C1-1

Machine Learning
by Andrew Ng, Stanford Engineering

Xiaojie Zhou
szxjzhou@163.com
2016.8.6

第一集：机器学习的动机与应用

- 什么是机器学习
- 机器学习的四大部分

第一集：机器学习的动机与应用

- 什么是机器学习
  - Arthur Samuel(1959):Machine Learning: Field of study that gives computers the ability to learn without being explicitly programmed.
    - 例子：像Alpha Go这样的下棋程序可以让计算机自己和自己下棋，通过多个样本计算机最后可以知道怎样的棋局更接近胜利，从而以此引导自己下棋的动作。最终使得计算机的下棋能力超过了人类。在这里并没有直接告知计算机应该如何下，而是让计算机自己学习。
  - Tom M. Mitchell(1998): Well Proposed Learning Problem: 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.
    - 对应上一个例子：E对应着程序不断和自己下棋的经历，T为下棋，P为真实环境下下棋的胜率

第一集：机器学习的动机与应用

- 机器学习的四大部分
  - 监督学习(Supervised Learning)
    - 在监督学习中我们为算法提供了一组正确的数据集（通过“标准答案”监督算法的运行），使得算法通过学习这些正确的数据中输入和输出的关系，以得到未知输入的对应输出
    - 回归(regression)
      - 比如说我们已知某一个地方房屋价格的统计（房屋的面积与房屋价格的关系数据），要预测这个地方在给定某一房屋面积x的情况下对应的房价
      - 对此我们可以将所有的数据样本画出来，然后拿直线或者抛物线或者某一函数曲线去拟合，要让拟合的误差尽可能小（就像高中物理实验那样）
      - 在回归问题中需要衡量的变量是连续的（比如说这里的价格）
    - 分类(classification)
      - 比如说预测某个肿瘤是否为恶性的，我们已知肿瘤的大小和是否为恶性肿瘤的关系数据，要预测给定肿瘤大小的肿瘤是否为恶性
      - 对此我们依然可以画一条线(PLA)，也可以采取其他的划分方法
      - 在分类问题中需要衡量的变量是离散的（比如说这里的是否为恶性肿瘤）

第一集：机器学习的动机与应用

- 机器学习的四大部分
  - 学习理论(Learning Theory)
    - 解决问题：怎样评价一个学习型算法？为什么说这个学习型算法是有效的？需要多少的训练数据？
  - 无监督学习(Unsupervised Learning)
    - 在监督学习中一开始提供了“标准答案”，而在无监督学习中给你一组没有答案的数据（即：没有类标的数据），需要从中发现一些有趣的结构
    - 聚类是一种典型的无监督学习，比如说发现基因的奥秘（对于基因进行聚类）、对图像进行三维重构（对图像中的像素点进行聚类，从而将图片分为不同的区域）、社交网络分析、推荐系统、语音增强去噪声（将声音进行聚类，进行独立成分分析(ICA)）

第一集：机器学习的动机与应用

- 机器学习的四大部分
  - 强化学习(Reinforcement Learning)
    - 在监督学习中我们需要得出一次性的结论（比如马上就要得出肿瘤是否为恶性的推断），但在某些实际情况下并不是这样的，需要做出一系列的决策
      - 比如说控制机器人，不是直接发一个指令就完事的而是要发一系列的指令
      - 在强化学习中有一个回报函数。就像训练狗狗一样，做得好给颗糖，做不好打屁屁。狗狗就会学习到什么样的动作是好的从而拿到更多的糖，什么样的是不好的，从而留下越来越多好的动作
      - 控制机器人也是一样，做得好加分，做不好扣分，久而久之就会有更多的积极回报
      - 这个的问题在于如何定义什么行为是好的行为？然后就是选择合适的算法以获得更大的回报

第二集：监督学习应用与梯度下降

- 线性回归(Linear Regression)
- 梯度下降(Gradient Descent)
- 正规方程(Normal Equations)

第二集：监督学习应用与梯度下降

- 本集案例
  - 利用监督学习实现汽车的自动驾驶
    - 在里面涉及到神经网络、梯度学习
    - 训练方法：由真人控制车辆，车辆会记录下方向盘的动作和车辆的行驶路径（每隔一段时间采集前方道路图像再进行图像分析）间的关系
      - 因为驾驶员提供了在某个路径下车的行驶方向（比如说左弯的路，驾驶员会左打方向盘使车辆朝左行驶），因此称为监督学习

