day 1

目标：

- 协同过滤算法：基于用户的协同过滤-UserCF，基于商品的协同过滤-ItemCF
- 矩阵分解算法：隐向量，加强模型处理稀疏矩阵的能力，-->为后续Embedding做基础
- FM算法（Factorization Machines）：LR的应用改进，加上了特征交叉项
- Wide & Deep
- GBDT + LR：使用树模做特征交叉

一、推荐系统简介

1. what

对用户：帮助用户快速发现有用信息

对公司：增加公司产品与用户接粗，购买等行为概率

2. why

对用户：不需要用户明确的需求对信息进行过滤，利用用户各类历史信息做出推测

对公司：帮助产品最大限度的吸引用户，留存用户，增长用户黏性，提高用户转化率，帮助公司商业目标增长

3. who

考虑目标用户和对应的目标公司

二、常用评测指标

1. 用户满意度

线上行为统计（或用户调查）--购买率，点击率，停留时间，转化率

2. 预测准确度

离线评价指标

a. 评分预测

- 评分预测模型，通过用户历史物品评分建模，从而预测用户对未见过商品评分（计算：均方根误差RMSE，平均绝对误差MAE------对测试集中

$$用户u，和物品i$$

$$r_{ui}为用户u对商品i的实际评分$$

$$\hat{r_{ui}}为推荐模型预测的评分$$

$$RMSE = \sqrt{\frac{\sum_{u,i\in T}(r_u i - \hat{r}_{ui})^2}{|T|}}$$

$$MAE = \frac{\sum_{u,i\in T}|r_{ui} - \hat{r}_{ui}|}{|T|}$$

RMSE 中的平方项，对用户真实评分与推荐系统预测评分相差较大的用户加大惩罚，对系统要求更严格

b. TopN推荐

给用户一个列表的推荐物品，预测准确率指标一般是精确率（precision）和召回率（recall）

$$R(u)：推荐模型得到的推荐列表$$

$$T(u)：用户在实际场景中的行为列表（测试集）$$

精确率(precision)：分类正确的正样本占分类器判定为正样本的比例

$$precision = \sqrt{\frac{\sum_{u\in U}|R(u)\cap T(u)|}{\sum_{u\in U}|R(u)|}}$$

召回率(recall)：分类正确的正样本占真正样本的比例

$$Recall = \frac{\sum_{u\in U}|R(u)\cap T(u)|}{\sum_{u\in U}|T(u)|}$$

3. 覆盖率

推荐系统对物品长尾的发掘能力

简单的定义：推荐系统推荐商品集合占总物品集合的比例

具体来说：对相同的覆盖率，不同的物品数量分布、物品的流行度分布不同

为描述推荐系统挖掘长尾能力：统计不同物品出现次数的分布

挖掘长尾能力好：若所有物品都出现在推荐列表中，且次数差不多（研究物品在推荐列表中出现的次数分布），常用信息熵和基尼系数来定义

信息熵：

$$p(i)是物品i的流行度除以所有物品流行度之和$$

$$H=-\sum_{i=1}^{n}p(i)logp(i)$$

基尼系数(详情推导点击)

$$i_j是按照物品流行度p从小到大排序的物品列表中的第j个物品$$

$$G=\frac{1}{n-1}\sum_{j=1}^{n}(2j-n-1)p(i_j)$$

4. 多样性

尽可能包含用户的兴趣，目标：增加用户找到兴趣物品的概率，度量推荐列表中商品的多样性 == 度量列表中物品间的不相似性，通过不同的相似性函数度量推荐列表中的相似性，基于内容的相似，基于协同过滤的相似，得到不同角度的多样性

$$s(i,j)为物品i和物品j相似性$$

用户推荐列表的多样性

$$Diversity(R(u))=1-\frac{\sum_{i,j\in R(u)}s(i,j)}{\frac{1}{2}|R(u)|(|R(u)|-1)}$$

推荐系统整体多样性，所有用户推荐列表多样性的平均值

$$Diversity=\frac{1}{U}\sum_{u\in U}Diversity(R(u))$$

5. 新颖性

流行度越低，用户可能觉得新颖

6. AUC曲线

AUC(Area Under Curve), ROC(Receiver Operating Characteristic Curve)曲线下与坐标轴围成的面积

（混淆矩阵，召回率，准确率，ROC曲线）

> TP 真的真了
>
> FN 真的假了

> FP 假的真了
>
> TN 假的假了

$$Recall = \frac{TP}{TP+FN} \ ; Precise = \frac{TP}{TP+FP}$$

ROC：横坐标为假阳性率（FP），纵坐标为真阳性率（TPR）

三、召回

1 召回层在推荐系统架构中的位置及作用

基于工程考虑，将核心算法层分为召回层和排序层

召回层(少量特征，简单模型)：负责将海量候选集快速缩小到几千到几万的规模；候选集合大，计算速度快，模型简单，特征小，快速召回用户感兴趣的物品----要权衡计算速度和召回率，主流方法：采用多个简单策略叠加的多路召回策略

排序层(更多特征，复杂模型)：对缩小后的候选集进行精准的排序；目标得到精准的排序结果，处理物品数量少，有较多可利用特征，使用复杂模型

2. 多路召回策略

采用不同的策略、特征、简单模型，分别召回一部分候选集，混合后供后续排序模型使用，权衡计算速度和召回率，通过多线程并发进行，召回层与业务强相关（e.g. 兴趣标签，兴趣topic，兴趣实体，协同过滤，热门）（对视频推荐：热门视频，导演召回，演员召回，最近上映，流行趋势，类型召回）

