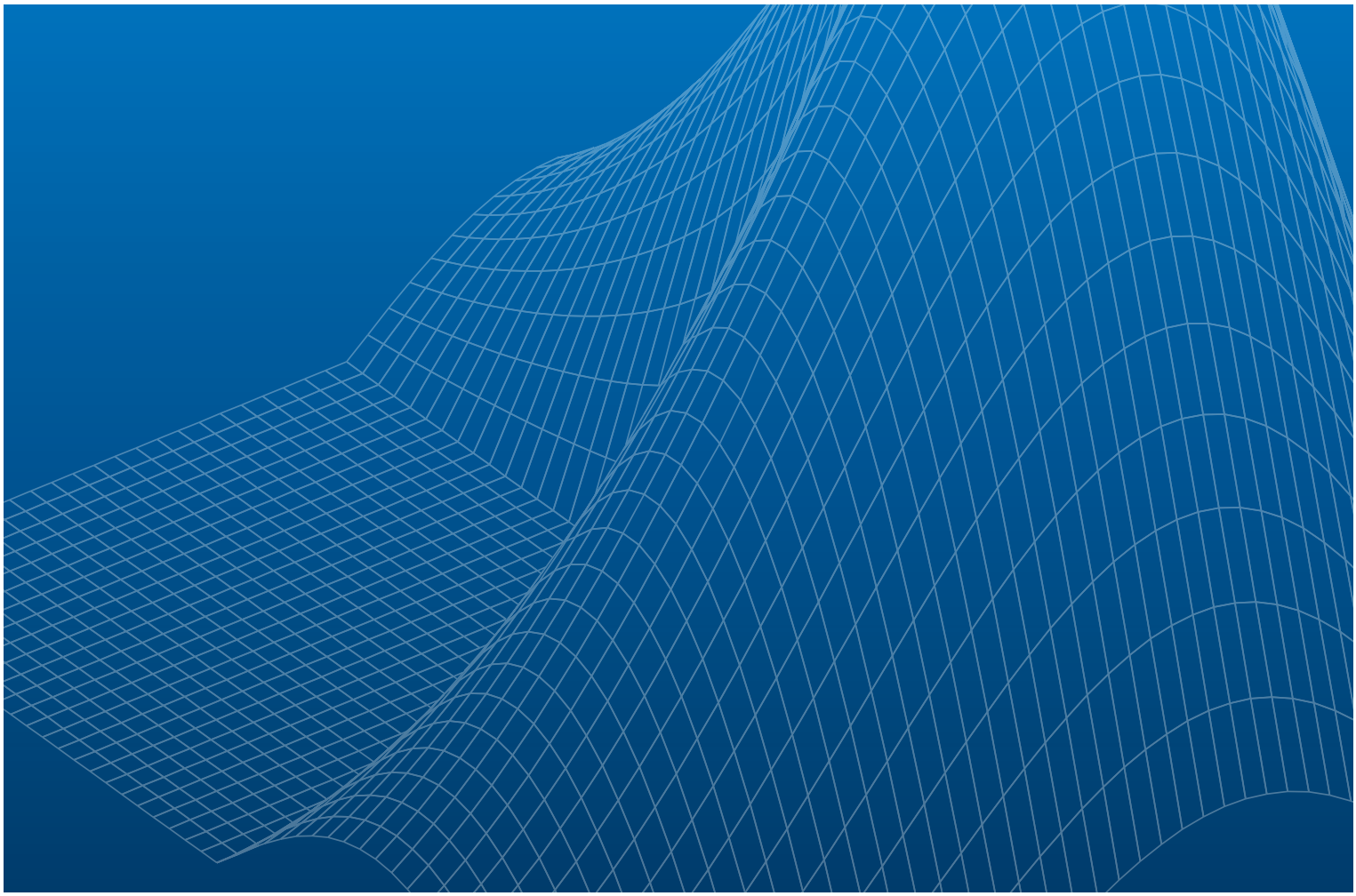


Netflix Prize与机器学习：行家看点

WHITE PAPER

MathWorks®

2006年，Netflix宣布启动著名的机器学习与数据挖掘竞赛—“Netflix Prize”，该竞赛奖金为100万美元，于2009年颁发给了获奖者。

在所有公众和媒体的关注下，胜出的解决方案最终前景如何？有没有在生产中得以运用？如果没有，原因是什么？—摘自Netflix的博文 *Netflix Recommendations: Beyond the 5 stars*不仅针对推荐系统，还针对现实世界中的商业机器学习等重要事项发表了实际见解。此白皮书详细介绍了从中吸取的经验教训。