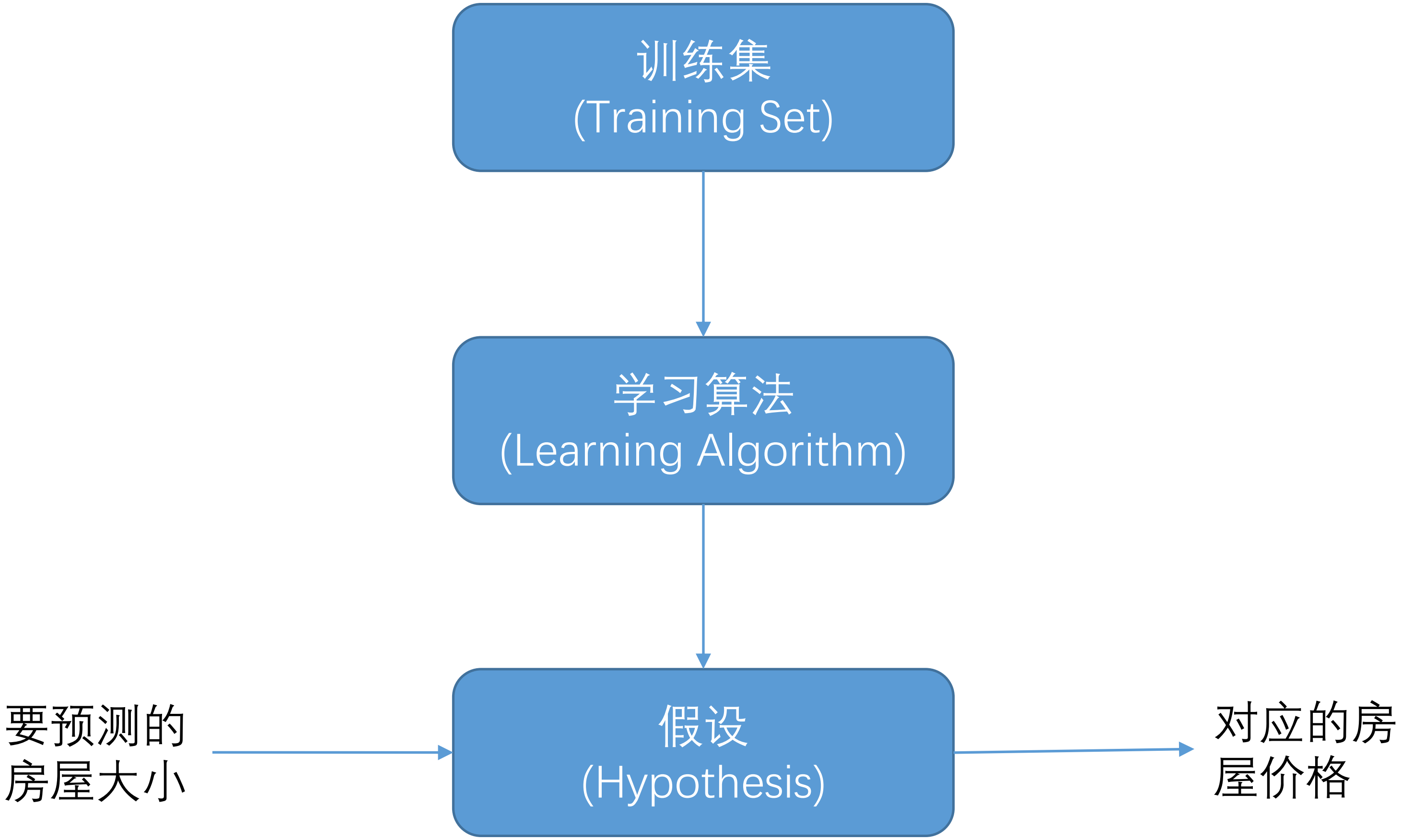
第二集：监督学习应用与梯度下降

- 线性回归
  - 假设下面这组房屋大小和价格的关系样本（暂时先不考虑几房的问题），现在要根据这些样本预测某个大小的房屋对应的价格
  - 符号定义：m表示样本的个数，x表示输入变量/特征（在这里即房屋的大小），y表示输出变量（在这里即要房屋对应的价格），(x,y)表示一组样本），(xi,yi)表示第i组样本，n表示学习问题中特征的数目

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

第二集：监督学习应用与梯度下降

- 线性回归



第二集：监督学习应用与梯度下降

- 线性回归
  - 现在的关键在于如何表示出这样的假设，在线性回归这里将假设用线性函数表达了出来
    - 即h(x) = θ0 + θ1x，其中θ0、θ1为系数，x为输入变量（在这里即房屋的大小）
    - 但是一般的回归问题可能不止一个输入特征，比如将上面问题的房间数考虑进来
      - 这时表达式就变成了h(x) = θ0 + θ1x1 + θ2x2 ，其中θ0、θ1 、θ2为系数， x1表示房屋大小，x2表示房数（在这里h(x) 也可写作hθ (x) ，以说明θ为参数）

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

第二集：监督学习应用与梯度下降

- 线性回归
  - h(x) = θ0 + θ1x1 + θ2x2可以看成x0=1，因此我们也可将其写成

$$h(x)=\sum_{i=0}^{n}\theta_i x_i=\theta^T x$$

其中n表示学习问题中特征的数目，在这里为2

    - 而在这里学习算法的任务在于怎样算出这里的参数θ
    - 由于我们已经拥有了训练集，我们要做的就是使得训练集的输入变量x通过h(x)求出的值尽量接近对应的标准答案y，即要选出使得下式J(θ)取最小值的θ

