

第一周笔记

1.1 引言

- 机器学习在日常生活中得到了广泛的应用。
- 机器学习也逐步在工业领域发挥作用。

1.2 机器学习的应用

- 传统：遇到问题-->分析问题-->编写算法-->运行解决问题
- 机器学习：遇到问题-->搜集或创造足够多的解决样例-->交由机器学习-->机器得以用于解决问题。

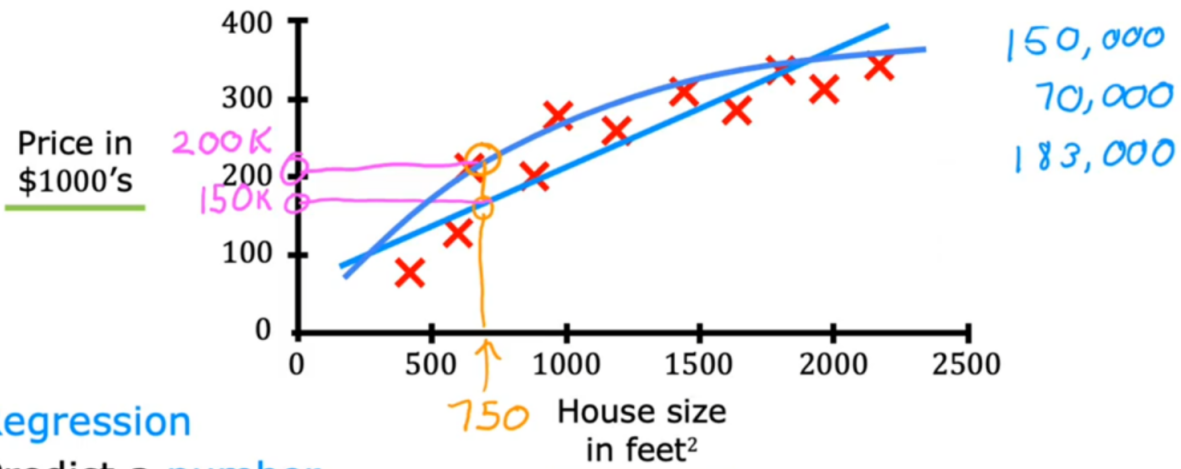
2.1 机器学习是什么

- 塞缪尔的定义：“给予电脑不需要明确编程的自学习能力的研究领域”。
- 机器学习的主要分类：
 - 监督学习（使用最多，而且进步速度最快）
 - 无监督学习
 - 强化学习（不在本次课程的范围内）

2.2 监督学习part1

- 监督学习中最为常见的是预测类别的学习,也就是回归问题

Regression: Housing price prediction



Regression

Predict a **number**

infinitely many possible outputs

- 输入x经过函数f的运算得到输出y

$$x \xrightarrow{f} y$$

这个f可以是任意形式的函数，所以无法直接求解通过已知的 x_0, y_0, \dots 等输入输出对，算出一个接近的 f'

$$ML(x_0, y_0, \dots) \rightarrow f'$$

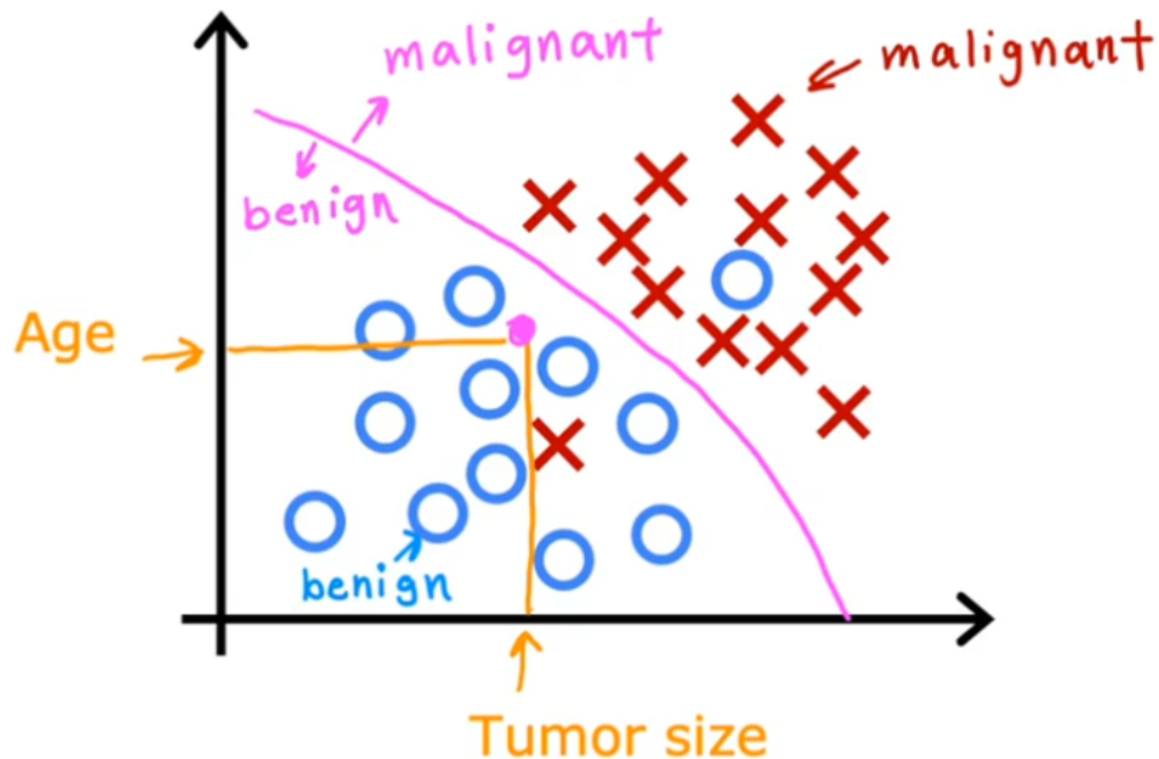
最后通过 f' 来预测新输入 x' 的输入 y'

