说明文档

2.2.1

$$J(\theta)=\frac{1}{m}(y-X\theta)^T(y-X\theta)$$

2.2.2

$$\frac{\partial J(\theta)}{\partial\theta}=\frac{2}{m}X^T\left(X\theta-y\right)$$

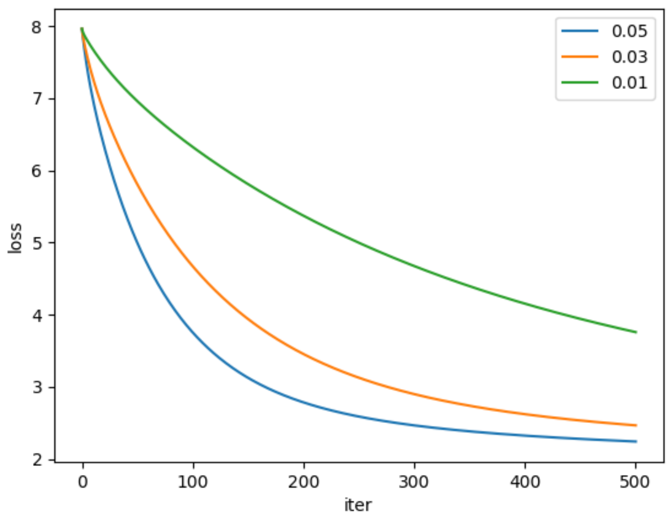
2.2.3 设当前梯度为 $g=\nabla_\theta J(\theta)$，则

$$J(\theta+\eta h)-J(\theta)\approx\eta h^Tg$$

2.2.4

$$\theta^{t+1}\leftarrow\theta^t-\frac{2\eta}{m}X^T\left(X\theta-y\right)$$

2.2.9 学习率大于 0.05 时损失函数发散，实验中在学习率为 0.05、0.03、0.01 时损失函数值随训练时间变化如下图：



2.3.1

$$\frac{\partial J(\theta)}{\partial\theta}=\frac{2}{m}X^T\left(X\theta-y\right)+2\lambda\theta$$

2.3.4 经过 500 轮训练，最终的验证集 Loss 值为：

λ \ 步长	0.05	0.03	0.01
100	发散	发散	发散
10	3.2813	4.9106	4.9106
1	2.3561	4.4245	4.4245
10^{-1}	2.9586	2.9676	3.2113
10^{-3}	2.6408	2.5108	2.7847
10^{-5}	2.6660	2.5194	2.7805
10^{-7}	2.6662	2.5195	2.7805

2.4.1

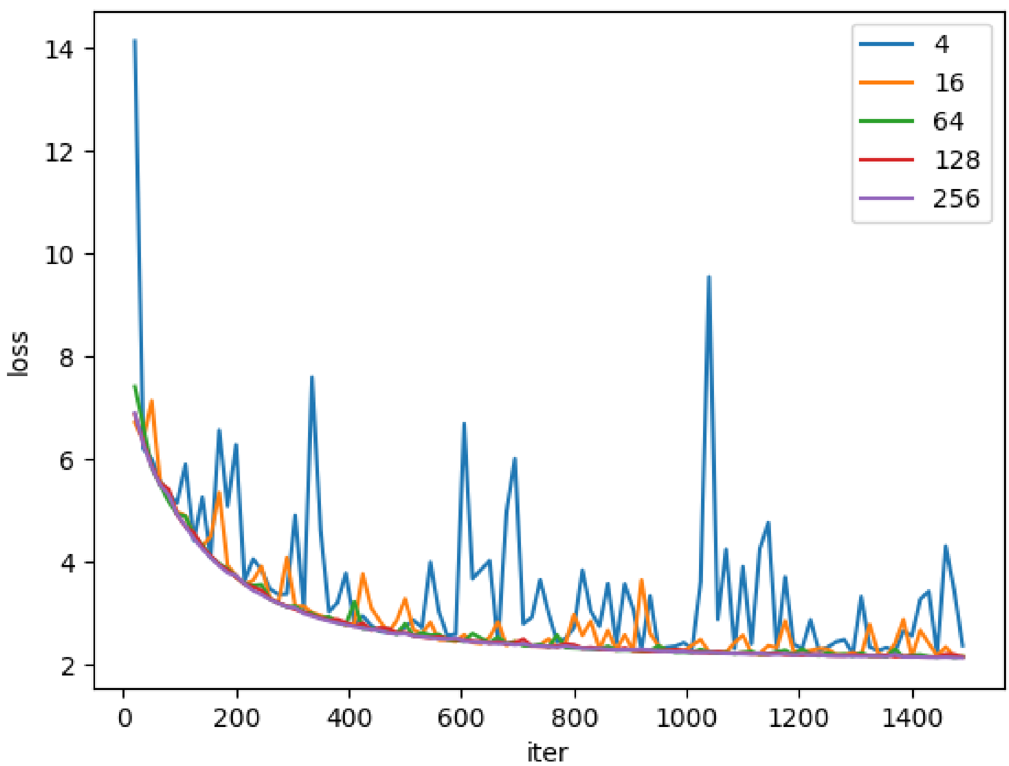
$$f_i(\theta) = (h_\theta(x_i) - y_i)^2 + \frac{\lambda}{m}\theta^T\theta$$

