CS229-note1-学习心得

第一部分

线性回归

简介：

1.背景：地区 *房间大小和价格* 数据，推进如何执行监督学习，以及确定参数 *thetaθ* 。

将 **y** 假设为关于 **x** 的线性函数，**θ** 代表拟合参数

于是：

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2$$

为了简单起见，假定 **X0 = 1**

于是：

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

如何确定 *θ*？

于是：

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

注：1/2参考方差

2.LMS算法-最小均方，也称为 Widrow-Hoff 学习规则，也即最小梯度下降

为了使J(θ) 更小，直到我们希望收敛到使 J(θ) **最小化**的 θ 值

于是：

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

注：α 称为学习率。这是一种非常自然的算法，它会在 J 的最陡下降方向上重复迈出一步

公式2-1

假设我们有一个训练实例（x,y）,我们需要计算出右侧的偏导数

于是：

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

公式 2-2

接着，结合**公式2-1和2公式-2**

于是：

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

注：LMS更新规则（LMS代表“最小均方”），也称为 Widrow-Hoff 学习规则。

如果在拟合过程中，我们的预测值和实际值有较大出入，以下是两种解决方法：

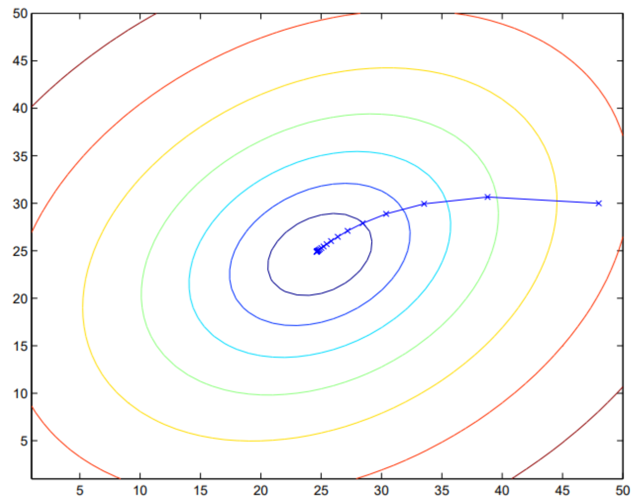
***First*.对于包含多个训练集，替换为以下算法，重复直到收敛：**

$$\begin{aligned}
&\text{Repeat until convergence } \{ \\
&\qquad \theta_j := \theta_j + \alpha \textstyle\sum_{i=1}^{m} \left(y^{(i)} - h_\theta(x^{(i)})\right) x_j^{(i)} \qquad (\text{for every } j). \\
&\}
\end{aligned}$$

注：求和部分为 *∂J(θ)/∂θj (for the original definition of J).*

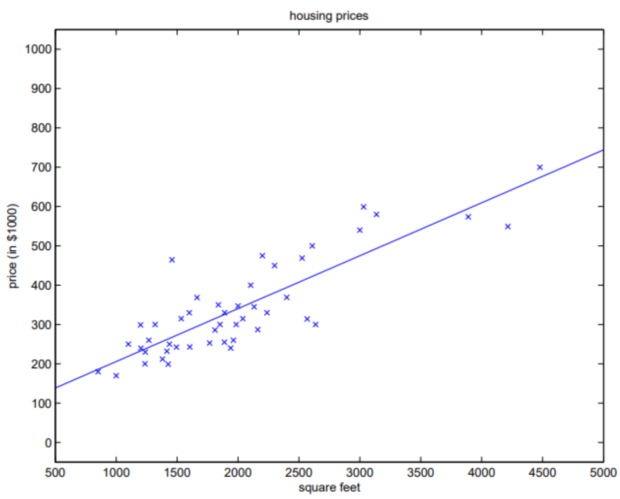
- 此方法被称为**批梯度下降**，每一步都扫描整个**训练集**的每个实例
- 此方法受**局部极小值**影响大，但是**线性回归问题**只有一个全局最优值，因此此梯度下降总是**收敛**
- 假设 **学习率 α** 不太大

下面是对比两种方法的直观图：



The ellipses shown above are the contours of a quadratic function. Also shown is the trajectory taken by gradient descent, which was initialized at (48,30). The x's in the figure (joined by straight lines) mark the successive values of θ that gradient descent went through.