对k的选取需要实验（离线评估，A/B测试）来考虑决定，不同任务具体策略的选择是人工基于经验，选择的策略间信息割裂

Embedding 召回是一个综合性强切计算速度也能满足需求的召回方法

3. Embedding召回

a. 什么是Embedding

一种思想，目的：将稀疏的向量（one-hot编码）转化为稠密的向量，即对one-hot做了平滑，相当于对Embedding做了max pooling

b. 常用的Embedding技术

主要分为三类：

- text Embedding(最常用)
- image Embedding
- graph Embedding

文本特征直接可以使用 text Embedding，对于非文本的 id 类特征，先将其转化为 id 序列，在使用 text embedding 的技术获取 id 的 Embedding 再做召回

对 text Embedding 常用技术

- 静态向量：word2vec，fasttext，glove
- 动态向量：ELMO，GPT，BERT

image Embedding针对有图或视频的特征，大部分是卷积通过各种连接技巧搭建的高效模型，使用现有的预训练模型提取图像或视频的向量特征，然后用于召回

graph embedding 对社交网络相关的推荐，用户、商品间存在复杂的图解钩关系，经典模型有Deep Walk，Node2Vec，LINE，EGES graph Embedding（alibaba 2018）

---------- 课后思考 ----------

四、AUC 的价值

1. 优势

不关注具体得分，只关注排序，适合做推荐排序的评估

2. 理论最高AUC（Max AUC）

存在GC（God Classifier）必须犯的错误，这时Max AUC < 1

主要因素：样本的不确定性，特征值完全相同的样本，对应标签不一定相同，这个不确定性的程度决定了Max AUC

最低值为 0.5

3. 贝叶斯错误率（BER）

'必须犯的错误' - irreducible error 对应数据中的不可约错误

BER：任意一个分类器在数据集上能取得的最低的错误率（Bayes Error Rate 贝叶斯错误率）

五、如何使用 Embedding 做召回

1. 什么是 Embedding

稠密向量的表达形式：Embedding相当于oneHot做了平滑，oneHot相当于对embedding做了max pooling

一般是神经网络倒数第二层的参数权重，只具有整体意义和相对意义

2. embedding 发展

word2vec之前：

MF矩阵分解 -- 》 word2vec

item2vec，wide and weep，youtube之后，embedding也用于特征工程，画像构建召回排序

意义：

> embedding表示：吧自然语言转化为数字，用于自然语言计算
>
> 替代oneHot，降低了特征的维度
>
> 替代协同矩阵，减低计算复杂度

3. item embedding

item的向量化：文本和图片向量化的过程

动态词向量相对于静态词向量，更充分利用了上下文信息，解决一词多意的问题

4. img embedding

相关技术

> resnet：图片向量
>
> image caption：图片的中文描述
>
> facenet：识别明星
>
> OCR：漫画文字

Embedding的实际意义：

> 图像：集合元素，简单图形，复杂图形
>
> 文本：字特征，句法特征，语义特征

5. user embedding

使新闻和用户可以在相同的向量空间下做文算，从用户画像中筛选出对排序模型重要的特征做向量化（如，标签tag，媒体号mid，一级分类cat1，二级分类cat2，主题topic等特征对用户是否点击某篇文章影响最大）中期使用了更多特征，模型采用了DSSM（确保user和item在同一向量空间

tag emb --> dssm emb --> bert + lstm(对用户的行为序列进行建模

6. embedding 召回

得到item，user向量后，基于向量的召回，大多数是：单embedding

单Embedding向量召回性价比高，多Embedding向量召回存储是瓶颈

> TDM：深度树匹配
>
> SDM，NIRSA：长期和短期兴趣建模
>
> EGES，Grapg-sage，GAT：graph Embedding
>
> MIND，CrossTag：多Embedding向量召回
>
> YouTube，DSSM：单Embedding向量召回

7. u2i 召回算法初步

u2i 召回算法：uese2vec,word2vec 个性化，crosstag，DSSM 个性化等召回算法；

> user2vec ：用户的 tag Embedding和文章的 tag Embedding求相似度，来做召回
>
> DSSM 个性化是：拿用户的 DSSM Embedding和文章的 DSSM Embedding求相似度，来做召回
>
> crosstag ：相当于多个 user2vec，需要把用户的 tag 按类别进行统计，每个类取 K 个 tag，共获取 m 组 tag，然后各组分别做 user2vec，最后汇总得到用户的推荐列表。

8. u2i 召回算法进阶

uese2vec 在做召回的初级阶段，做的一些朴素的尝试，该算法虽然简单暴力见效快，存储压力大。每个 user 都存储一个推荐列表，在产品初期 DAU 不多时，矛盾还不明显，但随着 DAU 不断提升，存储严重，可行的解决策略有两条

a. 把离线提前计算再存储转为线上即时计算不存储

b. 把按人推荐转化为分群推荐

9. 特征Embedding化

离散、连续、多值Embedding：某种程度上，对网络结构优化，即对Embedding的运算优化

10. Embedding缺点

长尾、稳定性、扩展性、多模态、分布（Embedding空间分布，影响模型泛化误差）

11. Embedding前言

Embedding表示优化

a. 簇内共享聚类中心，高度相似

b.分解与组合

Embedding结构优化

a. 为高频有效特征分配更多位置

b. 不重要特征共享位置

12. Embedding总结

> 表示：MF -- Seq Embedding -- Grap Embedding
>
> 特征抽取：DNN -- CNN -- RNN -- Transformer -- Bert
>
> 融合：拼接 -- 线性组合 -- Attention
>
> 交叉：Bit-wise -- Element-wise -- Vector-wise -- 高阶
>
> 优化：交叉方式优化 -- 表示方式优化 -- 结构优化