CS229_1

CS22

线性回归

我们在上一节房屋售价数据集的基础上，增添房间数量这一特征变量，如下图所示：

Living area (feet^2)	#bedrooms	Price (1000$s)
2104	3	400
1600	3	330
2400	3	369
1416	2	232
3000	4	540
⋮	⋮	⋮

因此，特征变量x变为了维度为2的向量，记作$x \in R^2$，其中$x_1^{(i)}$表示数据集中第i个房屋的房屋面积，则$x_2^{(i)}$表示数据集中第i个房屋的房间数量。

对于此监督学习问题，若我们采用线性回归模型，其假设函数$h(x)$为：

$$h(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 = \sum_{i=0}^{m} \theta_i x_i = h_\theta(x)$$

其中，$h_\theta(x)$表示以θ为参数。为了便于向量化，我们令$x_0 = 0$，则上式可改写为：

$$h_\theta(x) = \theta^T x$$

从上式可知，θ为未知变量。那么我们该如何根据数据集计算出θ的值呢？我们不妨回想一下假设函数$h_\theta(x)$的定义。从上一小节可知，假设函数$h_\theta(x)$是我们从给定数据集中学习得到的，

其输出的值与数据集中的y越相近越好。因此，我们可以定义如下的代价函数（Cost Function）：

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^i)^2$$

当代价函数$J(\theta)$最小时，其参数θ的值为我们所要的，从而得到了拟合训练集的最佳参数。