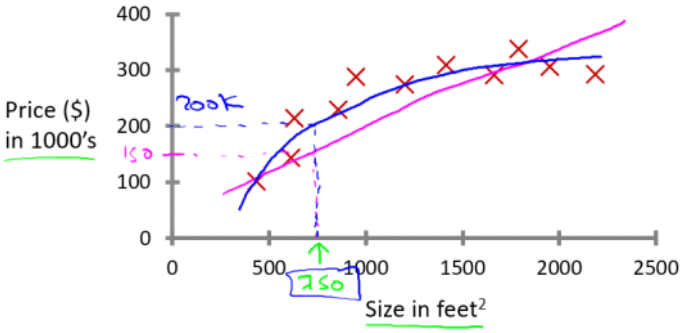
# 1 监督学习与无监督学习

2022年10月14日　　20:27

## 1.1监督学习

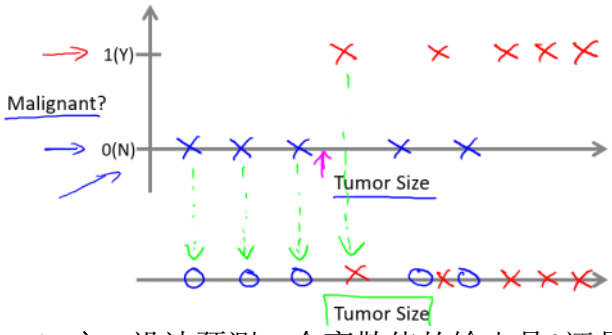
**监督学习**(Supervised Learning)：给出数据集，其中包含正确的答案。算法的目的是为了给出更多正确的答案。

一个关于房价预测的例子：



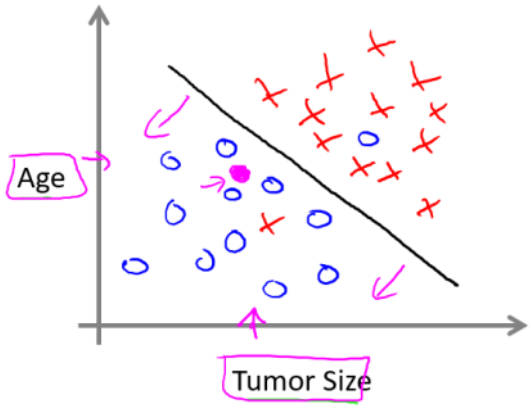

**回归问题**(Regression)：设法预测连续值的属性。

一个关于预测乳腺癌恶性还是良性的例子：




**分类问题**(Classification)：设法预测一个离散值的输出是0还是1(更多输出值)。

将上述肿瘤预测模型扩展，考虑多特征情况：




支持向量机算法是一个处理无穷多特征的分类模型的例子。

一个问题判断，应当将两个problem分别归为分类问题还是回归问题？

Problem 1: You have a large inventory of identical items. You want to predict how many of these items will sell over the next 3 months.
Problem 2: You'd like software to examine individual customer accounts, and for each account decide if it has been hacked/compromised. 0 – not hacked, 1 – hacked

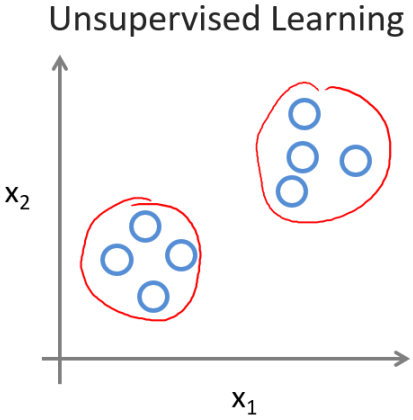
Should you treat these as classification or as regression problems?

- ○ Treat both as classification problems.
- ○ Treat problem 1 as a classification problem, problem 2 as a regression problem.
- → ○ Treat problem 1 as a regression problem, problem 2 as a classification problem.
- ○ Treat both as regression problems.

答案为c。

## 1.2无监督学习

**无监督学习：**只有一个数据集，需要找到其中的某种结构。

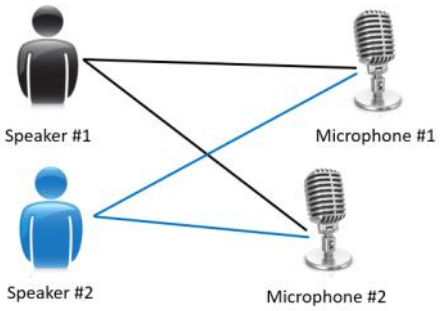



**聚类算法：**无监督学习判断数据集中包含不同的簇。

应用聚类算法的例子：
1.谷歌新闻：将成千上万条新闻进行分簇，有关同一主题的新闻被分到一起；
2.DNA微阵列数据：检测不同个体是否含有特定基因，运行聚类算法可以将个体归入不同的类或不同类型的人。

无监督学习的其他应用：
1.组织大型计算机集群：寻找哪些机器协同工作从而将其放在一起保证高效的运算工作；
2.社交网络分析：识别和我联系密切的人的社交圈，判断“可能认识的人”；
3.市场细分：对大量客户进行市场分割并分配到不同的细分市场，有针对地销售；
4.天文数据分析：帮助与分析星系形成理论。

一个例子，鸡尾酒会问题：




算法可以识别并分离叠加到一起的音频。
Octave中一句函数即可实现：

```
[W,s,v] = svd((repmat(sum(x.*x,1),size(x,1),1).*x)*x');
```

一个问题：下面四种情况哪种是无监督学习算法？

- ☐ Given email labeled as spam/not spam, learn a spam filter.
- ☐ Given a set of news articles found on the web, group them into set of articles about the same story.
- ☐ Given a database of customer data, automatically discover market segments and group customers into different market segments.
- ☐ Given a dataset of patients diagnosed as either having diabetes or not, learn to classify new patients as having diabetes or not.

答案为B,C。