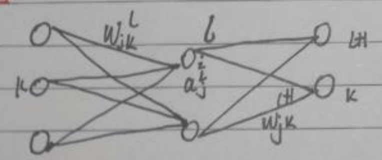
手推公式

2022年5月14日　　9:23

- 反向传播

• 手推反向传播



• 第l层第j个神经元产生的误差为 δ_j^l

$$\delta_j^l = \frac{\partial C}{\partial z_j^l} = \frac{\partial C}{\partial z_k^{l+1}} \cdot \frac{\partial z_k^{l+1}}{\partial a_j^l} \cdot \frac{\partial a_j^l}{\partial z_j^l}$$

$$= \delta_k^{l+1} \cdot \frac{\partial \sum_k w_{kj}^{l+1} a_j^l + b_k^{l+1}}{\partial a_j^l} \cdot \frac{\partial a_j^l}{\partial z_j^l}$$

$$= \delta_k^{l+1} \cdot \Sigma w_{kj}^{l+1} \cdot \frac{\sigma(z_j^l)}{\partial z_j^l}$$

$$= \delta_k^{l+1} \cdot \sum_k w_{kj}^{l+1} \cdot \sigma'(z_j^l)$$

$$= \sum_k \delta_k^{l+1} \cdot w_{kj}^{l+1} \cdot \sigma'(z_j^l)$$

• w_{jk}^l 的梯度

$$\frac{\partial C}{\partial w_{jk}^l} = \frac{\partial C}{\partial z_j^l} \cdot \frac{\partial z_j^l}{\partial w_{jk}^l} = \delta_j^l \cdot \frac{\partial \sum_k w_{jk}^l a_k^{l-1} + b_j^l}{\partial w_{jk}^l}$$

$$= \delta_j^l \cdot \sum_k a_k^{l-1} = \delta_j^l \cdot a_k^{l-1} \quad (\text{具体的}k)$$

NO.　/　/　Date.　/　/

• b_j^l 的梯度

$$\frac{\partial C}{\partial b_j^l} = \frac{\partial C}{\partial z_j^l} \cdot \frac{\partial z_j^l}{\partial b_j^l}$$

$$= \delta_j^l \cdot \frac{\sum_k w_{jk}^l a_k^{l-1} + b_j^l}{\partial b_j^l}$$

$$= \delta_j^l$$

- BN层计算

  - **BN层的计算过程**

    对于上一层的输出 $X=\{x_1, x_2, \dots x_m\}$，学习参数 γ, β

    均值： $\mu = \frac{1}{m}\sum_{i=1}^{m} x_i$

    方差： $\sigma = \frac{1}{m}\sum_{i=1}^{m}(x_i-\mu)^2$

    归一化： $\hat{x}_i = \frac{(x_i-\mu)}{\sqrt{\sigma^2+\epsilon}}$ ϵ为无穷小项

    重构： $\hat{y}_i = \gamma\hat{x}_i+\beta = BN_{\gamma,\beta}(x_i)$

- BN层反向传播

  - **BN层反向传播**

    - $\mu = \frac{1}{m}\sum_{i=1}^{m} x_i$ $\sigma = \frac{1}{m}\sum_{i=1}^{m}(x_i-\mu)^2$

      $\hat{x}_i = \frac{x_i-\mu}{\sqrt{\sigma+\epsilon}}$ $y_i = \gamma\hat{x}_i+\beta$

    - $\frac{\partial L}{\partial \gamma} = \sum_{i=1}^{m}\frac{\partial L}{\partial y_i}\cdot\hat{x}_i$

    - $\frac{\partial L}{\partial \beta} = \sum_{i=1}^{m}\frac{\partial L}{\partial y_i}$

    - $\frac{\partial L}{\partial \hat{x}_i} = \frac{\partial L}{\partial y_i}\cdot\frac{\partial y_i}{\partial \hat{x}_i} = \frac{\partial L}{\partial y_i}\cdot\gamma$