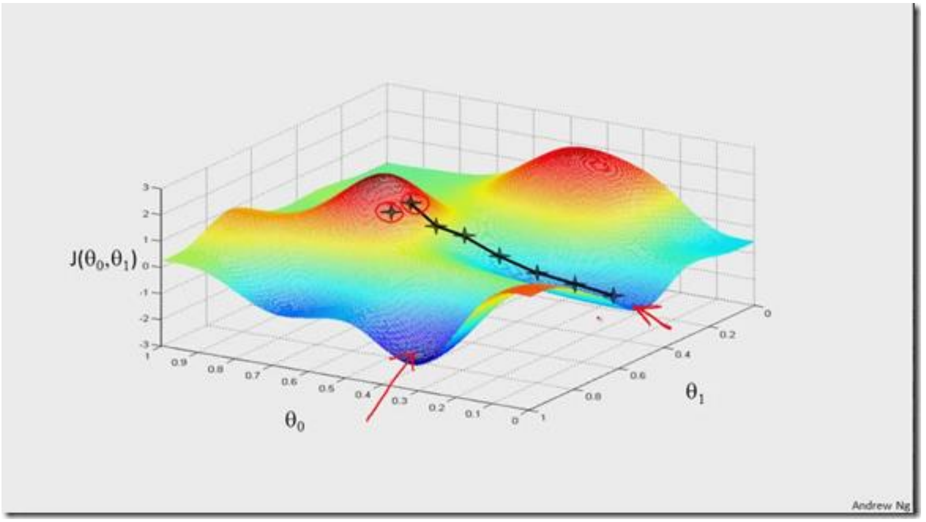
$$J(\theta)=\frac{1}{2}\sum_{i=1}^{m}(h_\theta(x^{(i)})-y^{(i)})^2.$$

其中m表示样本的数目，距离衡量采用的是2-范数的衡量方法，前面的1/2用以简化运算（后面会讲到）

      - 这种衡量方法也被称为线性最小二乘，这就是线性回归问题的一般表达形式，接下来将讨论如何在x、y已知的情况下求解出这里的θ

第二集：监督学习应用与梯度下降

- 梯度下降
  - 接下来的问题在于如何求出J(θ)取最小值的θ
    - 一种办法是搜索出这样的θ
      - 首先将θ初始化为0向量（或者其他向量都可以）
      - 接下来要通过改变θ的值使得J(θ)越来越小
    - 这种方法也称为梯度下降方法
      - 但是这种方法非常受到初始点选择的影响，最后到达的点不一定是全局最小值点，只是一个局部最小值点（极小值点）
      - 这种算法一定会结束（梯度为0说明结束，不为0说明还有下降的空间继续下降）

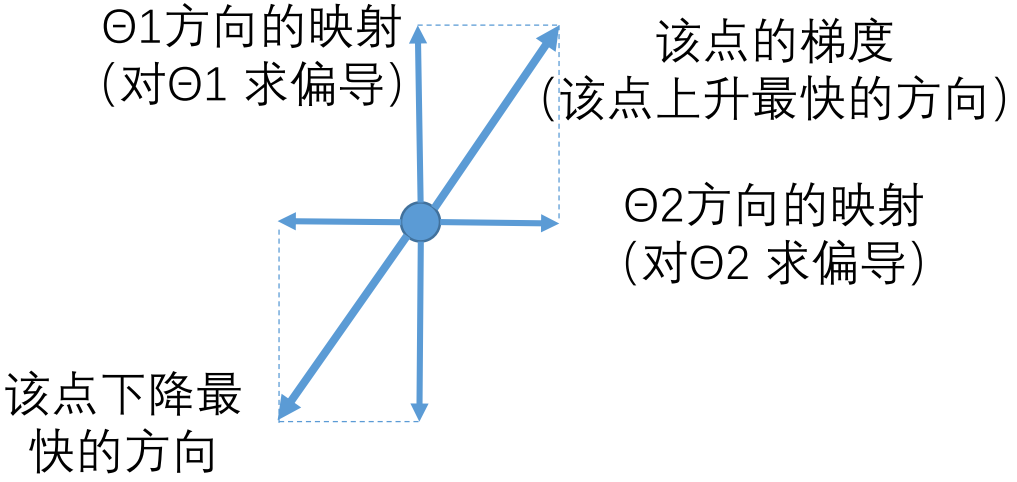


在这里纵坐标表示目标函数的值（即J(θ) 的值），两个横坐标分别表示两个参数θ1和θ2。从初始点开始总是往下降速度最快的地方走，这个地方也就是该点的梯度（就好像在山坡上，这个坡的坡度应该是指向这一点下降最快的方向的）