2.4.2 已知 $J(\theta) = \frac{1}{m}\sum_{i=1}^{m} f_i(\theta)$，故:

$$\begin{aligned}\mathbb{E}\left[\nabla f_i(\theta)\right] &= \frac{1}{m}\sum_{i=1}^{m}\nabla f_i(\theta)\\ &= \nabla\left(\frac{1}{m}\sum_{i=1}^{m} f_i(\theta)\right)\\ &= \nabla J(\theta)\end{aligned}$$

2.4.4 当 batch_size 为 4、16、64、128、256 时的训练曲线如下图：（由于 batch_size 数值小的情况下曲线抖动十分剧烈，因此 batch_size = 1 的曲线没有绘制在图中；为了避免训练中偶然出现的巨大 loss 数值导致图片难以绘制，因此从第 20 轮开始、每隔 15 轮绘制一次 loss 值)

可见随着 batch_size 的增大，训练损失曲线越来越平滑。



3.1.1 根据次梯度定义，有：

$$\partial f(x) = \bigcap_{z \in \mathrm{dom}\, f} \left\{ g \mid f(z) \geq f(x) + g^T(z-x) \right\}$$

设 $f_k(x)$ 在区间 D 上满足 $f(x) = f_k(x)$，而由于 $g \in \partial f_k(x)$，在区间 D 上有 $f_k(z) \geq f_k(x) + g^T(z-x)$，其中 $z \in \mathrm{dom} f$，$x \in D$。而在 $\mathrm{dom} f$ 上 $f(z) \geq f_k(z)$，故 $f(z) \geq f(x) + g^T(z-x)$，其中 $z \in \mathrm{dom} f$，$x \in D$，即在 D 上 $g \in \partial f(x)$。而 $\mathrm{dom} f = \bigcup_{1 \leq i \leq m} D_i$，故 $g \in \partial f(x)$。

3.1.2

$$\begin{aligned}
\frac{\partial J(w)}{\partial w} &= \frac{\partial \max\{0, 1 - y w^T x\}}{\partial x} \\
&= \begin{cases} \frac{0}{\partial w} & 1 - y w^T x < 0 \\ \frac{\partial (1 - y w^T x)}{\partial w} & 1 - y w^T x \geq 0 \end{cases} \\
&= \begin{cases} 0 & y w^T x \geq 1 \\ -yx & y w^T x < 1 \end{cases}
\end{aligned}$$

3.2.1 $\ell(\hat{y}_i, y_i)$ 的次梯度如下：

$$\begin{aligned}
\frac{\partial \ell(\hat{y}_i, y_i)}{\partial w} &= \frac{\partial \max\{0, -\hat{y}_i y_i\}}{\partial w} \\
&= \frac{\partial \max\{0, -w^T x_i y_i\}}{\partial w} \\
&= \begin{cases} \frac{\partial 0}{\partial w} & w^T x_i y_i < 0 \\ \frac{\partial (-w^T x_i y_i)}{\partial w} & w^T x_i y_i \geq 0 \end{cases} \\
&= \begin{cases} 0 & y_i w x_i^T < 0 \\ -y_i x_i & y_i w x_i^T \geq 0 \end{cases}
\end{aligned}$$

