6-4-cn

1 本节我们学习推荐系统的第二部分，关于矩阵分解算法

2 数据处理系统提供大数据计算处理能力和应用开发平台。从计算架构的角度，将数据处理系统分为数据算法层、计算模型层、计算平台层、计算引擎层等。

与大数据相关的计算算法包括机器学习算法和数据挖掘算法。计算模型是指不同类型的大数据在不同场景下的处理方式，包括批处理、流计算、结构化数据

的大规模并发处理(MPP)模型、内存计算模型和数据流图模型。

计算平台和引擎**提供各种开发套件和操作环境**，我们选取 Spark Mllib 和 TensorFlow 为例

数据应用系统我们以推荐系统和社交网络分析为例进行讲解，本节我们继续讨论推荐系统典型应用的第 2 部分

3 推荐系统是信息过滤系统的一个子类，它试图预测用户对某项物品的评级或偏好。

简单地说，就是向用户推荐相关物品的算法。推荐系统被广泛应用于许多领域，如 Netflix，亚马逊，京东，淘宝，QQ 音乐等。

推荐算法大致可以分为三类。协同过滤，基于内容和基于知识的过滤算法。而协同过滤算法和基于内容的算法最大的区别在于，基于内容的算法依赖于物品本身的特征，而协同过滤依赖于其他用户对同一物品的反应。

协同过滤可以进一步分为基于邻域和基于模型，基于邻域的算法包括基于用户和基于物品的算法，我们将在这一节讨论。基于模型的算法包括隐语义模型、图模型等。在隐语义模型中，我将以矩阵分解为例进行说明。

基于内容的推荐算法基于项目特征对项目进行推荐，包括结构化特征和非结构化特征。

第三类是基于知识的算法。

4 让我们看一个视频“推荐系统如何工作(netflix 亚马逊)”

5 **隐语义模型 LFM(Latent factor model)**

从视频中我们知道评级矩阵中隐藏着一些模式，

我们想找出一些物品可能具有的特征。

将评价矩阵分解为物品-角色评价和用户-角色评价。

另外，我们并不想知道哪些特征的含义，这是抽象模型，我们只假定特征的数量。

6 **奇异值分解(SVD-Singular Value Decomposition)**

在线性代数中，奇异值分解(SVD)是实矩阵或复矩阵的因式分解。

它将具有标准正交特征基的方阵的特征分解推广到任意 m*n 矩阵。

它与极性分解有关。

具体来说，m*n 复矩阵 M 的奇异值分解是 U 的因式分解，

其中 UΣV*，其中 U 是 m*m 复酉矩阵，

Σ 是一个 m*n 的矩形对角线矩阵对角线上有非负实数，

V 是一个 n*n 的复酉矩阵。

如果 M 是实数，U 和 V 也可以保证是实数正交矩阵。

在这种情况下，SVD 通常表示为 UΣ(V 转置)

7 **矩阵分解**

奇异值分解需要稠密矩阵，即矩阵没有缺失值。

显然，用户物品评价矩阵有很多缺失的值。

所以用矩阵分解代替奇异值分解。

使用类似 SVD 的矩阵分解方法。

8 将矩阵分解为两个矩阵，即R=P乘以Q，其中R为$m\times n$用户-物品打分矩阵，P为$m\times k$用户-隐语义 lf (Latent factor)矩阵，Q为$k\times n$物品- 隐语义 lf (Latent factor)矩阵。

对于 u-user 和 i-item，它们的评分如图所示：

如果得到两个稠密矩阵P和Q，则从R=P乘以Q可以预测R中缺失的值。

那么如何计算P和Q呢?

9 计算 P 和 Q

定义如图公式所示的损失函数，并利用该损失函数来评价 P 和 Q 的较优选择

我们只利用用户已经给出的打分值计算损失函数。

损失函数的第一部分是预测额定值 Rui 尖尖 hat 和真实值 Rui 的最小方差。

代价函数的第二部分是正则值，防止过拟合。

10 最小化损失函数

我们进行迭代以使损失函数最小化。两种最小化损失函数的方法:

1 .用 ALS(交替最小二乘)最小化损失函数:

即固定 P，计算 Q 使损失函数 c 最小；然后，固定 Q，计算 P 使损失函数 c 最小;

直到达到最大迭代或 c 满足阈值条件。

求 C 对 Pu 的偏导数，使公式为 0，得到Q。

类似地获取P。

2 利用梯度下降最小化代价函数，即计算 C 对 Pu 的偏导数

以及 C 对 Qi 的偏导数。

然后使用幻灯片中的公式进行迭代(其中α为步长)

每次迭代结束后，更新 Pu 和 Qi，直到达到最大迭代或 c 满足阈值条件。

11 动手实验

为了更好的了解推荐系统，设计了一系列实验，

它包括基于用户的过滤推荐和矩阵分解。

在基于用户的过滤推荐中，包括预处理和协同过滤。

在 1.1 预处理中，包括加载数据并关联两个原表，创建一个新的 data.csv 文件，并通过删除重复记录来生成字典。

在 1.2 协同过滤中，首先计算用户相似度，然后列出与当前用户相似度前 10 位的用户，并进行推荐。

矩阵分解实验需要导入库 surprise，使用的数据集包括 10 万用户对电影的评分。

相关的训练模型包括 Funk 或 Bias SVD, Grid Search。

目标是在最佳模型上进行训练和测试，得到 SVD 的最佳参数，

这个过程将是

1.) 导入库

2 ）.导入数据

3）.Grid 搜索 SVD 训练

4.) 利用网格搜索得到的最佳参数进行训练和预测

5.) 最后将结果可视化。

平台上提供了所有的实验材料，包括手册和代码，可以帮助大家完成实验。

12 本节我们学习了推荐系统中的矩阵分解算法，今天内容就到这里，谢谢大家