第二集：监督学习应用与梯度下降

- 梯度下降
  - 由此我们可以通过下式表达梯度下降的过程

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta).$$



根据多元函数微分学中梯度的定义，标量场中某一点上的梯度指向标量场增长最快的方向（前提是要可微），梯度的长度是这个最大的变化率。因此我们需要对$J(\theta)$的每一个元（在这里就是$\theta_1, \theta_2 \cdots$）求偏导数，乘上对应的方向向量后相加即可（从而得到梯度在该方向上的映射，因为梯度是个方向导数，即模为1的一个向量，方向指向标量场增长最快的方向）。由于是梯度下降，因此我们需要取个负数，来表示该方向上的最速下降。那么究竟要下降多少，这个由α值进行控制。如果α值太大，可能会出现不下降反而上升的情况；如果α值太小，收敛很慢，很容易陷入局部最优。

第二集：监督学习应用与梯度下降

- 梯度下降
  - 现在我们将原表达式J(θ)代进来
    - 先假设一种简单的情况，只有一组样本(x,y)，要对于第j个参数θj求偏导数

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

$$\begin{aligned}
\frac{\partial}{\partial\theta_j}J(\theta) &= \frac{\partial}{\partial\theta_j}\frac{1}{2}\left(h_\theta(x) - y\right)^2 \\
&= 2\cdot\frac{1}{2}\left(h_\theta(x) - y\right)\cdot\frac{\partial}{\partial\theta_j}\left(h_\theta(x) - y\right) \\
&= \left(h_\theta(x) - y\right)\cdot\frac{\partial}{\partial\theta_j}\left(\sum_{i=0}^{n}\theta_i x_i - y\right) \\
&= \left(h_\theta(x) - y\right)x_j
\end{aligned}$$

$$\theta_j := \theta_j - \alpha\frac{\partial}{\partial\theta_j}J(\theta).$$

- 代入后可得（该方法也称为LMS update rule/Widrow-Hoff learning rule）

$$\theta_j := \theta_j + \alpha\left(y^{(i)} - h_\theta(x^{(i)})\right)x_j^{(i)}.$$

第二集：监督学习应用与梯度下降

- 梯度下降
  - 单组样本(x,y)进行梯度下降的问题解决了，但如果现在不止一组样本怎么办？
    - 一种最直观的办法，根据定义将J(θ)代进去求微分

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

$$\theta_j := \theta_j - \alpha\frac{\partial}{\partial\theta_j}J(\theta).$$

$$\longrightarrow \quad \theta_j := \theta_j + \alpha \sum_{i=1}^{m} \left(y^{(i)} - h_\theta(x^{(i)})\right) x_j^{(i)}$$

    - 由此可得算法（该方法也称为batch gradient descent）

Repeat until convergence {

$$\theta_j := \theta_j + \alpha \sum_{i=1}^{m} \left(y^{(i)} - h_\theta(x^{(i)})\right) x_j^{(i)} \qquad \text{(for every } j\text{)}.$$

}

第二集：监督学习应用与梯度下降

- 梯度下降
  - 在本课的例子里面J实际上是一个凸的二次函数
    - 某个函数f(x)为凸函数当且仅当定义域为凸集，f(kx+(1-k)y)≤kf(x)+(1-k)f(y)（对于任意定义域上的x,y和0≤k≤1恒成立）--见《凸优化》第三章
      - 定义域为R，一定为凸集--见《凸优化》第二章

由于

$$\mathrm{J}(\theta)=\frac{1}{2}\sum_{i=1}^{m}(h\theta(\mathrm{xi})-\mathrm{yi})^2=\frac{1}{2}\sum_{i=1}^{m}(\theta^T(\mathrm{xi})-\mathrm{yi})^2$$

因此

$$\mathrm{J}(\mathrm{k}\theta1+(1-\mathrm{k})\theta2)=\frac{1}{2}\sum_{i=1}^{m}\left[(\mathrm{k}\theta1+(1-\mathrm{k})\theta2)^T(\mathrm{xi})-\mathrm{yi}\right]^2$$

$$=\frac{1}{2}\sum_{i=1}^{m}[\mathrm{k}(\theta1)^T(\mathrm{xi})+(1-\mathrm{k})(\theta2)^T(\mathrm{xi})-\mathrm{yi}]^2$$

$$\mathrm{kJ}(\theta1)+(1-\mathrm{k})\mathrm{J}(\theta2)$$

$$=\frac{1}{2}k\sum_{i=1}^{m}(\theta1^T(\mathrm{xi})-\mathrm{yi})^2+\frac{1}{2}(1-k)\sum_{i=1}^{m}(\theta2^T(\mathrm{xi})-\mathrm{yi})^2$$

要证

$$\mathrm{J}(\mathrm{k}\theta1+(1-\mathrm{k})\theta2)\leq\mathrm{kJ}(\theta1)+(1-\mathrm{k})\mathrm{J}(\theta2)$$

只需对于任意的 i 下式成立（令$(\theta1)^T(\mathrm{x})$为 a，$(\theta2)^T(\mathrm{x})$为 b）

$$[\mathrm{ka}+(1-\mathrm{k})\mathrm{b}-\mathrm{y}]^2\leq k(\mathrm{a}-\mathrm{y})^2+(1-k)(\mathrm{b}-\mathrm{y})^2$$

