机器学习笔记

空修菜

1 线性回归（Linear Regression）

1. 线性回归模型可以使用梯度下降法求得全局最优解 (迭代是收敛的) 是因为假设函数损失函数 $J(\theta)$ 是凸的. 为方便说明 $J(\theta)$ 是凸函数, 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 $\alpha \in [0, 1]$,

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

   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. □

2. 说明损失函数的凸性.

   Theorem 1.1. *Let $x \in \mathbb{R}^{d+1}$ and*

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

   and

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

   证明 $J(\theta)$ 是凸的.

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

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

由 Lemma1.1以及 J 的定义，有

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

所以，J 关于 θ 是凸的。 □

3. 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 $J''(\theta) \ge 0$ 进行证明.

4. θ 的值通过梯度下降法迭代产生. 若 $x \in \mathbb{R}$, 则 θ 是一个二维向量 (还有截距项). 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

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

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.

5. 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, 函数的梯度.

6. 方向导数的定义.

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

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

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。

7. 梯度的定义.

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,

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

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,

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

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.

8. 迭代公式的矩阵表示. 当 $x \in \mathbb{R}^n$ 时, $\theta \in \mathbb{R}^{n+1}$. $\theta = (\theta_0, \theta_1, \ldots, \theta_n), 0 \leq j \leq n$, 由于

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

$$\begin{aligned}(\theta_0, \ldots, \theta_n) &:= (\theta_0, \ldots, \theta_n) - \frac{\alpha}{m} \Big(h_\theta(x^{(1)}) - y^{(1)}, \cdots, h_\theta(x^{(m)}) - y^{(m)} \Big) \begin{pmatrix} x_0^{(1)} & \cdots & x_n^{(1)} \\ \vdots & \ddots & \vdots \\ x_0^{(m)} & \cdots & x_n^{(m)} \end{pmatrix} \\ &= (\theta_0, \ldots, \theta_n) - \frac{\alpha}{m} \Big(h_\theta(x^{(1)}) - y^{(1)}, \cdots, h_\theta(x^{(m)}) - y^{(m)} \Big) X\end{aligned}$$

9. 高斯分布的定义.

Definition 1.3. 若随机变量 X 的密度函数是

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < \mu < \infty$$

则称 X 服从以 $\sigma > 0$ 和 μ 为参数的高斯分布.

10. 概率的解释.

(1). 当 θ 已知, $h_\theta(x^{(i)})$ 估计的值与实际值可能存在一个随机的误差 $\varepsilon^{(i)}$, 即

$$y^{(i)} = \theta^T x^{(i)} + \varepsilon^{(i)},$$

这个随机误差一般服从以 0 为均值的高斯分布, 即 $\epsilon \sim \mathcal{N}(0, \sigma^2)$, ϵ 表示误差的随机变量, 它的取值是 ε;

(2). y 的分布. 因为 $y - \theta^T X = \epsilon$, 由 ϵ 的密度函数, 有

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{\epsilon^2}{2\sigma^2}} = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-\theta^T X)^2}{2\sigma^2}} = p(y),$$

由高斯分布的密度函数的定义可知, y 服从均值为 $\theta^T X$, 方差为 σ^2 的高斯分布, 即

$$p(y \mid X; \theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\Big(- \frac{(y - \theta^T X)^2}{2\sigma^2} \Big).$$

11. 似然函数 (Likelihood function). $X = (x^{(1)}, \cdots, x^{(m)})$, $x^{(i)}, y^{(i)} \in \mathbb{R}^n$, $y = (y^{(1)}, \cdots, y^{(m)})$, 假设 $(x^{(i)}, y^{(i)})$ 与 $(x^{(j)}, y^{(j)})$ 之间是独立同分布的 (idd), 由乘法原理,

$$\begin{aligned}\mathcal{L}(\theta) = p(y \mid X; \theta) &= \prod_{i=1}^{m} p(y^{(i)} \mid x^{(i)}; \theta) \\ &= \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\sigma} \exp\Big(- \frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2} \Big).\end{aligned}$$

12. 最大似然估计.

    (1). 现在的目标是选取 θ 使得函数 $\mathcal{L}(\theta)$ 的值最大. 由于 $\mathcal{L}(\theta)$ 有阶乘, 所以对 $\mathcal{L}(\theta)$ 取对数. 实际上, θ_0 使得 $\mathcal{L}(\theta)$ 取到最大值, 也必然使 $\log \mathcal{L}(\theta)$ 取得最大值, 两者等价;

    (2). 对数似然函数.

$$\begin{aligned}\ell(\theta) = \log \mathcal{L}(\theta) &= \log \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\sigma} \exp\Big(- \frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2} \Big) \\ &= m \log \frac{1}{\sqrt{2\pi}\sigma} + \sum_{i=1}^{m} - \frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2} \\ &= m \log \frac{1}{\sqrt{2\pi}\sigma} - \sum_{i=1}^{m} \frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2},\end{aligned}$$

    (3). 由于 $m \log \frac{1}{\sqrt{2\pi}\sigma}$ 不含参数 θ, 所以 $m \log \frac{1}{\sqrt{2\pi}\sigma}$ 关于 θ 是一个常数. 若它为正, $\ell(\theta)$ 要取最大值, 就要使后一项尽可能地小; 即让 $\sum_{i=1}^{m} (y^{(i)} - \theta^T x^{(i)})^2$ 尽可能地小;

    (4). 对损失函数 $J(\theta)$

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

    取最小值的解释的角度有两个:

    - 损失函数 $J(\theta)$ 的最小二乘法;
    - 概率分布的极大似然.