No.

Date

Recommendation Systems + Boosting

关于前面算法介绍，详见另一文件

(unconstrained!)* 啥意思？后面将会知道

Matrix Factorization (用于 Latent factor methods) (MF)

有许多方式可将一个 matrix 表示为多个 matrix 的乘法，比如 SVD

而 MF 就是一个套入了梯度下降式学习的方法：

define model → define 目标 func → optimize with SGD

High level idea: rating matrix (R) ⇒ $U \times V$

where: $U \to$ users, $V \to$ items

If $R \in \mathbb{R}^{m\times n}$, and R has rank $k \ll \min(m,n)$, then: Insight.

$\exists\, U \in \mathbb{R}^{m\times k}$ and $V \in \mathbb{R}^{n\times k}$, s.t., $R = UV^T$

若 R rank 为 l, $l>k$ 呢？则是 $\exists\, U_{(m\times k)}\ V_{(n\times k)}$, $R \approx UV^T$

这就是低秩分解的优点！(low rank matrix's rank k: $k \ll \min(m,n)$)

很幸运的是，User-Item 矩阵 R 被普遍认为低秩！那么如何找 U. V？

Unconstrained MF: Let $E = R - UV^T$

欲 E 接近零矩阵！故定义 Objective Function：

$$J(U,V) = \frac{1}{2}\|E\|_2^2 = \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n} E_{i,j}^2 = \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n}\left(R_{i,j} - U_{j,\cdot}\, V_{\cdot,i}^T\right)^2$$

问题在于，这便要 $R_{i,j}$ 均已知；但现实是：User 只给极少 item 打分！

即：有 set：$Z = \{(i,j) : R_{i,j} \text{ is known}\}$, $J = \frac{1}{2}\sum_{(i,j)\in Z}\left(R_{i,j} - U_{j,\cdot}\, V_{\cdot,i}^T\right)^2 + \frac{\lambda}{2}\left(\|U\|_2^2 + \|V\|_2^2\right)$

则：$\nabla_{U_{j,\cdot}} J_{i,j}(U,V) = -E_{i,j} V_{\cdot,i}^T + \lambda U_{j,\cdot}$

$\nabla_{V_{\cdot,i}^T} J_{i,j}(U,V) = -E_{i,j} U_j + \lambda V_{\cdot,i}$　正则化

$U_{j,\cdot} \leftarrow U_{j,\cdot} - \eta \nabla_{U_{j,\cdot}} J_{i,j}(U,V)$; $V_{\cdot,i}^T \leftarrow V_{\cdot,i}^T - \eta \nabla_{V_{\cdot,i}^T} J_{i,j}(U,V)$

上述便是用 SGD 优化！每一次，挑一个 $(i,j) \in Z$，以更新参数

No.

Date

Trick：若知 $U.V^T$ 中一个，则 solve 是十分简单的！回想 $J(\theta)=\frac{1}{2}\sum_{i=1}^{N}(y^{(i)}-\theta^T x^{(i)})^2$

则可考虑：初始化 $U.V^T$ 后，while not converged：

- Fix V^T and solve U
- Fix U and solve V^T

SVD for collaborative filtering

SVD：$A=U\Lambda V^T$，where Λ：diagonal，$U.V$：orthogonal

考虑：对于 User-item R matrix，若已知：

$$R=Q\Sigma P^T$$

then truncate each of $Q.\Sigma$ and P s.t.：$Q.P$ has k columns.

Σ has $k\times k$ entries：$R\approx Q_k\Sigma_k P_k^T$

Then let $U\triangleq Q_k\Sigma_k$，$V\triangleq P_k$

$\Rightarrow U,V=\underset{U,V}{\operatorname{argmin}}\frac{1}{2}\|R-UV^T\|_2^2$

Theorem：R 是 fully observed 时，SVD 优化出的 U.V 与 MF 的，是一样的

Non-negative MF：

通常公司收集的 R 矩阵中是无负数的！（如：打星：1~5 颗，无负！）

故问题变成了：$U,V=\underset{U,V}{\operatorname{argmin}}\frac{1}{2}\|R-UV^T\|_2^2$

s.t. $U_{ij}\geq 0$，$V_{ij}\geq 0$

与先前介绍的 Unconstrained MF 不同，此处有限制！

Ensemble method：集成学习

Weighted Majority Algorithm：（example of ensemble method）

Given：pool of binary classifier

Goal：design a new learner that uses predictions of pool to make new prediction

则加权 majority vote 流程如下：

Campus

初始化 classifier weights: $\alpha_t = 1$, $\forall t \in \{1, \cdots, T\}$

对于每个样本 $(\vec{x}, y)$, do:

$\hat{h}(x) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$.

if $\hat{h}(x) \neq y$:

for each classifier $h_t \in \{1, \cdots, T\}$, do:

if $h_t(x) \neq y$, $\alpha_t \leftarrow \beta \alpha_t$. hyperparameter

根据个体学习器生成方式，目前 ensemble learning 可分两类：

① 个体学习器间存在依赖关系，必须串行生成 ⇒ boosting 家族

② ······ 不存在强依赖关系，也可并行化式 ⇒ Random forest

Adaboost: ⚠ 只作二分类 基模型个数

流程: Dataset $D = \{(x_1, y_1), \cdots, (x_m, y_m)\}$, 基学习算法 $\mathcal{L}$

$D_1(x) = 1/m$ $x = [\vec{x}_1, \vec{x}_2, \cdots, \vec{x}_m]$

for $t = 1, \cdots, T$. do: $D_t(x)$: 样本权值分布

(寻找当前误差最小基分类器为 h_t) $h_t = \mathcal{L}(D_t, D)$ 用 D_t 分布从 D 训练一个弱分类器 h_t.