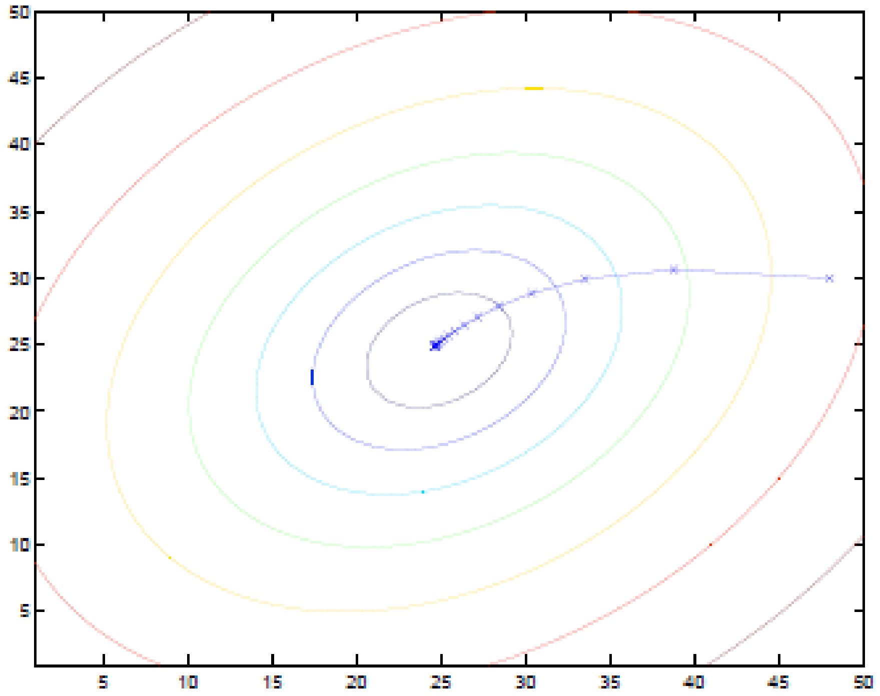
完全平方展开后左式减去右式可得

$$k^2a^2+(1-k)^2b^2+2k(1-k)ab-\mathrm{k}a^2-(1-k)b^2$$
$$=\mathrm{k}a^2(k-1)-k(1-k)b^2+2k(1-k)ab$$
$$=k(k-1)(a-b)^2$$

该结果小于等于0$(0\leq k\leq 1)$

第二集：监督学习应用与梯度下降

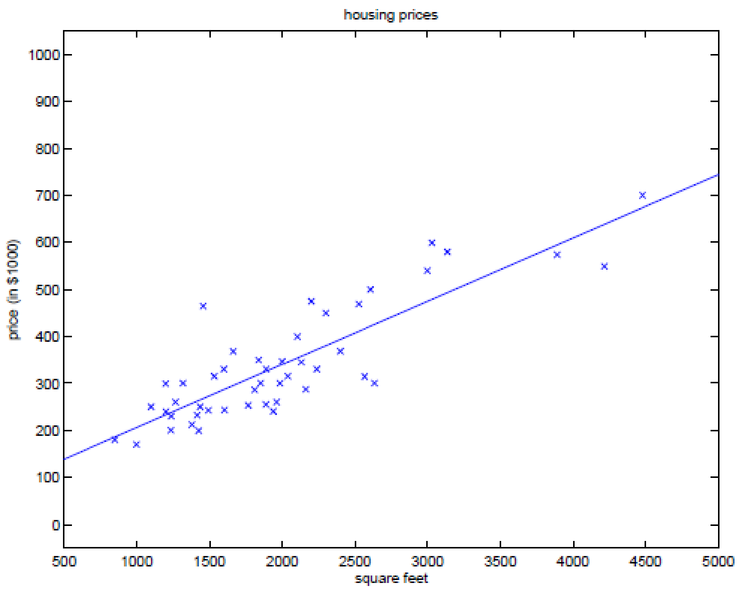
- 梯度下降
  - 由于在凸函数中任何极小值也是最小值，因此梯度下降法所求到的解一定是最小值解（也就是全局最优解）



就好比本次的凸二次函数，形状近似碗形，用梯度下降即可得最优解

第二集：监督学习应用与梯度下降

- 梯度下降
  - 通过上述的梯度下降方法即可求出给定房屋面积、房数的情况下预测房价的最佳拟合的直线



If the number of bedrooms were included as one of the input features as well, we get $\theta_0 = 89.60, \theta_1 = 0.1392$, $\theta_2 = -8.738$.

