1 Introduction引言

1.1 Welcome

第一个视频主要讲了什么是机器学习，机器学习能做什么。

- Grew out of work in AI
- New capability for computers

机器学习案例：

- Database mining 数据库挖掘
  Large datasets from growth of automation/web.
  E.g.,Web click data, medical records, biology, engineering
- Applications can't program by hand.
  E.g.,Autonomous helicopter,(直升机自主飞行程序) handwriting recognition, most of Natural Language Processing(NLP), Computer Vision
- Self-customizing programs
  E.g., Amazon, Netflix product recommendations

1.2 What is Machine Learning?什么是机器学习

第一个机器学习的定义来自于 Arthur Samuel。他定义机器学习为，在不进行特定编程的情况下，给予计算机学习能力的领域。

Arthur Samuel described it as: "the field of study that gives computers the ability to learn without being explicitly programmed." This is an older, informal definition.

另一个年代比较近的定义，Tom Mitchell定义的机器学习是，一个好的学习问题，定义如下，他说，一个程序被认为能从经验E中学习，解决任务T，达到性能度量值 P，当且仅当，有了经验E后，经过P评判，程序在处理T时的性能有所提升。

Tom Mitchell provides a more modern definition: "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.

Example: playing checkers.
E = the experience of playing many games of checkers
P = the probability that the program will win the next game.
In general, any machine learning problem can be assigned to one of two broad classifications: Supervised learning and Unsupervised learning.

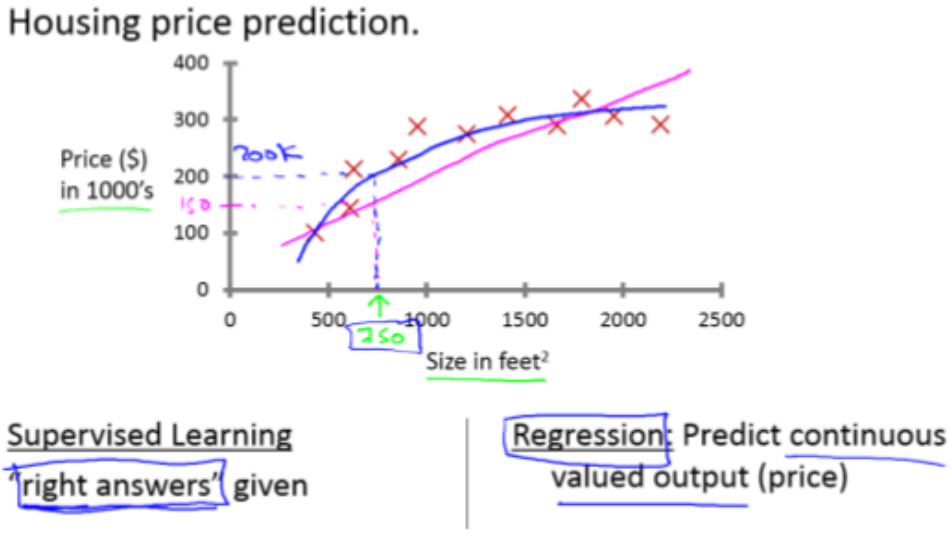
目前存在几种不同类型的学习算法。主要的两种类型被我们称之为监督学习（supervised learning）和无监督学习（unsupervised learning）。

其他的有，强化学习（Reinforcement learning），推荐系统（recommender systems）。也会提到应用学习算法的实用建议（Practical advice for applying learning algorithms）。

1.3 Intorduction Supervised Learning 监督学习

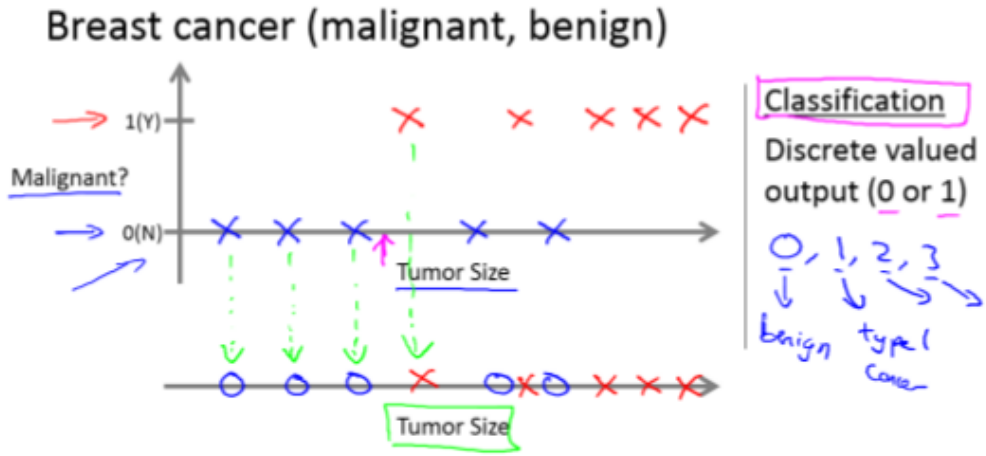
Supervised Learning "right answers"given.监督学习指的就是我们给学习算法一个数据集，这个数据集由"正确答案"组成。