梯度下降实例



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

批梯度下降实例

***Second.*考虑以下算法：**

$$
\begin{aligned}
&\text{Loop } \{ \\
&\qquad \text{for i=1 to m, } \{ \\
&\qquad\qquad \theta_j := \theta_j + \alpha \left(y^{(i)} - h_\theta(x^{(i)}) \right) x_j^{(i)} \qquad (\text{for every } j). \\
&\qquad \} \\
&\}
\end{aligned}
$$

- 根据**单个训练样例**相关误差梯度更新参数
- **随机梯度下降**
- 不用每次都扫描**整个训练集**
- 更快 ***接近*** **θ最小值**，可能永远无法到达最小值，在**最小值附近振荡**

3.正规方程

不依靠**迭代算法**求最小值的算法

- **学习率** **α** 减小到 0
- **参数** **θ** 收敛到最小值而**非振荡**

3.1矩阵导数

为了简化书写以及，避免大量的导数矩阵出现

于是：

函数 f 对于 **矩阵 A** 的导数：

$$
\nabla_A f(A) = \begin{bmatrix} \frac{\partial f}{\partial A_{11}} & \cdots & \frac{\partial f}{\partial A_{1n}} \\ \vdots & \ddots & \vdots \\ \frac{\partial f}{\partial A_{m1}} & \cdots & \frac{\partial f}{\partial A_{mn}} \end{bmatrix}
$$

公式1

我们定义 **矩阵A:**

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

函数 f：

$$f(A) = \frac{3}{2}A_{11} + 5A_{12}^2 + A_{21}A_{22}.$$

公式2

因此由**公式1和2**得出：

$$\nabla_A f(A) = \begin{bmatrix} \frac{3}{2} & 10A_{12} \\ A_{22} & A_{21} \end{bmatrix}.$$

如果 矩阵A 和 矩阵B相乘，使得矩阵AB为方阵，那么有：**tr*AB* = tr*BA***

以下再简单介绍几条性质，**（4）仅适用于非奇异方阵**

$$\begin{aligned} \nabla_A \mathrm{tr} AB &= B^T && (1) \\ \nabla_{A^T} f(A) &= (\nabla_A f(A))^T && (2) \\ \nabla_A \mathrm{tr} ABA^T C &= CAB + C^T AB^T && (3) \\ \nabla_A |A| &= |A|(A^{-1})^T. && (4) \end{aligned}$$

（1）阐述：

$$B \in Rn \times m.$$

$$f : Rm \times n \to R$$

$$f(A) = trAB.$$

$$A \in Rm \times n$$

矩阵A，B相乘为**方阵**，于是在矩阵B中，N行都被赋予矩阵A的M_i。类似的，列元素也是，对函数求导后，A的每一项求导为1，即**m行，n列被保留。**即为**B的转置。**

（2）阐述：根据（1），想象一下，如果求导对象是A的转置呢？

（3）阐述：根据（1），（2）。

3.2最小二乘再访

给定一个矩阵 X 为 m * n

把要训练的实例输入X的行中：

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

目标值定义为 Y 向量：

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

于是：

$$h\theta(x(i)) = (x(i))T\theta$$

所以：

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

根据 $z^T z = \sum_i z_i^2$：

我们有：

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

结合（2），（3）我们得出：

$$\nabla_{A^T}\operatorname{tr}ABA^TC = B^TA^TC^T + BA^TC \tag{5}$$

推出：

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

注：

- 推导步骤3，**实数的迹就是实数**
- 推导步骤4，**矩阵A的迹就是矩阵A转置的迹**
- 推导步骤5，使用上述等式（5）
- 其中 **AT = θ, B = BT = XT X, and C = I**
- 然后使用了上述等式（1）

接下来将 **J(θ)** 的导数设为 **0**，得到：

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

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

4.概率解释

5.局部加权线性回归

第二部分

分类和逻辑回归