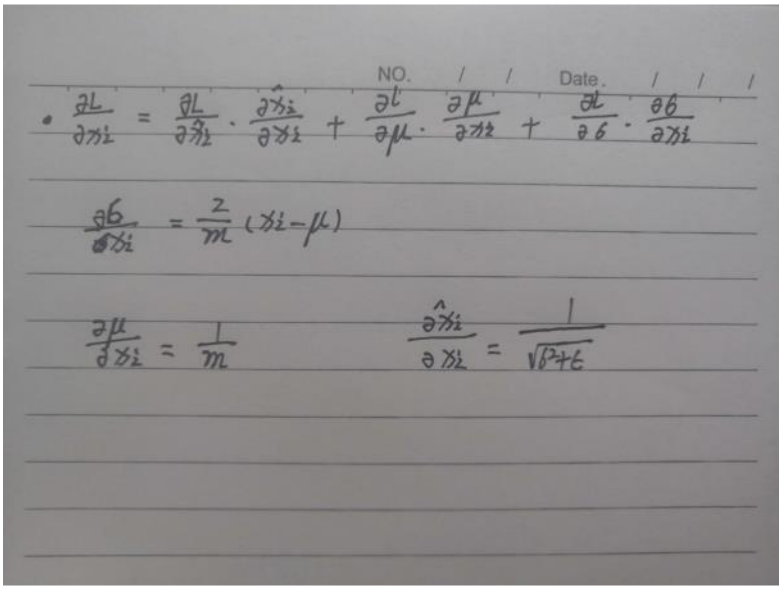
    - $\frac{\partial L}{\partial \sigma} = \sum_{i=1}^{m}\frac{\partial L}{\partial \hat{x}_i}\cdot\frac{\partial \hat{x}_i}{\partial \sigma}$

      $$\frac{\partial \hat{x}_i}{\partial \sigma} = \frac{-(x_i-\mu)[(\sigma^2+\epsilon)^{\frac{1}{2}}]'}{((\sigma^2+\epsilon)^{\frac{1}{2}})^2} = \frac{(\mu-x_i)\cdot\sigma(\sigma^2+\epsilon)^{-\frac{1}{2}}}{\sigma^2+\epsilon}$$

      $$= \sigma(\mu-x_i)(\sigma^2+\epsilon)^{-\frac{3}{2}}$$

    - $\frac{\partial L}{\partial \mu} = \sum_{i=1}^{m}\frac{\partial L}{\partial \hat{x}_i}\cdot\frac{\partial \hat{x}_i}{\partial \mu} + \frac{\partial L}{\partial \sigma}\cdot\frac{\partial \sigma}{\partial \mu}$

      $\frac{\partial \hat{x}_i}{\partial \mu} = \frac{-1}{\sqrt{\sigma^2+\epsilon}}$ $\frac{\partial \sigma}{\partial \mu} = -\frac{2}{m}\sum_{i=1}^{m}(x_i-\mu)$