采用步长 η 为 1 的 SSGD 算法时，

$$w_{t+1} = w_t - \eta \frac{\partial \ell(\hat{y}_i, y_i)}{\partial w_t} = \begin{cases} w_t & y_i w x_i^T < 0 \\ w_t + y_i x_i & y_i w x_i^T \geq 0 \end{cases}$$

与感知机算法代码的逻辑一致。

3.2.2 初始时，$w_0 = 0$，不妨设 $w_0 = \sum_{i=1}^n \alpha_{0i} x_i$，其中 $\alpha_{0i} = 0, 1 \leq i \leq n$。

假设 $w_t = \sum_{i=1}^n \alpha_{ti} x_i$，则用数据 (x_j, y_j) 将 w_t 更新为 w_{t+1} 时，若 $y_i x_i w_t > 0$，则 $w_{t+1} = w_t$，故 $w_{t+1} = \sum_{i=1}^n \alpha_{ti} x_i$；若 $y_i x_i w_t \leq 0$，则 $w_{t+1} = w_t + y_j x_j = \sum_{i=1, i \neq j}^{i=n} \alpha_{ti} x_i + (\alpha_{tj} + y_j) x_j$。

由数学归纳法可知，最终输出的 w 同样可表示为 $\sum_{i=1}^n \alpha_i x_i$。

3.3.1 令 $\xi_i = 1 - y_i(w^T x_i + b)$，则原问题可重新表述为：

$$\min_{w,b,\xi} \frac{\lambda}{2} \|w\|^2 + \frac{1}{m} \sum_{i=1}^m \xi_i, \text{ s.t. } y_i(w^T x_i + b) \geq 1 - \xi_i,\ \xi_i \geq 0, 1 \leq i \leq m$$

对应的拉格朗日函数：

$$L(w, b, \xi, \alpha, \mu) = \frac{\lambda}{2} \|w\|^2 + \frac{1}{m} \sum_{i=1}^m \xi_i + \sum_{i=1}^m \alpha_i \left(1 - \xi_i - y_i(w^T x_i + b)\right) - \sum_{i=1}^m \mu_i \xi_i,\ \alpha_i \geq 0, \mu_i \geq 0, 1 \leq i \leq m$$

3.3.2 令

$$-\frac{\partial L}{\partial w} = 0, -\frac{\partial L}{\partial b} = 0, -\frac{\partial L}{\partial \xi_i} = 0$$

得到：

$$\lambda w = \sum_{i=1}^{m} \alpha_i y_i x_i$$

$$\sum_{i=1}^{m} \alpha_i y_i = 0$$

$$\frac{1}{m} = \alpha_i + \mu_i$$

故对偶形式为：

$$\max_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2\lambda} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \left(x_i \cdot x_j\right),\ \text{s.t.}\ \sum_{i=1}^{m} \alpha_i y_i = 0, 0 \le \alpha_i \le \frac{1}{m}, 1 \le i \le m$$

3.3.3 原问题：

$$\min_{w,b,\xi} \frac{\lambda}{2} \|w\|^2 + \frac{1}{m} \sum_{i=1}^{m} \xi_i,\ \text{s.t.}\ y_i(w^T \Phi(x_i) + b) \ge 1 - \xi_i,\ \xi_i \ge 0, 1 \le i \le m$$

对偶问题：

$$\max_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2\lambda} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j k\left(x_i, x_j\right),\ \text{s.t.}\ \sum_{i=1}^{m} \alpha_i y_i = 0, 0 \le \alpha_i \le \frac{1}{m}, 1 \le i \le m$$

3.3.4 损失函数为

$$J_i(w) = \frac{\lambda}{2} \|w\|^2 + \max\{0, 1 - y(w^T x_i + b)\} = \frac{\lambda}{2} \|w\|^2 + \text{hinge}(x_i)$$