$$x' \xrightarrow{f'} y''$$

2.3 监督学习part2

- 如果输出是受限的，比如只有有限的选择，那么回归问题会变为分类问题。因此，输出可以不再是数，而是由分类方式决定的。

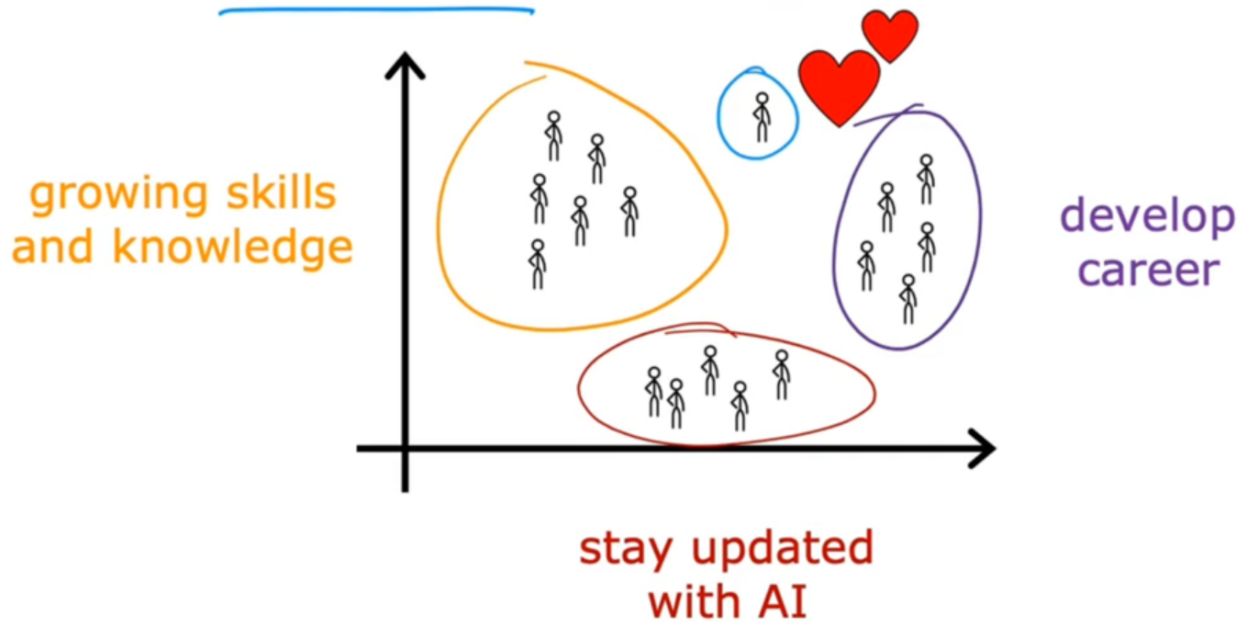
Two or more inputs



2.4 无监督学习part1

- 无监督学习不再有预先准备好的正确数据，而是交由机器去发掘数据中呈现出的特殊性。
- 代表算法为聚类算法（在大量数据中，划分出（给定或不给定数量的）多块，这些块内部更加相关，而且互相之间差异较大）

Clustering: Grouping customers




2.5 无监督学习part2

- 异常检测也是一种无监督学习
- 数据压缩：将一个大的数据集减小到一个较小的数据集

2.6 jupyter入门

本节课主要引入了jupyter notebook的应用。

 [lab1-0.ipynb](#)：尝试最为基本的ipynb文件

3.1 线性回归模型part1

介绍了一些机器学习中的基本结构及术语：

- Training Data 用于进行模型训练的数据集
- feature 特征值，也就是输入的 x （可以是多维）
- target 目标值，也就是输出的 y （可以是多维）
- (x, y) 数据对，代表了一一对应的输入输出
- (x^i, y^i) 第 i 组训练数据

3.2 线性回归模型part2

进一步介绍了一些结构术语

- f 函数，一般指经过训练得到的函数
- $\hat{y} = f(x)$,指预测得到的输出，和真实的 y 之间存在差距

 [lab1-1.ipynb](#): 用python绘制一个简单的线性函数

3.3 损失函数

对于一个最基本的线性模型 f ，最常用的损失函数构造为：

$$f = wx + b$$
$$J(w, b) = \frac{1}{2m} \sum_1^m (\hat{y} - y)^2$$

这里除以一个 m 是为了让损失函数值与训练的数据量无关，再除以一个 2 是为了让之后的求导变的更简洁