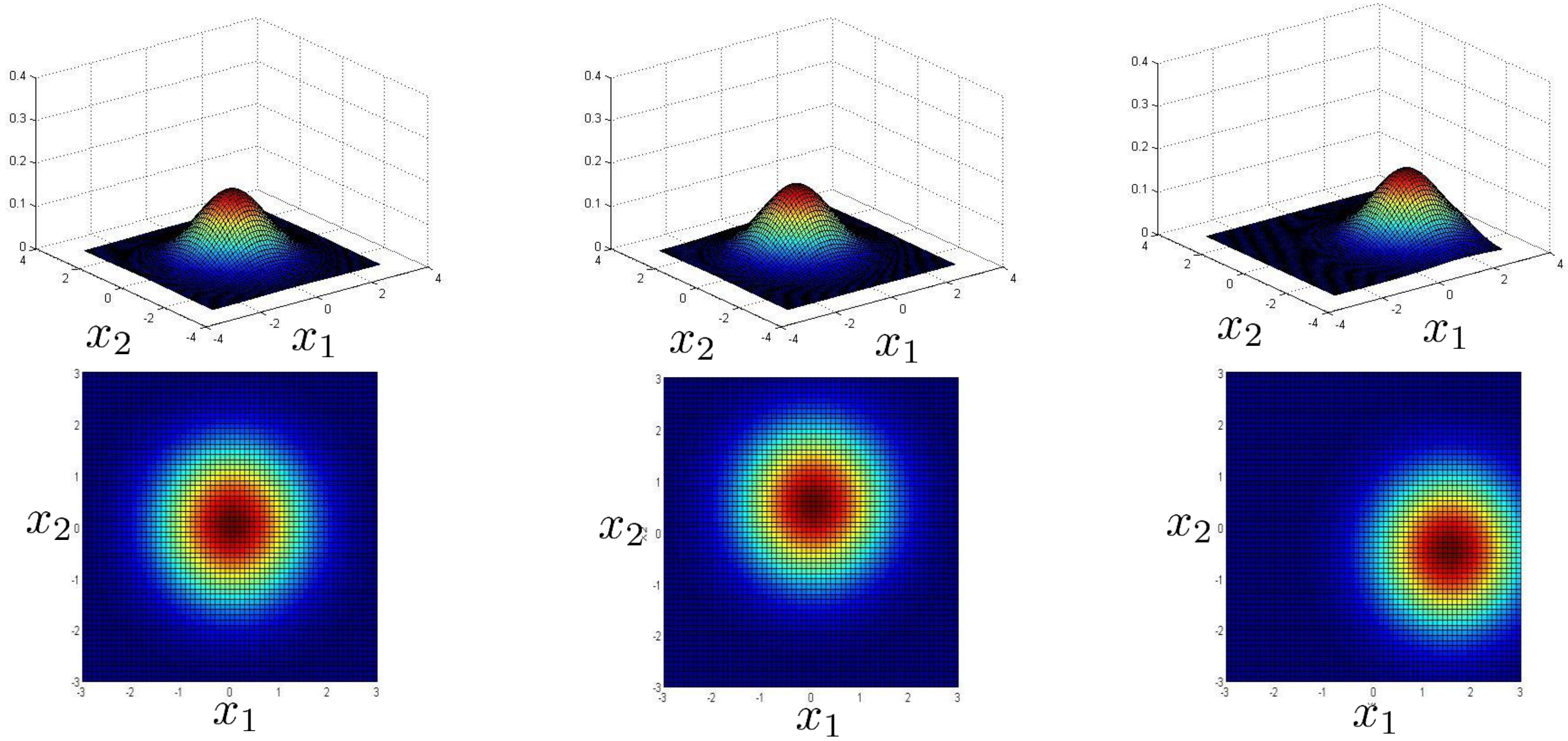




多元高斯分布

多元高斯（正态）分布参数影响

$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad \mu = \begin{bmatrix} 0 \\ 0.5 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad \mu = \begin{bmatrix} 1.5 \\ -0.5 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

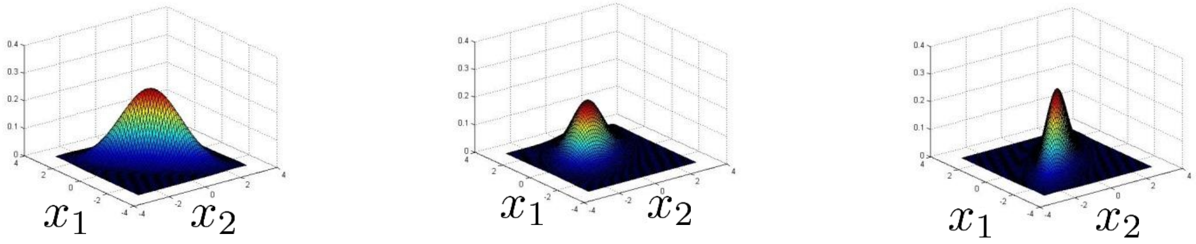


使用多元高斯分布进行异常检测

多元高斯（正态）分布

高斯参数：$\mu, \Sigma \qquad \mu \in \mathbb{R}^n, \Sigma \in \mathbb{R}^{n\times n}$

$$p(x;\mu,\Sigma) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$



对于训练集：$\{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\} \qquad x \in \mathbb{R}^n$