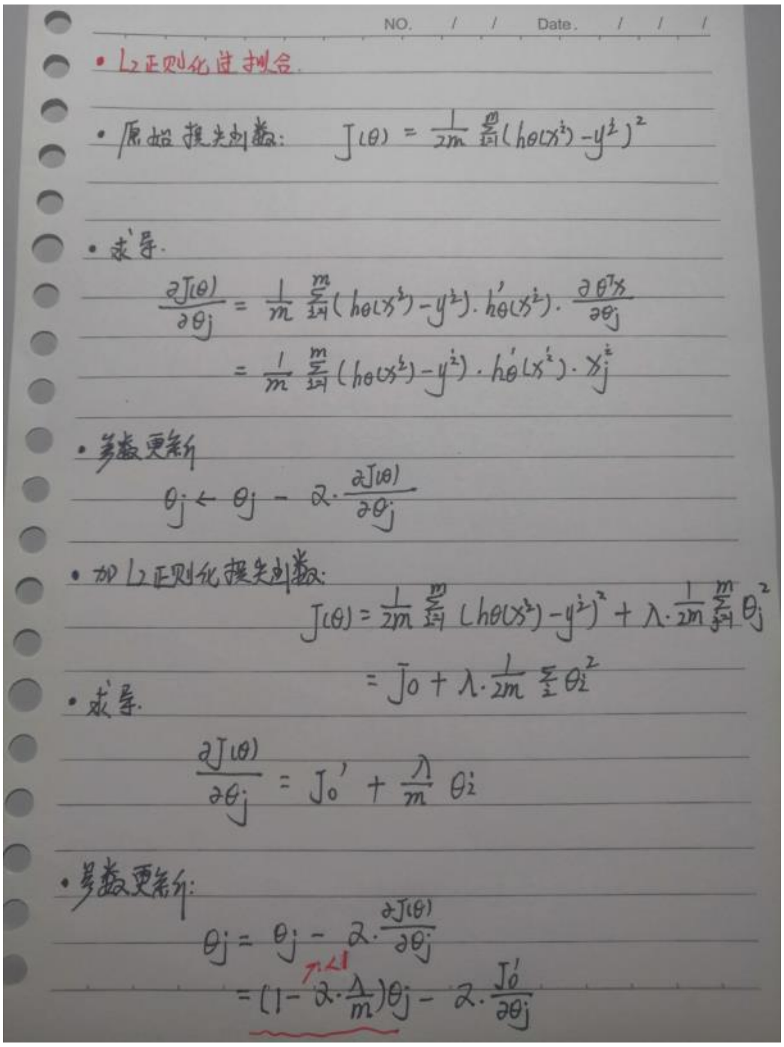
NO. / / Date. / / /

- $\frac{\partial L}{\partial x_i} = \frac{\partial L}{\partial \hat{x}_i}\cdot\frac{\partial \hat{x}_i}{\partial x_i} + \frac{\partial L}{\partial \mu}\cdot\frac{\partial \mu}{\partial x_i} + \frac{\partial L}{\partial \sigma}\cdot\frac{\partial \sigma}{\partial x_i}$

$$\frac{\partial \sigma}{\partial x_i} = \frac{2}{m}(x_i-\mu)$$

$$\frac{\partial \mu}{\partial x_i} = \frac{1}{m} \qquad \frac{\partial \hat{x}_i}{\partial x_i} = \frac{1}{\sqrt{\sigma^2+\epsilon}}$$

- L2正则化（防止过拟合）



NO. / / Date. / / /

- L2正则化过拟合
- 原始损失函数: $J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^i)-y^i)^2$
- 求导.

$$\frac{\partial J(\theta)}{\partial \theta_j} = \frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^i)-y^i)\cdot h_\theta'(x^i)\cdot\frac{\partial \theta^T x}{\partial \theta_j}$$
$$= \frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^i)-y^i)\cdot h_\theta'(x^i)\cdot x_j^i$$

- 参数更新

$$\theta_j \leftarrow \theta_j - \alpha\cdot\frac{\partial J(\theta)}{\partial \theta_j}$$

- 加L2正则化损失函数:

$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^i)-y^i)^2 + \lambda\cdot\frac{1}{2m}\sum_{j=1}^{m}\theta_j^2$$
$$= J_0 + \lambda\cdot\frac{1}{2m}\sum\theta_i^2$$

- 求导.

$$\frac{\partial J(\theta)}{\partial \theta_j} = J_0' + \frac{\lambda}{m}\theta_i$$

- 参数更新:

$$\theta_j = \theta_j - \alpha\cdot\frac{\partial J(\theta)}{\partial \theta_j}$$
$$= (1-\overset{<1}{\alpha\cdot\frac{\lambda}{m}})\theta_j - \alpha\cdot\frac{J_0'}{\partial \theta_j}$$

- L2正则化服从高斯分布

- L2正则化服从高斯分布

  • 高斯分布： $f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

  • 参数 $p(w)$ 若服从均值为0的高斯分布，则有：

  $$p(w_i) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{w_i^2}{2\sigma^2}}$$

  • 极大似然估计的对数形式：

  $$\begin{aligned}
  \log p(w) &= \log \prod_i \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{w_i^2}{2\sigma^2}} \\
  &= \sum_i \log \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{w_i^2}{2\sigma^2}} \\
  &= \sum_i \log e^{-\frac{w_i^2}{2\sigma^2}} + \sum_i \log \frac{1}{\sigma\sqrt{2\pi}} \\
  &= -\frac{1}{2\sigma^2} \sum_i w_i^2 + \sum_i \log \frac{1}{\sqrt{2\pi}\cdot\sigma} \\
  &= \underline{-\frac{1}{2\sigma^2} \sum_i w_i^2 + C'}
  \end{aligned}$$