第二集：监督学习应用与梯度下降

- 梯度下降
  - 在之前讲到的batch gradient descent方法中每次迭代都需要遍历整个训练集合

Repeat until convergence {

$$\theta_j := \theta_j + \alpha \sum_{i=1}^{m} \left(y^{(i)} - h_\theta(x^{(i)})\right) x_j^{(i)} \qquad \text{(for every } j\text{)}.$$

}

  - 这导致一个问题，如果训练集合非常大，这样计算耗时很长
    - 因此后来改进了一种称为“随机梯度下降”(stochastic gradient descent/incremental gradient descent)的方法

Loop {

for i=1 to m, {

$$\theta_j := \theta_j + \alpha \left(y^{(i)} - h_\theta(x^{(i)})\right) x_j^{(i)} \qquad \text{(for every } j\text{)}.$$

}

}

第二集：监督学习应用与梯度下降

- 梯度下降
  - 通过比较batch gradient descent发现，stochastic gradient descent可以取得比batch gradient descent更快的接近全局最优值的效果。但这样也导致了一个问题：可能会在全局最优值的附近徘徊
    - 因为过程近似于我们只考虑一个样本点的情况，就好比要去上海交大的闵行校区。一开始要去上海，几乎所有人都知道哪里是上海，因此你就不需要听那么多人的建议直奔上海。但是很多人不知道交大在哪里，所以问第一个人他会告诉你一个地点。走过去发现并不是。再问第二个又告诉你一个地点，走过去好像还不是…就这样很可能在交大附近徘徊
    - 但是回过头来说这样的点在实际情况下往往比batch gradient descent发现的看上去最优的点要好，因为在机器学习中过分的准确相当于不准，因为噪声点普遍存在，数据并不是乖乖听话的

第二集：监督学习应用与梯度下降

- 正规方程
  - 之前的梯度下降是一种迭代方法，同时还有一种正规方程的方法也可以用来求解最小二乘拟合问题
  - 在此之前先复习一些矩阵的运算方法
    - 矩阵求导(matrix derivatives)的六种形式
    - 矩阵的迹(trace)

第二集：监督学习应用与梯度下降

- 正规方程
  - 广义的矩阵求导有如下六种形式

Types of Matrix Derivatives

Types	Scalar	Vector	Matrix
Scalar	$\frac{\partial y}{\partial x}$	$\frac{\partial \mathbf{y}}{\partial x}$	$\frac{\partial \mathbf{Y}}{\partial x}$
Vector	$\frac{\partial y}{\partial \mathbf{x}}$	$\frac{\partial \mathbf{y}}{\partial \mathbf{x}}$	
Matrix	$\frac{\partial y}{\partial \mathbf{X}}$		

- 其中最熟悉的是标量对标量求导（从高中开始学的那个就是）

第二集：监督学习应用与梯度下降

- 正规方程
  - 标量对向量求导仍为向量，该向量称为切向量(tangent vector)
    - 计算方法

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} \qquad \frac{\partial \mathbf{y}}{\partial x} = \begin{bmatrix} \frac{\partial y_1}{\partial x} \\ \frac{\partial y_2}{\partial x} \\ \vdots \\ \frac{\partial y_m}{\partial x} \end{bmatrix}$$

    - 例子：比如告诉你一个点在时间t内在x轴方向移动的距离和y轴方向移动的距离，问这个点在这段时间的在两个方向上的速度v是多少；或者问在在x轴、y轴方向移动的速度，问这个点在这段时间的在两个方向上的加速度a是多少

第二集：监督学习应用与梯度下降

- 正规方程
  - 向量对标量求导也同样为向量，该向量称为梯度(gradient)
    - 计算方法：

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \qquad \frac{\partial y}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y}{\partial x_1} & \frac{\partial y}{\partial x_2} & \cdots & \frac{\partial y}{\partial x_n} \end{bmatrix}$$

    - 例子：比如告诉你某一空间内某一点及其邻域的电势（电势是标量，没有方向的），问该点处电场（电场的方向指向电势变化最大的方向）如何？
    - 该运算也可记为$\bigtriangledown_u f(x)$

第二集：监督学习应用与梯度下降

- 正规方程
  - 向量对向量求导为矩阵，该矩阵称为Jacobian矩阵
    - 计算方法

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} \qquad \frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y_m}{\partial x_1} & \frac{\partial y_m}{\partial x_2} & \cdots & \frac{\partial y_m}{\partial x_n} \end{bmatrix}$$

    - 对该矩阵求行列式即为Jacobian行列式
    - 该运算也可记为df(x)

第二集：监督学习应用与梯度下降

- 正规方程
  - 标量对矩阵求导仍为矩阵，该向量称为切矩阵(tangent matrix)
    - 计算方法

$$\frac{\partial \mathbf{Y}}{\partial x} = \begin{bmatrix} \frac{\partial y_{11}}{\partial x} & \frac{\partial y_{12}}{\partial x} & \cdots & \frac{\partial y_{1n}}{\partial x} \\ \frac{\partial y_{21}}{\partial x} & \frac{\partial y_{22}}{\partial x} & \cdots & \frac{\partial y_{2n}}{\partial x} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y_{m1}}{\partial x} & \frac{\partial y_{m2}}{\partial x} & \cdots & \frac{\partial y_{mn}}{\partial x} \end{bmatrix}$$

