7 优化算法 8采样

2020年5月31日 星期日　　上午9:11

机器学习 = 模型表征 + 模型评估 + 优化算法

1. 有监督的损失函数

1.1 二分类问题 $Y=\{-1,1\}$.

指示函数：$L_{0-1}(f,y) = 1_{fy \le 0}$

Hinge：$L_{hinge}(f,y) = \max\{0, 1-fy\}$

Logistic (Sigmoid)：$L_{logistic}(f,y) = \log_2(1+\exp(-fy))$

Cross Entropy：$L_{CE}(f,y) = -\log_2\left(\frac{1+fy}{2}\right)$

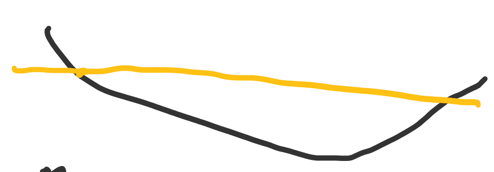
1.2 回归问题 平方损失函数

$L_{square}(f,y) = (f-y)^2$

2. 优化问题

Q：哪些是凸优化问题，哪些不是（机器学习中）

A：凸函数 $L(\lambda x + (1-\lambda) y) \le \lambda L(x) + (1-\lambda) L(y)$



逻辑回归是凸优化；主成分分析不是凸优化

3. 经典优化算法

Q：无约束问题的优化方法有哪些？

A：假设 $\min_{\theta} L(\theta)$

经典优化算法：直接法和迭代法

迭代法：一阶泰勒展开 — 梯度下降法

二阶泰勒展开 — 牛顿法

4. 梯度验证

5. 随机梯度下降法

$L(\theta) = E_{(x,y)\sim P_{data}} L(f(x,\theta), y)$

$\rightarrow \theta = \arg\min_{\theta} L(\theta)$

梯度下降：$L(\theta) = \frac{1}{m}\sum_{i=1}^{M} L(f(x_i,\theta), y_i)$　　$\theta_{i+1} = \theta - \alpha \cdot \nabla L(\theta)$

随机梯度下降：$L(\theta) = \frac{1}{m}\sum_{j=1}^{m} L(f(x_{i_j},\theta), y_{i_j})$　　$\nabla L(\theta) = \frac{1}{m}\sum_{j=1}^{m} \nabla L(f(x_{i_j},\theta), y_{i_j})$

取 m 一般 2 的幂次，先对所有数据进行随机排序

6. 随机梯度算法的加速

Q：SGD 失效原因

A：SGD 相对于大样本接受的步少，而且波动剧烈甚至不收敛，局部最优

解决办法：引入惯性，动量 (Momentum)

SGD：$\theta_{t+1} = \theta - \eta g_t$　$\Rightarrow$　$v_t = \gamma v_{t-1} + \eta g_t$，$\theta_{t+1} = \theta_t - v_t$　{速度更快，更收敛}

— AdaGrad：$\theta_{t+1,i} = \theta_{t,i} - \frac{\eta}{\sqrt{\sum_{k=0}^{t} g_{k,i}^2 + \epsilon}} g_{t,i}$　（环境感知）

— Adam：集合 adagrad 和 momentum 的优点.

$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t$

$v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2$

$\theta_{t+1} = \theta_t - \frac{\eta \cdot m_t}{\sqrt{v_t + \epsilon}}$

β_1, β_2 衰减系数

m_t 一阶矩 — 无偏估计 $E(g_t)$

v_t 二阶矩 — $E(g_t^2)$

7. 正则化和稀疏性

Q：L1 正则化使模型具有稀疏性的原因

A：

8. 采样：从特定的概率分布中抽取样本点的过程

均匀分布随机数