- L1正则化服从拉普拉斯分布

NO. / / Date. / /

L1正则化服从拉普拉斯分布

• 拉普拉斯分布: $f(x) = \frac{1}{2b} e^{-\frac{|x-\mu|}{b}}$

• 参数 p(w) 服从均值为0，参数为b的拉普拉斯分布:

$$p(w) = \frac{1}{2b} e^{-\frac{|w|}{b}}$$

• 极大似然对数形式:

$$\log p(w) = \log \prod \frac{1}{2b} e^{-\frac{|w|}{b}}$$

$$= \sum \log \frac{1}{2b} e^{-\frac{|w|}{b}}$$

$$= \sum \log e^{-\frac{|w|}{b}} + \sum \log \frac{1}{2b}$$

$$= -\sum \frac{|w|}{b} + \sum \log \frac{1}{2b}$$

$$= -\frac{1}{b} \sum |w| + \underbrace{\sum \log \frac{1}{2b}}_{} \approx C'$$

● 逻辑回归（预测函数、损失函数、梯度更新）

- LR手推、损失函数、梯度更新

①. LR的本质仍然为线性回归，加入了sigmoid函数将连续值映射为0 or 1、

②. sigmoid函数：

$$g(z) = \frac{1}{1+e^{-z}} \qquad g'(z) = g(z)(1-g(z))$$

③ LR预测函数：

$$h_\theta(x) = g(\theta^T x)$$

④ 概率表达式：

$$p(y=1|x) = h_\theta(x)$$
$$p(y=0|x) = 1-h_\theta(x)$$
$$\Rightarrow \quad p(y|x) = h_\theta(x)^{y} \cdot (1-h_\theta(x))^{1-y}$$

⑤. 极大似然函数：

$$L(\theta) = \prod_{i=1}^{m} p(y=y^{(i)} | x^{(i)})$$
$$= \prod_{i=1}^{m} h_\theta(x^{(i)})^{y^{(i)}} \cdot (1-h_\theta(x^{(i)}))^{1-y^{(i)}}$$

⑥. log似然函数：

$$l(\theta) = \sum_{i=1}^{m} y^{(i)} h_\theta(x^{(i)}) + (1-y^{(i)}) \cdot (1-h_\theta(x^{(i)}))$$
$$= \sum y \log(h_\theta(x)) + (1-y)\log(1-h_\theta(x))$$

NO. / / Date.

- ③. 求导

$$\frac{\partial L(\theta)}{\partial \theta_j} = \frac{\partial L(\theta)}{\partial h_\theta(x)} \cdot \frac{\partial h_\theta(x)}{\partial \theta_j}$$

$$= \left(y \cdot \frac{1}{h_\theta(x)} - (1-y) \frac{1}{1-h_\theta(x)}\right) \cdot \frac{\partial h_\theta(x)}{\partial \theta_j}$$

$$= \left(y \cdot \frac{1}{g(\theta^T x)} - (1-y) \frac{1}{1-g(\theta^T x)}\right) \cdot \frac{\partial g(\theta^T x)}{\partial \theta_j}$$

$$= \left(y \cdot \frac{1}{g(\theta^T x)} - (1-y) \cdot \frac{1}{1-g(\theta^T x)}\right) \cdot g(\theta^T x)(1-g(\theta^T x)) \cdot \frac{\partial (\theta^T x)}{\partial \theta_j}$$

$$= \left(y(1-g(\theta^T x)) - (1-y)\, g(\theta^T x)\right) \cdot x_j$$

$$= (y - g(\theta^T x)) \cdot x_j$$

$$= (y - h_\theta(x)) \cdot x_j$$

④. 参数更新（极大似然 → 梯度上升）

$$\theta_j \leftarrow \theta_j + \alpha \cdot \frac{\partial L(\theta)}{\partial \theta_j} = \theta_j + \alpha \cdot (y^{(i)} - h_\theta(x^{(i)})) \cdot x_j^{(i)}$$

