

1 线性回归 (Linear Regression)

1.1 一些理论

由于下面要说明 $J(\theta)$ 是凸函数，为了方便说明，先证明 $f(x) = x^2$ 是凸的。

Lemma 1.1. *Let $x \in \mathbb{R}, f(x) = x^2$. Show f is convex.*

Proof. 对任意的 $\alpha \in [0, 1]$,

$$\begin{aligned} f(\alpha x_1 + (1 - \alpha)x_2) - \alpha f(x_1) - (1 - \alpha)f(x_2) \\ = (\alpha^2 - \alpha)(x_1 + x_2)^2, \end{aligned}$$

由于 $\alpha \in [0, 1]$, 所以 $\alpha^2 - \alpha \leq 0$ 。因此 $(\alpha^2 - \alpha)(x_1 + x_2)^2 \leq 0$ 。我们得到

$$f(\alpha x_1 + (1 - \alpha)x_2) \leq \alpha f(x_1) + (1 - \alpha)f(x_2),$$

由凸函数的定义立即得知 f 是凸的。 □

之所以可以对 loss 函数运用梯度下降法得到最优解，正是由于 loss 的凸性（凹性）才保证了全局最优解的存在性。

Question 1. Let $x \in \mathbb{R}^{d+1}$ and

$$J(\theta) = \frac{1}{2} \sum_{i=1}^n (h_{\theta}(x^{(i)}) - y^{(i)})^2 \quad (1.1)$$

and

$$h_{\theta}(x) = \sum_{i=0}^d \theta_i x_i, \quad (1.2)$$

证明 $J(\theta)$ 是凸的。

Proof. 对任意的 $\theta^1, \theta^2 \in \mathbb{R}^{d+1}, \alpha \in [0, 1]$

$$\begin{aligned} h_{\alpha\theta^1 + (1-\alpha)\theta^2}(x) &= \sum_{i=0}^d (\alpha\theta_i^1 + (1 - \alpha)\theta_i^2)x_i \\ &= \alpha \sum_{i=0}^d \theta_i^1 x_i + (1 - \alpha) \sum_{i=0}^d \theta_i^2 x_i \\ &= \alpha h_{\theta^1}(x) + (1 - \alpha) h_{\theta^2}(x). \end{aligned}$$

由 Lemma 1.1 以及 J 的定义，有

$$\begin{aligned}
J(\alpha\theta^1 + (1-\alpha)\theta^2) &= \frac{1}{2} \sum_{i=1}^n \left(h_{\alpha\theta^1 + (1-\alpha)\theta^2}(x) - y^{(i)} \right)^2 \\
&= \frac{1}{2} \sum_{i=1}^n \left(\alpha h_{\theta^1}(x^{(i)}) + (1-\alpha)h_{\theta^2}(x^{(i)}) - y^{(i)} \right)^2 \\
&= \frac{1}{2} \sum_{i=1}^n \left(\alpha [h_{\theta^1}(x^{(i)}) - y^{(i)}] + (1-\alpha)[h_{\theta^2}(x^{(i)}) - y^{(i)}] \right)^2 \\
&\leq \frac{1}{2} \sum_{i=1}^n \alpha \left(h_{\theta^1}(x^{(i)}) - y^{(i)} \right)^2 + (1-\alpha) \left(h_{\theta^2}(x^{(i)}) - y^{(i)} \right)^2 \\
&= \alpha J(\theta^1) + (1-\alpha)J(\theta^2).
\end{aligned}$$

所以， J 关于 θ 是凸的。 \square

除了可以用凸函数的定义证明，还可以用凸函数的等价条件 $J''(\theta) \geq 0$ 进行证明。

θ 的值通过梯度下降法迭代产生。若 x 是一个一维数，则 θ 是一个二维向量（还有截距项）。下面的算法在具体编写代码时，还需要确定迭代的次数 n ，可以设置大一点。还要确定学习率 α ，学习率很大则迭代时会很快，但也容易出错。

$$\theta = (\theta_0, \theta_1),$$

$$\theta_j := \theta_j - \alpha \frac{1}{m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)}) x_j^{(i)},$$

when dim is 2, $j = 0, 1$. X is a matrix of vector of x , like this

$$X = (x^{(1)}, x^{(2)}, x^{(3)}, \dots, x^{(m)}),$$

and

$$x = (x_0^{(1)}, x_1^{(1)}),$$

here we suppose x is 2-dm. $x_0^{(1)}$ and $x_1^{(1)}$ are the first and second elements of $x^{(1)}$ respectively, both of them are scalars.

1.2 梯度下降法

梯度下降法可以在凸（凹）函数中找到最值点，要理解这两个概念：函数 f 的方向导数，行数的梯度。

Definition 1.1. 设三元函数 f 在点 $P_0(x_0, y_0, z_0)$ 的邻域 $U(P_0)$ 内有定义, l 为从点 P_0 出发的射线, $P(x, y, x)$ 为 l 上且含于 $U(P_0)$ 的任意点, ρ 是 P 与 P_0 之间的距离。若极限

$$\lim_{\rho \rightarrow 0^+} \frac{f(P) - f(P_0)}{\rho}$$

存在, 则称此极限为函数 f 在点 P_0 沿方向 l 的方向导数, 记为 $f_l(P_0)$ 。

$u = (u_1, u_2, u_3)$ 是射线 l 的单位方向当函数 f 在点 $P_0(x_0, y_0, z_0)$ 可微时, 方向导数写为

$$f_l(P_0) = \sum_{i=1}^3 f_i(P_0)u_i,$$

其中, $f_i(P_0)$ 是函数 f 关于 P_0 的第 i 个分量的偏导。

Definition 1.2. 若函数 $f(x, y, z)$ 在 P_0 对所有自变量的偏导都存在, 则偏导向量 ∇f 称为 f 在点 P_0 的梯度,

$$\nabla f = (f_x(P_0), f_y(P_0), f_z(P_0)).$$

因此, 函数 f 在点 P_0 的方向导数可以写为

$$f_l(P_0) = f_x(P_0)u_1 + f_y(P_0)u_2 + f_z(P_0)u_3 = \nabla f \cdot u.$$

因此,

$$f_l(P_0) = \nabla f \cdot u = |\nabla f| \cos \theta,$$

其中, θ 是射线方向与梯度的夹角。所以, 当 $\theta = 0$ 时, 也就是梯度方向 ∇f 是 f 的值增长最快的方向, 同理, 负梯度方向 $-\nabla f$ 是 f 减小最快的方向。

1. 由于要求关于 θ 的函数 $J(\theta)$ 的极小值, 所以 θ 沿着方向 $-\nabla J$ 变化, 速度最快。
2. 因此, 下一个会使得 J 的值更小的点是 $\theta + (-\nabla J)$ 。
3. J 的凸性保证了最小值点存在。