进行参数估算：

$$\mu = \frac{1}{m}\sum_{i=1}^{m} x^{(i)} \qquad \Sigma = \frac{1}{m}\sum_{i=1}^{m}(x^{(i)}-\mu)(x^{(i)}-\mu)^T$$

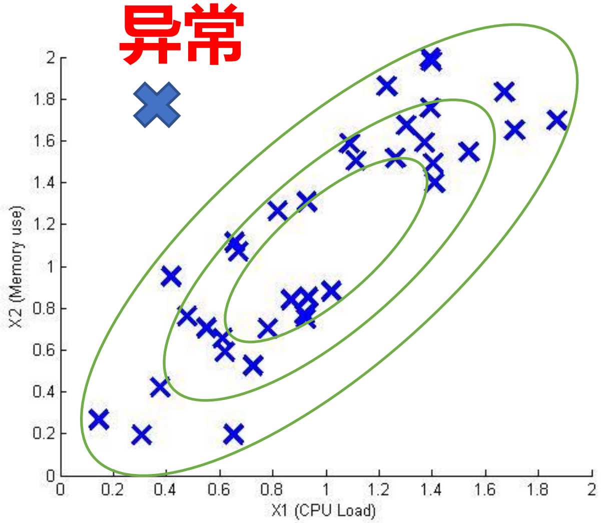
使用多元高斯分布进行异常检测

多元高斯分布异常检测算法

1. 通过参数估算建立模型： $p(x)$

$$\mu = \frac{1}{m}\sum_{i=1}^{m} x^{(i)}$$

$$\Sigma = \frac{1}{m}\sum_{i=1}^{m}(x^{(i)} - \mu)(x^{(i)} - \mu)^T$$



2. 对于测试样本 x，计算：

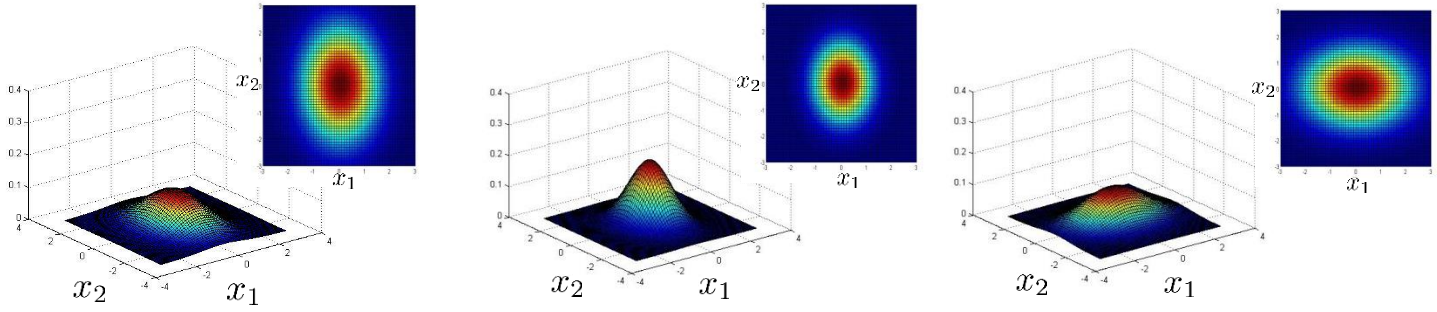
$$p(x) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$

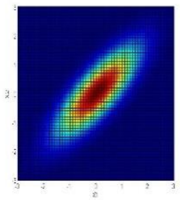
当 $p(x) < \varepsilon$ 判断为异常。

使用多元高斯分布进行异常检测

与原高斯分布模型的关系

原模型: $$p(x) = p(x_1; \mu_1, \sigma_1^2) \times p(x_2; \mu_2, \sigma_2^2) \times \cdots \times p(x_n; \mu_n, \sigma_n^2)$$





$$\Sigma = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}$$

对应多元高斯模型：

$$p(x; \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right)$$

其中:

$$\Sigma = \begin{bmatrix} \sigma_1^2 & & \\ & \sigma_2^2 & \\ & & \ddots & \\ & & & \sigma_n^2 \end{bmatrix}$$

使用多元高斯分布进行异常检测

原高斯分布模型 vs. 多元高斯分布模型

原高斯分布模型

$$p(x_1;\mu_1,\sigma_1^2)\times\cdots\times p(x_n;\mu_n,\sigma_n^2)$$

- 不能捕捉特征之间的相关性 但可以通过将特征进行组合的方法来解决

$$x_5=\frac{x_3}{x_4}=\frac{\text{CPU负载}}{\text{网络通讯量}}$$

计算代价低，适应大规模的特征

$n=10,000 \quad n=100,000$

- 当样本数m很小时，也能计算

多元高斯分布模型

$$p(x;\mu,\Sigma)=\frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$

- 自动捕捉特征之间的相关性

$\Sigma\in\mathbb{R}^{n\times n} \quad \Sigma^{-1}$

计算代价较高

- 必须要有 $m>n$，不然的话协方差矩阵Σ 不可逆

通常需要 $m>10n$

特征冗余也会导致协方差矩阵不可逆

谢谢