$\epsilon_t = \underline{P_{x \sim D_t}(h_t(x) \neq f(x))}$ → h_t 误差估计

(if $\epsilon_t > 0.5$, break) (若>0.5, 则比随机还差!)

$\alpha_t = \frac{1}{2} \ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right)$ 规范化，确保 $D_{t+1}(x)$ 也是分布

$$D_{t+1}(x) = \frac{D_t(x)}{Z_t} \times \begin{cases} \exp(-\alpha_t), & \text{if } h_t(x) = f(x) \\ \exp(\alpha_t), & \text{if } \neq \end{cases}$$

$$= \frac{D_t(x)}{Z_t} \exp(-\alpha_t h_t(x) f(x))$$

end for

输出: $F(x) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$

*: 用 D_t 分布，可重采样出一个新数据集，在这个数据集上，优化 h_t 参数 to minimize loss. (题目中，此步可能不进行)

No.

Date

理解："加性模型"．$H(x)=\sum_{t=1}^{T}\alpha_t h_t(x)$

to minimize 指数损失函数：$\ell_{exp}(H|D)=\mathbb{E}_{x\sim D}[e^{-f(x)H(x)}]$

推导：$\frac{\partial \ell_{exp}(H|D)}{\partial H(x)}=-e^{-H(x)}P(f(x)=1|x)+e^{H(x)}P(f(x)=-1|x)$.
$=0$

Δ: $\frac{\partial e^{-H(x)}}{\partial H(x)}=-e^{-H(x)}$ $\frac{\partial e^{H(x)}}{\partial H(x)}=e^{H(x)}$.

$\ell_{exp}(H|D)=\sum_{i=1}^{|D|}D(x_i)(e^{-H(x_i)}\mathbb{I}(f(x_i)=1))+D(x_i)(e^{H(x_i)}\mathbb{I}(f(x_i)=-1))$

$=\sum_{i=1}^{|D|}(e^{-H(x_i)}P(f(x_i)=1|x_i)+e^{H(x_i)}P(f(x_i)=-1|x_i))$

$\Rightarrow H(x)=\frac{1}{2}\ln\frac{P(f(x)=1|x)}{P(f(x)=-1|x)}$，则 $\text{sign}(H(x))=\begin{cases}1, & P(f(x)=1|x)>P(f(x)=-1|x)\\ -1, & \text{otherwise}\end{cases}$

$=\arg\max_{y\in\{-1,1\}}P(f(x)=y|x)$.

补：$Z_t=\sum_{i=1}^{n}D_t(i)\exp(-\alpha_t y_i h_t(x_i))$

$=e^{\alpha_t}\sum_{i=1}^{n}D_t(i)\mathbb{I}(y_i\neq h_t(x_i))+e^{-\alpha_t}\sum_{i=1}^{n}D_t(i)\mathbb{I}(y_i=h_t(x_i))$

$=e^{\alpha_t}\epsilon_t+e^{-\alpha_t}(1-\epsilon_t)$

训后：最终：

$\epsilon=\frac{1}{n}\sum_{i=1}^{n}\begin{cases}1, & \text{if } y_i(\sum_{t=1}^{T}\alpha_t h_t(x_i))\le 0\\ 0, & \text{else}\end{cases}\le\frac{1}{n}\sum_{i=1}^{n}\exp(-y_i(\sum_{t=1}^{T}\alpha_t h_t))$

而 $D_{T+1}(i)=\frac{D_T(i)}{Z_T}\exp(-\alpha_T y_i h_T(x_i))$. 左右连乘：

$\epsilon\le\frac{1}{n}\sum_{i=1}^{n}\exp(-y_i(\sum_{t=1}^{T}\alpha_t h_t(x_i)))=\prod_{t=1}^{T}Z_t\cdot\underbrace{\sum_{i=1}^{n}D_{T+1}(i)}_{=1}$

$\therefore\epsilon\le\prod_{t=1}^{T}Z_t$, Upper bound $\downarrow$，又 $Z_t=e^{\alpha_t}\epsilon_t+e^{-\alpha_t}(1-\epsilon_t)$

$\therefore\frac{\partial Z_t}{\partial\alpha_t}=0$ $(\frac{\partial^2 Z_t}{\partial\alpha_t^2}>0$, convex$)\Rightarrow\boxed{\alpha_t=\frac{1}{2}\log\frac{1-\epsilon_t}{\epsilon_t}}$

再代回 Z_t：$Z_t=2\sqrt{\epsilon_t(1-\epsilon_t)}$

Campus