- SVM（原形式、对偶形式）

- 手推SVM，原形式，对偶形式.
- 超平面可表示为:

$$w \cdot x + b = 0$$

w：法向量

- 分类决策函数:

$$f(x) = \mathrm{sign}(w^* \cdot x + b^*)$$

- 函数间隔:（表示分类的正确性及确信度）

$$\hat{\gamma}_i = y_i(w \cdot x_i + b)$$

- 几何间隔:

$$\gamma_i = \frac{y_i(w \cdot x_i + b)}{\|w\|}$$

$\|w\|$ → 规则化因子

$$\gamma_i = \frac{\hat{\gamma}_i}{\|w\|}$$

- 最大化几何间隔:

$$\max_{w,b} \ \gamma = \frac{\hat{\gamma}}{\|w\|}$$

$$\text{s.t.} \quad \frac{y_i(w \cdot x_i + b)}{\|w\|} \geq \gamma \qquad i = 1, 2, \cdots N$$

γ：最小几何间隔.

$$\Downarrow$$

$$\min_{w,b} \ \frac{1}{2}\|w\|^2$$

$$\text{s.t.} \quad y_i(w \cdot x_i + b) - 1 \geq 0$$

（凸二次规划问题）

• 转化为对偶问题（拉格朗日函数）

原: $\min\limits_{w,b} \frac{1}{2}\|w\|^2$

s.t. $y_i(w x_i + b) - 1 \ge 0$

$\Downarrow$

$$L(w,b,\alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{N} \alpha_i y_i (w \cdot x_i + b) + \sum_{i=1}^{N} \alpha_i$$

• 转化为极大极小问题:

$$\max_{\alpha} \min_{w,b} L(w,b,\alpha)$$

• 求 $\min\limits_{w,b} L(w,b,\alpha)$ (偏导)

令 $\frac{\partial L(w,b,\alpha)}{\partial w} = 0$ 则:

$$\therefore w - \sum_{i=1}^{N} \alpha_i y_i x_i = 0 \Rightarrow w = \sum_{i=1}^{N} \alpha_i y_i x_i$$

令 $\frac{\partial L(w,b,\alpha)}{\partial b} = 0$ 则:

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

• 代回原 $L(w,b,\alpha)$:

$$L(w,b,\alpha)=\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j(x_i\cdot x_j)-\sum_{i=1}^{N}\alpha_i y_i\left(\sum_{j=1}^{N}\alpha_j y_j(x_i\cdot x_j)+b\right)+\sum_{i=1}^{N}\alpha_i$$

$$=-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j(x_i\cdot x_j)+\sum_{i=1}^{N}\alpha_i$$

• 求极大

$$\max_{\alpha} L(w,b,\alpha)=-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j(x_i\cdot x_j)+\sum_{i=1}^{N}\alpha_i$$

$$s.t.\quad \sum_{i=1}^{N}\alpha_i y_i=0$$

$$\alpha_i\geq 0\qquad i=1,2,\dots N$$

$$\Downarrow$$

变极小：

$$\min_{\alpha}\ \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j(x_i\cdot x_j)-\sum_{i=1}^{N}\alpha_i$$

$$s.t.\quad \sum_{i=1}^{N}\alpha_i y_i=0$$

$$\alpha_i\geq 0\qquad i=1,2,\dots N$$

• 求 w^*, b^* (KKT条件)

$$w^*=\sum_{i=1}^{N}\alpha_i^* y_i x_i$$

$$y=wx+b\ \rightarrow\ b^*=y_j-\sum_{i=1}^{N}\alpha_i^* y_i(x_i\cdot x_j)$$

• 分离超平面(最优)

$$\sum_{i=1}^{N}\alpha_i^* y_i(x_i\cdot x)+b=0$$

- • SVM核函数

  线性核函数: $k(x_i, x_j) = (x_i^T \cdot x_j)$

  多项式核函数: $K(x_i, x_j) = (x_i^T \cdot x_j + R)^d$

  高斯核函数: $k(x_i, x_j) = e^{\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)}$

  $$f(x) = sgn(w^* x + b^*) = sgn\left(\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i^* y_i k(x_i, x_j) + b^*\right)$$

  • SVM损失函数:

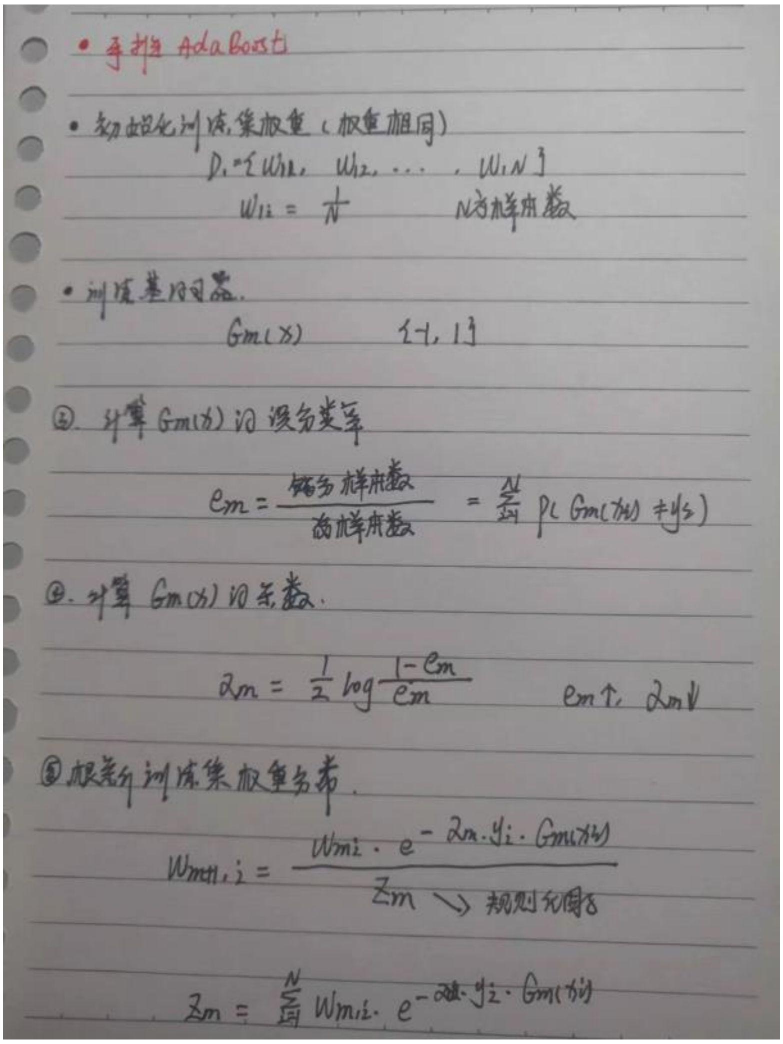
  $$L = \sum_{i=1}^{N} max(0, 1 - y_i(wx_i + b)) + \frac{1}{2}\|w\|^2$$

  • SVM硬间隔, 软间隔表达式.

  $$\min_{w,b} \frac{1}{2}\|w\|^2 \qquad s.t.\ y_i(wx_i + b) \geq 1$$

  $$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi_i \qquad s.t.\ y_i(wx_i + b) \geq 1 - \xi_i,\ \xi_i \geq 0$$