例1：售楼的案例的数据是实现准确得到的，一个学生从波特兰俄勒冈州的研究所收集了一些房价的数据。你把这些数据画出来,看起来是这个样子:横轴表示房子的面积,单位是平方英尺,纵轴表示房价,单位是千美元。那基于这组数据,假如你有一个朋友,他有一套 750 平方英尺房子,现在他希望把房子卖掉,他想知道这房子能卖多少钱。



用术语来讲，这叫做回归问题。回归(Regression)这个词的意思是，我们在试着推测出这一系列连续值属性(Predict continuous valued outputR(price))。

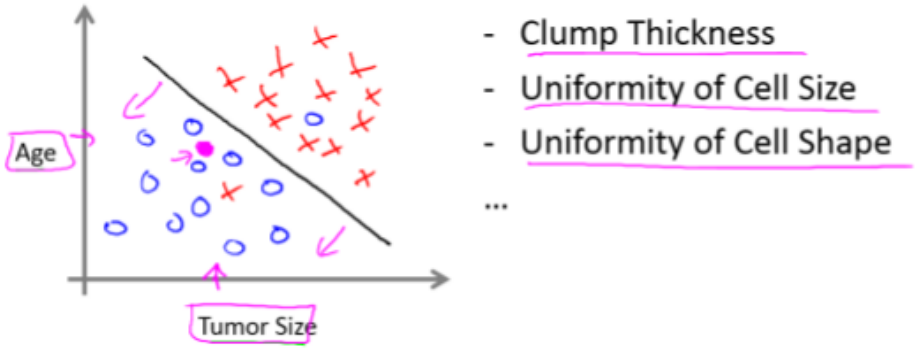
例2：肿瘤样本的例子属于分类问题，其目标是推出一组离散结果。实际的分类问题中输出会不止两个值。预测肿瘤的恶性与否用到的特征，在机器学习问题中，可能会遇到不止一种特征。



现在我用不同的符号来表示这些数据。既然我们把肿瘤的尺寸看做区分恶性或良性的特征,那么我可以这么画,我用不同的符号来表示良性和恶性肿瘤。或者说是负样本和正样本 现在我们不全部画 X,良性的肿瘤改成用 O 表示,恶性的继续用 X 表示。来预测肿瘤的恶性与否。

如果想用无限多种特征好让你的算法可以利用大量的特征， 或者说线索来做推测。如何处理怎么存出这些特征都存在问题，电脑内存会不够。用支持向量机来解决，可以让计算计处理无限多个特征。

在其它一些机器学习问题中,可能会遇到不止一种特征。举个例子,我们不仅知道肿瘤的尺寸,还知道对应患者的年龄。在其他机器学习问题中,我们通常有更多的特征,我朋友研究这个问题时,通常采用这些特征,比如肿块密度,肿瘤细胞尺寸的一致性和形状的一致性等等,还有一些其他的特征。这就是我们即将学到最有趣的学习算法之一。那种算法不仅能处理 2 种 3 种或 5 种特征,即使有无限多种特征都可以处理。



我列举了总共 5 种不同的特征,坐标轴上的两种和右边的 3 种,但是在一些学习问题中,你希望不只用 3 种或 5 种特征。相反,你想用无限多种特征,好让你的算法可以 利用大量的特征,或者说线索来做推测。那你怎么处理无限多个特征,甚至怎么存储这些特征都存在问题,你电脑的内存肯定不够用。我们以后会讲一个算法,叫支持向量机,里面有一个巧妙的数学技巧,能让计算机处理无限多个特征。

监督学习。其基本思想是,我们数据集中的每个样本都有相应的“正确答案”。再根据这些样本作出预测,就像房子和肿瘤的例子中做的那样。我们还介绍了回归问题,即通过回归来推出一个连续的输出,之后我们介绍了分类问题,其目标是推出一组离散的结果。

练习:

You're running a company, and you want to develop learning algorithms to address each of two problems. 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. 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.

正确

1.4 Intorduction Unsupervised Learning 无监督学习

无监督学习中的数据没有任何标签或者是有相同的标签或者就是没标签，没有基于预测结果的反馈。但无监督学习算法可能会把这些数据分成不同的簇，叫做聚类算法(clustering)。

无监督学习或者聚类算法在其他领域有着大量应用：组织大型的计算机群(organize computing clusters)、社交网络分析(social network analysis)、 市场分割中的应用(market segmentation)、天文数据分析(astronomical data analysis)。这些都是聚类的例子，聚类只是无监督学习的一种。

例1：基因问题，找到一种方法，将这些基因自动组合成类似或相关的不同变量，如寿命、位置、角色等。

例2：非聚类问题，鸡尾酒宴问题，允许你在混乱的环境中找到结构。就是从重叠的音频中分离音频。 代码：[W, s, v] = svd ((repmat(sum(x.∗x,1),size(x,1),1).∗x)∗x');