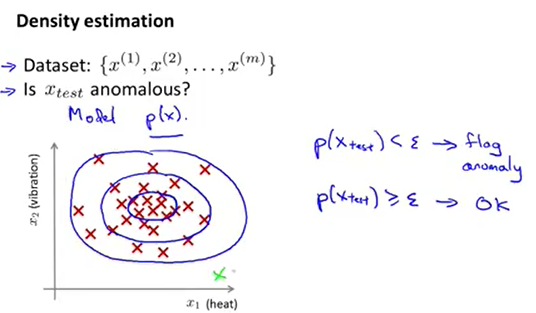
# 异常检测

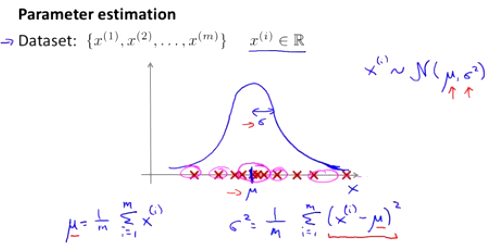
## 异常检测步骤



如上图，对训练数据集构建概率模型P(x)，若概率大于等于阈值则为正常数据，反之则为异常数据。

## 高斯分布

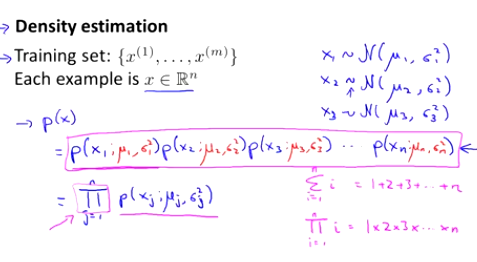
参数估计：



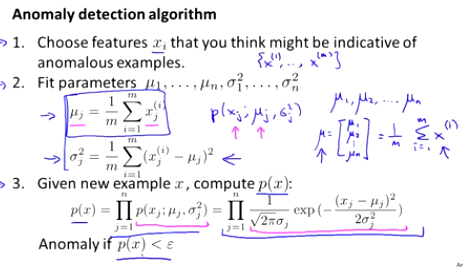
## 算法

高斯分布开发异常检测算法，从训练数据集构造概率模型P(x)，计算出哪些特征出现的概率比较高，哪些特征的概率比较低。因此，x是一个向量，概率模型p(x)等于各特征概率的乘积。

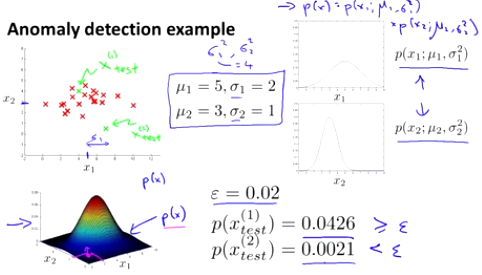
密度估计：



异常检测算法：

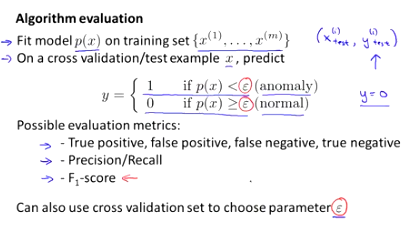


异常检测例子：



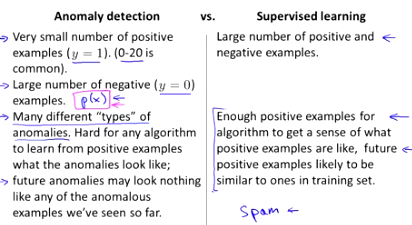
## 开发和评价异常检测系统

选择一个实数来评价算法的好坏具有很重要的作用，（1）假设训练集是无异常的数据集，用训练集拟合出高斯概率模型p(x)；（2）模型对交叉验证集和测试集进行概率预测，交叉验证集和测试集有已知的标签；（3）用评价准则来选择阈值；