- AdaBoost

- 

• 手推 AdaBoost

• 初始化训练集权重（权重相同）

$D_1 = \{w_{11}, w_{12}, \dots, w_{1N}\}$

$w_{1i} = \frac{1}{N}$ N为样本数

• 训练基分类器

$G_m(x)$ $\{-1, 1\}$

③. 计算 $G_m(x)$ 的误分类率

$$e_m = \frac{\text{错分样本数}}{\text{总样本数}} = \sum_{i=1}^{N} P(G_m(x_i) \neq y_i)$$

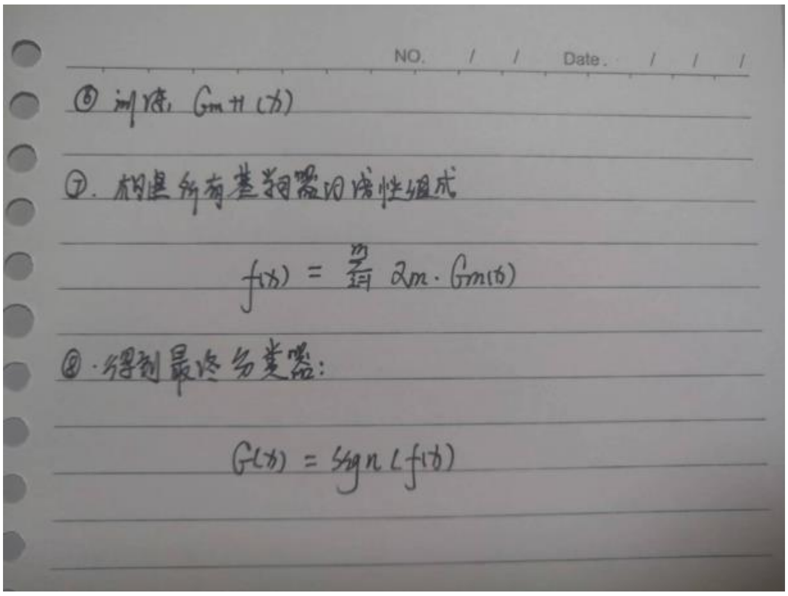
④. 计算 $G_m(x)$ 的系数

$$\alpha_m = \frac{1}{2}\log\frac{1-e_m}{e_m}$$ $e_m \uparrow, \alpha_m \downarrow$

⑤ 根据新训练集权重分布

$$w_{m+1,i} = \frac{w_{mi} \cdot e^{-\alpha_m \cdot y_i \cdot G_m(x_i)}}{Z_m}$$ 规则化因子

$$Z_m = \sum_{i=1}^{N} w_{mi} \cdot e^{-\alpha_m \cdot y_i \cdot G_m(x_i)}$$

- 

NO. / / Date. / / /

⑥ 训练 $G_{m+1}(x)$

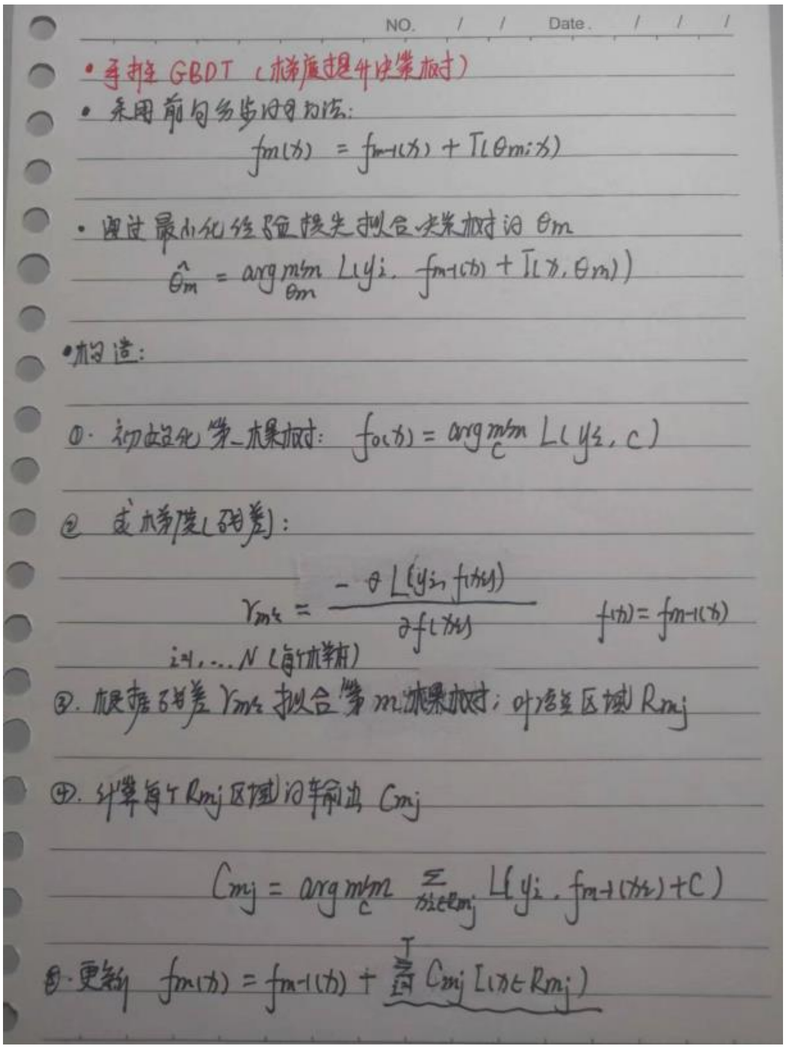
⑦. 构建所有基分类器的线性组成

$$f(x) = \sum_{i=1}^{m} \alpha_m \cdot G_m(x)$$

⑧. 得到最终分类器:

$$G(x) = sign(f(x))$$

- GBDT（梯度提升决策树）

- 

• 手推 GBDT（梯度提升决策树）

• 采用前向分步算法：

$$f_m(x) = f_{m-1}(x) + T(\theta_m; x)$$

• 通过最小化经验损失拟合决策树的 θ_m

$$\hat{\theta}_m = \arg\min_{\theta_m} L(y_i, f_{m-1}(x) + T(x, \theta_m))$$

•构造：

①. 初始化第一棵决策树：$f_0(x) = \arg\min_{c} L(y_i, c)$

②. 求梯度（残差）：

$$r_{mi} = \frac{-\partial L(y_i, f(x_i))}{\partial f(x_i)} \qquad f(x) = f_{m-1}(x)$$

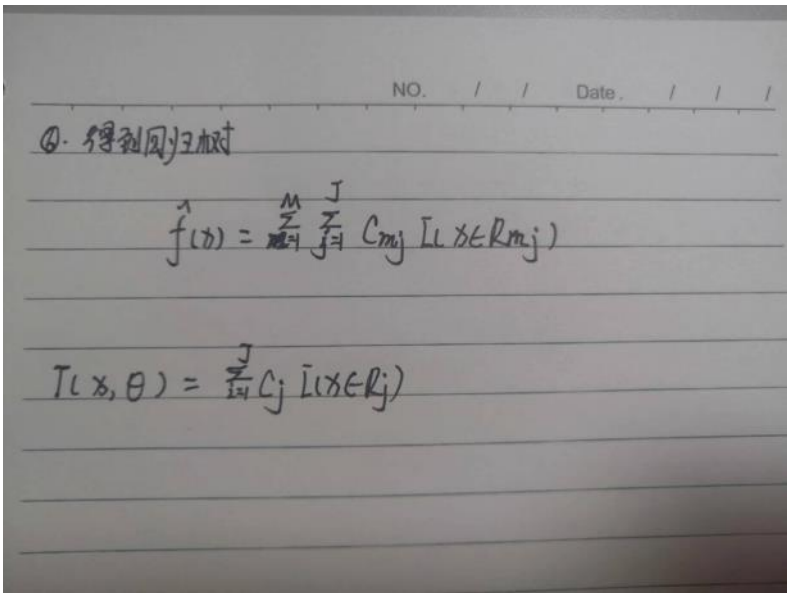
$i=1,\ldots,N$（每个样本）

③. 根据残差 r_{mi} 拟合第 m 棵决策树；叶子区域 R_{mj}

④. 计算每个 R_{mj} 区域的输出 C_{mj}

$$C_{mj} = \arg\min_{c} \sum_{x_i \in R_{mj}} L(y_i, f_{m-1}(x_i) + c)$$

⑤. 更新 $f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J} C_{mj} I(x \in R_{mj})$

- 

⑥. 得到回归树

$$\hat{f}(x) = \sum_{m=1}^{M} \sum_{j=1}^{J} C_{mj} I(x \in R_{mj})$$

$$T(x, \theta) = \sum_{j=1}^{J} C_j I(x \in R_j)$$

- 泰勒展开式

NO. / / Date. / / /

• 泰勒展开式.

$$f(x)=\frac{f(x_0)}{0!}+\frac{f'(x_0)}{1!}(x-x_0)+\frac{f''(x_0)}{2!}(x-x_0)^2+\cdots+\frac{f^{n}(x_0)}{n!}(x-x_0)^n$$

e^x 在 $x=0$ 处的泰勒展开:

$$e^x=1+\frac{e^0}{1!}(x-0)+\frac{e^0}{2!}(x-0)^2+\cdots+\frac{e^0}{n!}(x-0)^n$$

$$=1+x+\frac{x^2}{2!}+\frac{x^3}{3!}+\cdots+\frac{x^n}{n!}$$