第二集：监督学习应用与梯度下降

- 正规方程
  - 矩阵对标量求导也同样为矩阵，该向量称为梯度矩阵(gradient matrix)
    - 计算方法：

$$\frac{\partial y}{\partial \mathbf{X}} = \begin{bmatrix} \frac{\partial y}{\partial x_{11}} & \frac{\partial y}{\partial x_{21}} & \cdots & \frac{\partial y}{\partial x_{p1}} \\ \frac{\partial y}{\partial x_{12}} & \frac{\partial y}{\partial x_{22}} & \cdots & \frac{\partial y}{\partial x_{p2}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y}{\partial x_{1q}} & \frac{\partial y}{\partial x_{2q}} & \cdots & \frac{\partial y}{\partial x_{pq}} \end{bmatrix}$$

    - 这个矩阵常见于优化问题中，接下来将继续讨论这个梯度矩阵

第二集：监督学习应用与梯度下降

- 正规方程
  - 梯度矩阵的例子

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

$$f(A) = \frac{3}{2}A_{11} + 5A_{12}^2 + A_{21}A_{22} \longrightarrow \nabla_A f(A) = \begin{bmatrix} \frac{3}{2} & 10A_{12} \\ A_{22} & A_{21} \end{bmatrix}$$

第二集：监督学习应用与梯度下降

- 正规方程
  - 矩阵的迹：方阵A的迹被定义为矩阵主对角线上的所有元素之和，记为tr(A)
    - 矩阵的迹满足以下基本性质（证明详见线性代数）

$$\mathrm{tr}ABC = \mathrm{tr}CAB = \mathrm{tr}BCA,$$

$$\mathrm{tr}ABCD = \mathrm{tr}DABC = \mathrm{tr}CDAB = \mathrm{tr}BCDA$$

$$\mathrm{tr}A = \mathrm{tr}A^T$$

$$\mathrm{tr}(A+B) = \mathrm{tr}A + \mathrm{tr}B$$

$$\mathrm{tr}\, aA = a\mathrm{tr}A$$

第二集：监督学习应用与梯度下降

- 正规方程
  - 将矩阵的迹和梯度矩阵结合在一起，可以得到以下基本的性质
    - 1、

$$\nabla_A \mathrm{tr} AB = B^T$$

假设矩阵 A、B 为

$$A=\begin{bmatrix} a11 & \cdots & a1n \\ \vdots & \ddots & \vdots \\ an1 & \cdots & ann \end{bmatrix}, B=\begin{bmatrix} b11 & \cdots & b1n \\ \vdots & \ddots & \vdots \\ bn1 & \cdots & bnn \end{bmatrix}$$

可求得 AB 为

$$AB=\begin{bmatrix} \sum_{i=1}^{n} a1ibi1 & \cdots & \sum_{i=1}^{n} a1ibin \\ \vdots & \ddots & \vdots \\ \sum_{i=1}^{n} anibi1 & \cdots & \sum_{i=1}^{n} anibin \end{bmatrix}$$

因此

$$\mathrm{tr}(AB)=\sum_{j=1}^{n}\sum_{i=1}^{n} ajibij$$

因此

$$\nabla A\mathrm{tr}(AB)=\begin{bmatrix} \frac{\partial \mathrm{tr}(AB)}{\partial a11} & \cdots & \frac{\partial \mathrm{tr}(AB)}{\partial a1n} \\ \vdots & \ddots & \vdots \\ \frac{\partial \mathrm{tr}(AB)}{\partial an1} & \cdots & \frac{\partial \mathrm{tr}(AB)}{\partial ann} \end{bmatrix}=\begin{bmatrix} b11 & \cdots & bn1 \\ \vdots & \ddots & \vdots \\ b1n & \cdots & bnn \end{bmatrix}=B^T$$

第二集：监督学习应用与梯度下降

- 正规方程
  - 将矩阵的迹和梯度矩阵结合在一起，可以得到以下基本的性质
    - 2、

$$\nabla_{A^T} f(A) = (\nabla_A f(A))^T$$

假设矩阵 A 为

$$A = \begin{bmatrix} a11 & \cdots & a1n \\ \vdots & \ddots & \vdots \\ an1 & \cdots & ann \end{bmatrix}$$

可得

$$\nabla A^T f(A) = \begin{bmatrix} \frac{\partial f(A)}{\partial a11} & \cdots & \frac{\partial f(A)}{\partial an1} \\ \vdots & \ddots & \vdots \\ \frac{\partial f(A)}{\partial a1n} & \cdots & \frac{\partial f(A)}{\partial ann} \end{bmatrix} = \begin{bmatrix} \frac{\partial f(A)}{\partial a11} & \cdots & \frac{\partial f(A)}{\partial a1n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f(A)}{\partial an1} & \cdots & \frac{\partial f(A)}{\partial ann} \end{bmatrix}^T = \nabla A f(A)^T$$

第二集：监督学习应用与梯度下降

- 正规方程
  - 将矩阵的迹和梯度矩阵结合在一起，可以得到以下基本的性质
    - 3、

$$\nabla_A \mathrm{tr} ABA^T C = CAB + C^T AB^T$$

这个的证明要借助第一个定理

$$\nabla A\mathrm{tr}(AB) = B^T$$

但是在这个定理里面实际上存在两个 A，因此需要分开来求偏微分。当我们对于 A 进行偏微分的时候，将 A^T 看成是与 A 无关的系数。反之亦然，当我们对于 A^T 进行偏微分的时候，将 A 看成是与 A^T 无关的系数。