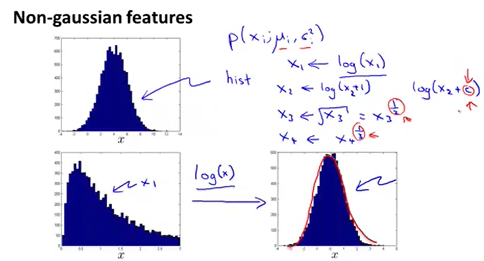
## 异常检测和监督学习

由于上一节当中介绍的异常检测系统有给出标记的交叉验证集和测试集，那么为什么不直接用监督学习的方法来构建模型，什么时候用异常检测，什么时候用监督学习，如下图所示。



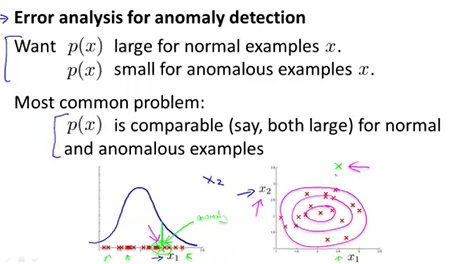
## 特征变量的选择

对于非高斯分布的特征，可以通过非线性方法转换为高斯分布。

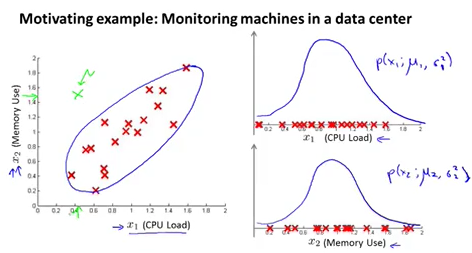


异常检测的误差分析：

如下图，异常检测遇到最大的问题是正常样本和异常样本具有可比的概率模型P(x)。异常样本特征x1的概率较大，若仅仅依靠x1特征来进行异常检测，会出现错误的结果，因此，通过增加新的特征变量x2来构建新的模型。

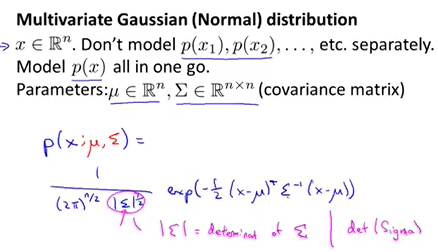


## 多元变量高斯分布



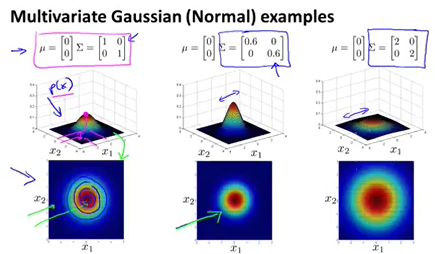
如上图的绿色测试点，若用单元变量分布则会得到是正常的数据，若用多元变量分布则会得到异常数据，原因在于两个变量具有相关性。

多元变量高斯分布：

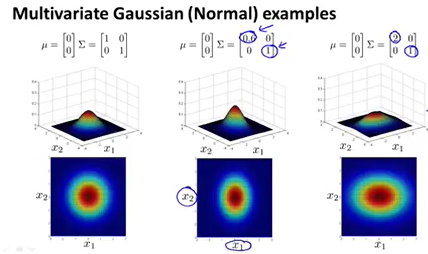


多元变量高斯分布图：

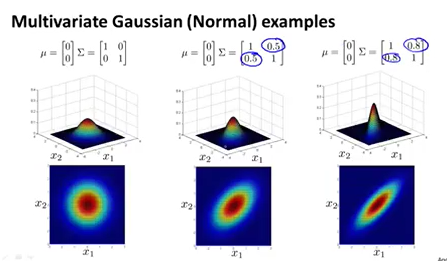
变量相同方差，且无相关性：



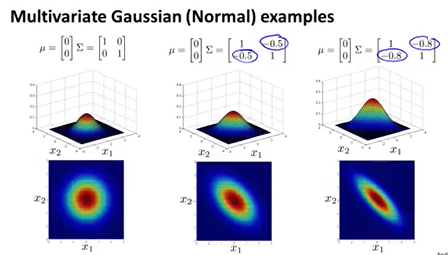
变量具有不相同方差，且无相关性：



多元变量具有正相关性：

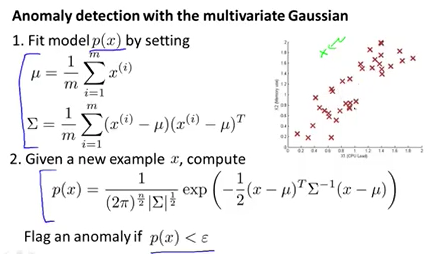


多元变量具有负相关性：

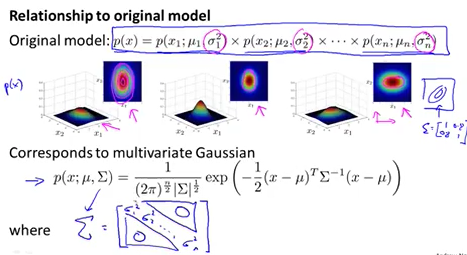


## 多元变量高斯分布的异常检测

多元变量高斯分布的异常检测步骤：



当多元变量相互独立时，多元变量的概率模型等于每个变量概率模型的乘积，多元变量的等高线轴是沿着轴向的高斯分布。



原始模型和多元变量高斯模型的区别：

1. 原始模型检测需要结合之前的特征来手动构建新的特征变量，多元变量高斯模型则自动获取变量间的相关关系。
2. 原始模型即使在特征变量很大的情况下运行速度也很快，多元高斯模型则需要花费较大的时间。
3. 原始模型对样本数无要求，多元变量高斯模型需要样本数大于特征数，否则协方差矩阵是奇异矩阵或不可逆矩阵。