其中 $\text{hinge}(x)$ 的次梯度已经由 3.1.2 给出，故梯度 g 为：

$$g = \frac{\partial J_i(w)}{\partial w} = \lambda w + \begin{cases} 0 & y_i(w^T x_i + b) \ge 1 \\ -y_i x_i & y_i(w^T x_i + b) < 1 \end{cases} = \begin{cases} \lambda w & y_i(w^T x_i + b) \ge 1 \\ \lambda w - y_i x_i & y_i(w^T x_i + b) < 1 \end{cases}$$

3.3.5 设 $J_j(w_t)$ 的次梯度为 g_{tj} 。在代码中，当 $y_j(w_t^T x_j + b) < 1$ 时，

$$\begin{aligned} w_{t+1} &= (1 - \eta_t \lambda) w_t + \eta_t y_j x_j \\ &= w_t - \eta_t(\lambda w_t - y_j x_j) \\ &= w_t - \eta \cdot g_{tj} \end{aligned}$$

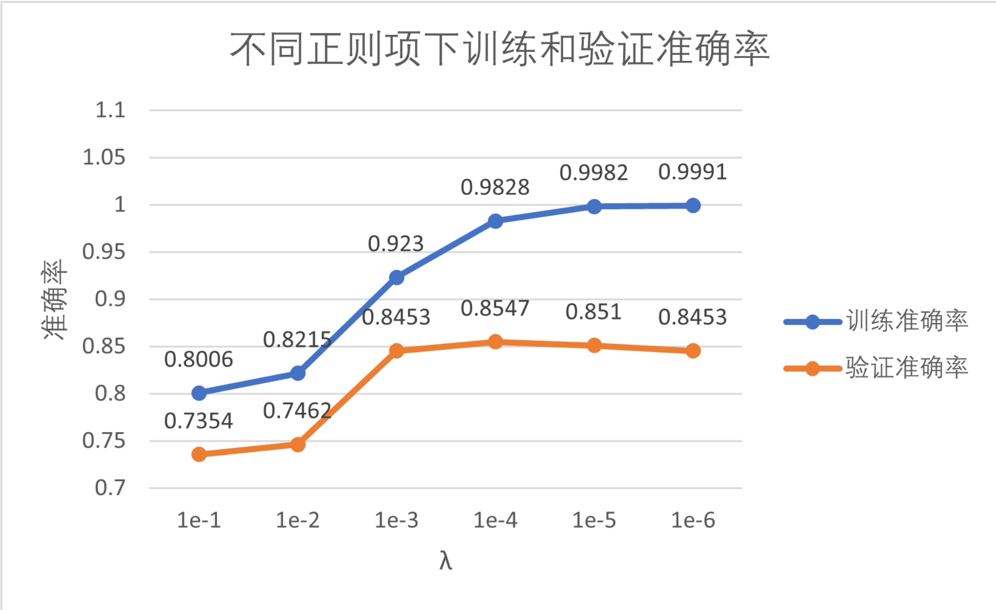
当 $y_j(w_t^T x_j + b) \ge 1$ 时：

$$\begin{aligned} w_{t+1} &= (1 - \eta_t \lambda) w_t \\ &= w_t - \eta_t \lambda w_t \\ &= w_t - \eta \cdot g_{tj} \end{aligned}$$

故该代码的实现就是利用 $J(w)$ 次梯度作为梯度值、步长为 $1/(\lambda t)$ 的 SGD 算法。

3.4.3 训练次数为 200 轮。

在步长策略为 $1/(\lambda t)$ 、Batch Size 为 512 的条件下，不同正则化参数值的影响：



在 Batch Size 为 512 、正则化参数值为 1e-4 的条件下，不同步长衰减策略的影响：

λ	**训练准确率**	**验证准确率**
$1/(\lambda t)$	0.9828	0.8547
固定步长衰减	0.9681	0.8619
指数衰减 $0.95^t\alpha_0$	0.9844	0.8713

3.4.4 测试集上的准确率：0.8539。不能提高准确率。核函数的本质是将数据映射到高维空间以使其线性可分，而当前训练集的数据经过词袋法向量化，各个单词已经映射到很高的维度上，并且数据分布十分稀疏，因此核函数无法起到效果。

3.4.5 验证集准确率：0.8713 。F1-Score：0.8741 。混淆矩阵：