因此

$$\nabla A\mathrm{tr}(ABA^T C) = \nabla A\mathrm{tr}(ABA^T C) + \nabla A\mathrm{tr}(A^T CAB) = (BA^T C)^T + CAB = C^T AB^T + CAB$$

第二集：监督学习应用与梯度下降

- 正规方程
  - 接下来将讲述如何通过本方法解决之前提到的θ的取值
    - 首先定义矩阵X为所有样本的输入，Y为所有样本的输出

$$X = \begin{bmatrix} — (x^{(1)})^T — \\ — (x^{(2)})^T — \\ \vdots \\ — (x^{(m)})^T — \end{bmatrix}$$

第一行表示第一个样本，第二行表示第二个样本，以此类推。然后列数取决于每个样本的自变量数（比如说这个例子中，有房屋大小和房费两个自变量）

$$\vec{y} = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(m)} \end{bmatrix}$$

第一行表示第一个样本，第二行表示第二个样本，以此类推。在这里只有一列表示预测的房价

- 由此可以计算h(X)的值

$$h(x) = \sum_{i=0}^{n} \theta_i x_i = \theta^T x \longrightarrow \begin{bmatrix} (x^{(1)})^T\theta \\ \vdots \\ (x^{(m)})^T\theta \end{bmatrix}$$

其实h(X)的每一行就是对应样本的h(x)值

第二集：监督学习应用与梯度下降

- 正规方程
  - 同样地，我们需要计算目标函数就需要将h(X)和Y作差后求内积

$$
\begin{aligned}
X\theta - \vec{y} &= \begin{bmatrix} (x^{(1)})^T\theta \\ \vdots \\ (x^{(m)})^T\theta \end{bmatrix} - \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(m)} \end{bmatrix} \\
&= \begin{bmatrix} h_\theta(x^{(1)}) - y^{(1)} \\ \vdots \\ h_\theta(x^{(m)}) - y^{(m)} \end{bmatrix}.
\end{aligned}
$$

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

$$
\begin{aligned}
\frac{1}{2}(X\theta - \vec{y})^T(X\theta - \vec{y}) &= \frac{1}{2}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2 \\
&= J(\theta)
\end{aligned}
$$

第二集：监督学习应用与梯度下降

- 正规方程
  - 再接下来，和之前一样需要求J(θ)的最小值
    - 求最小值的办法很简单，既然J(θ)是个凸函数（正在前面证明过），其极小值就是最小值，因此可以说导数（梯度）为0的点就是最小值点
    - 因此我们需要对J(θ)进行求导，其过程如下：

$$
\begin{aligned}
\nabla_\theta J(\theta) &= \nabla_\theta \frac{1}{2}(X\theta-\vec{y})^T(X\theta-\vec{y}) \\
&= \frac{1}{2}\nabla_\theta \left(\theta^T X^T X\theta - \theta^T X^T \vec{y} - \vec{y}^T X\theta + \vec{y}^T\vec{y}\right) \\
&= \frac{1}{2}\nabla_\theta \operatorname{tr}\left(\theta^T X^T X\theta - \theta^T X^T \vec{y} - \vec{y}^T X\theta + \vec{y}^T\vec{y}\right) \\
&= \frac{1}{2}\nabla_\theta \left(\operatorname{tr}\theta^T X^T X\theta - 2\operatorname{tr}\vec{y}^T X\theta\right) \\
&= \frac{1}{2}\left(X^T X\theta + X^T X\theta - 2X^T\vec{y}\right) \\
&= X^T X\theta - X^T\vec{y}
\end{aligned}
$$

1式到2式直接展开； 2式到3式使用了任意实数的迹等于它本身的性质(tr(a)=a, a∈R)； 3式到4式运用了迹的加法性质(trA+trB=tr(AB))、乘法性质(tr(ABC)=tr(BCA)=tr(CAB))和转置性质(trA^T=trA)，同时由于最后一项与θ无关，微分后必定为0，因此直接消去； 4式到5式使用了迹的梯度矩阵的性质1和性质3（这两个公式见下，在使用性质3时A,B,A^T,C,分别对应θ,I(单位向量), $θ^T$,X^TX)

$$
\begin{aligned}
\nabla_A \operatorname{tr}AB &= B^T \\
\nabla_A \operatorname{tr}ABA^TC &= CAB + C^TAB^T
\end{aligned}
$$

第二集：监督学习应用与梯度下降

- 正规方程
  - 由此得到了最终的结果，现在我们让导数为0，即可得到以前线性代数中学过的一个式子（用来求解最小二乘问题的关系式），这个式子也被称为正规方程

$$X^T X\theta = X^T \vec{y}$$

  - 那么对于θ的值可以一步完成求解

$$\theta = (X^T X)^{-1} X^T \vec{y}.$$