

CS229_0

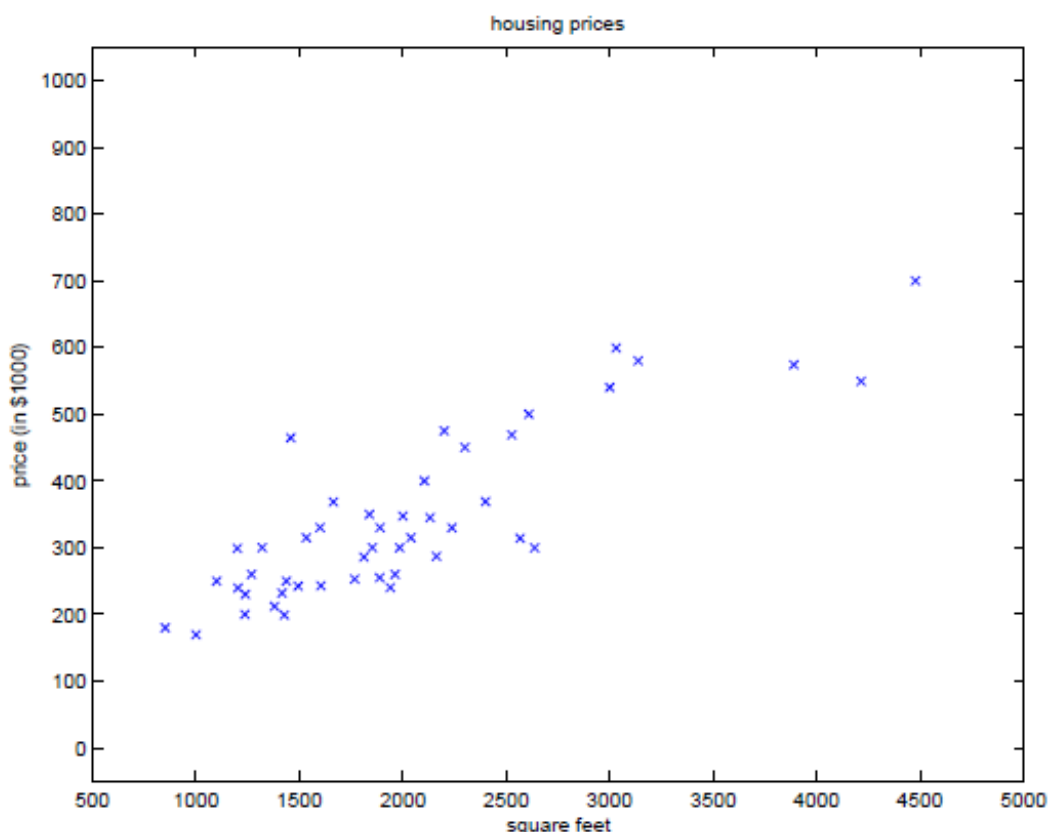
CS229

监督学习 (Supervised Learning)

首先我们通过一些例子来讨论监督学习问题吧。假设我们有如下数据集：

Living area (feet ²)	Price (1000\$)
2104	400
1600	330
2400	369
1416	232
3000	540
⋮	⋮

我们可以将上述数据集绘制成如下图所示：

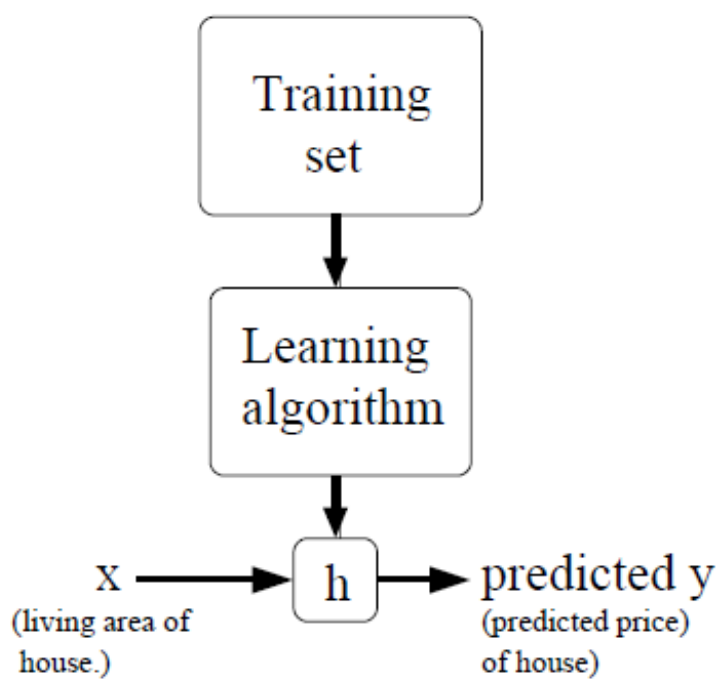


那么我们如何根据这些数据集预测其他房屋的价格呢？

为了今后更好地学习，以上例为例，我们将房屋面积定义为输入变量 $\boldsymbol{x}^{(i)}$ ，其亦称为特征变量；将房屋价格定义为输出变量 $\boldsymbol{y}^{(i)}$ ，其亦称为目标变量； $(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)})$ 表示一个训练实例，其中上标 i 表示第 i 个训练实例；集合 $\{(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}); i = 1, \dots, m\}$ 表示训练集，其中 m 表示训练集的大小。

当然，我们也可使用 $\boldsymbol{\chi}$ 表示输入变量空间， \boldsymbol{Y} 表示输出变量空间。在房屋价格此例中， $\boldsymbol{\chi} = \boldsymbol{Y} = \boldsymbol{R}$ 。

在监督学习问题中，我们对给定的数据集，学习到一个函数 $\boldsymbol{h} : \boldsymbol{\chi} \mapsto \boldsymbol{Y}$ ，该函数 $\boldsymbol{h}(\boldsymbol{x})$ 能够很好地预测目标变量。通常，我们将该函数 $\boldsymbol{h}(\boldsymbol{x})$ 称为假设函数（Hypothesis Function），其处理流程如下图所示：



当输出变量为连续值时，我们将其问题称为回归问题（Regression Problem）；当输出变量为离散值时，我们将其问题称为分类问题（Classification Problem）。