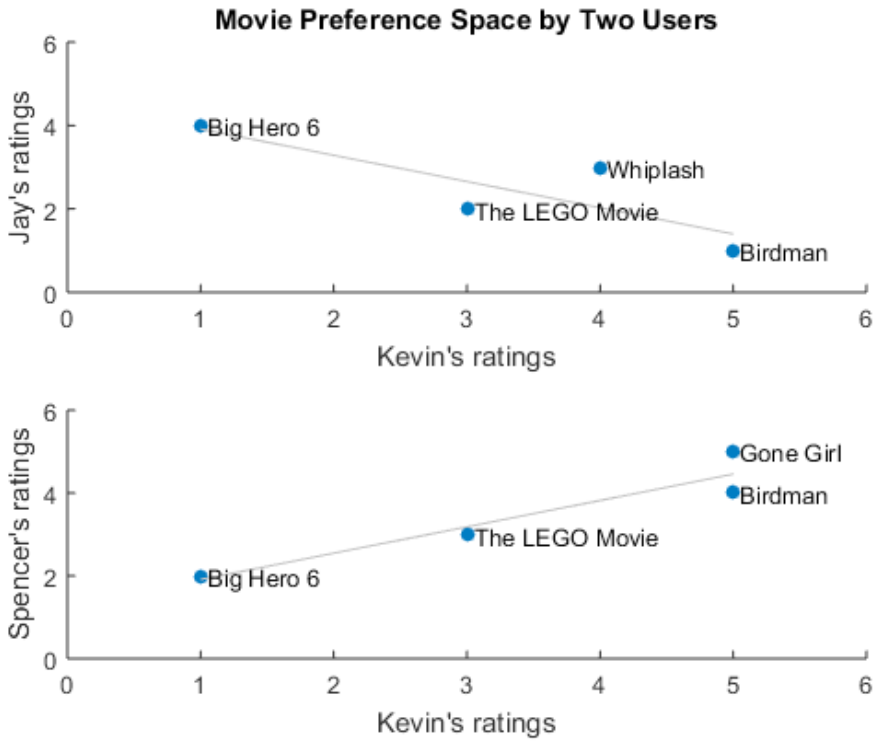
目录

Netflix竞赛 3
Netflix是否采用了胜出的解决方案 4
吸取的经验教训：新评测指标 5
吸取的经验教训：系统架构 5
避免双重实现 6
MATLAB的机器学习示例 8
关于MathWorks 8

Netflix竞赛

Netflix Prize的目标是征集电影推荐算法，该算法必须能够使现有系统的预测准确率提升10%以上。如果您使用Netflix，您会看到在“您可能喜欢的电影”或“更多此类电影”等下面列出的影片。此算法可增强Netflix的个性化用户体验。

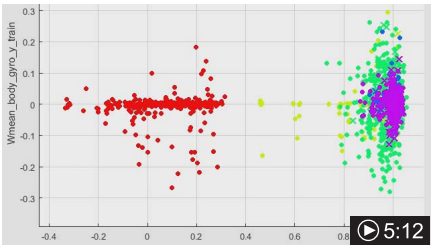
以下示例简单介绍该竞赛的甄选方法。协同过滤(*CF*)是用于推荐系统的基本算法，其所依据的理念是：您可通过趣味相似的用户给出的评分，对未评级项目进行五星评分预测。在此虚设的示例中，针对两名用户均评过分的影片，就二位给出的评分进行比较。如果您绘制一条最佳拟合线，会发现当用户评分相似时该线略微上升，用户评分不同时该线下降。



协同过滤可通过应用此相关性，综合类似用户对其他项目的评分，预测未评级项目的评分。

为确定获奖者，Netflix使用名为均方根误差*(RMSE)*的指标—算出预测评分（如1.4965星）与真实评分（如四星）的差值，取平均值后得出一个数值。如果真实评分与预测评分完全吻合，则RMSE为零。

然而，通常推荐结果都是在网站上列出排名前N的影片，并不提供原始预测的评分，比如1.4965星。因此RMSE指标是否有意义？由于是竞赛，Netflix不得不挑选出单一指标来确定获胜者。



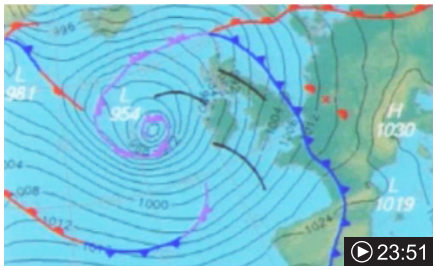
Classify Data Using the Classification Learner App

Netflix是否采用了胜出的解决方案

有两个解决方案在Netflix Prize中脱颖而出：获胜团队将预测准确率提升了10%，排名第二的团队开发的算法将预测准确率提升了8.43%。Netflix采用了可将推荐准确率提升8.43%的解决方案。Netflix没有采用提升10%的解决方案的原因是：将该解决方案应用到生产环境中所需的工程量远远超出了准确率提升所带来的利益。

此外，Netflix的商业模式从DVD租赁转向了流媒体服务，这反过来改变了数据收集方式以及进行推荐的方式。

1.57%的准确率提升带来的利益为何不值得付出努力，这是一个很有意思的思考问题。例如，您可以通过缩小低评分中的预测差距来改善RMSE。然而，Netflix是不会在用户界面上显示低评分影片的。在生产系统中，支持可扩展性和灵活性的改进措施比RMSE的影响更大。



How Weather and Pricing Affect Sales: Using MATLAB to Improve Tesco's Supply Chain

MathWorks®

吸取的经验教训：新评测指标

即使采用了该解决方案，Netflix也需要克服额外的工程挑战，因为RMSE指标鼓励参与者侧重准确性而非可扩展性和灵活性。

- 竞赛数据集中的评分数量为1亿，而实际生产系统中的数量超过了50亿
- 竞赛数据集是静态的，而生产系统中的评分数量却不断增加（该博文发表时每天增加400万）

当Netflix谈到他们当前的系统时，其突出的重点显而易见。

- “人们观看内容的75%或多或少均来自系统推荐。”
- “不断优化会员体验可显著提升会员满意度。”

Netflix认为最重要的是用户体验、用户满意度以及用户留存，这些都同商业目标相吻合，都优于RMSE。第二个要点是Netflix正在其实时生产系统上进行的A/B测试。这意味着他们在不断更新系统。

吸取的经验教训：系统架构

Netflix博主***Xavier Amatriain***与***Justin Basilico***写道：“提出一种软件架构，使其能够处理大量现有数据、响应用户交互并轻松试验新推荐方法的任务非常繁重”（要了解更多信息，请参阅此文章个性化和推荐系统架构）。

Netflix Prize胜出解决方案所应用的一项技术为集成法，其被称为“线性堆栈”。Netflix采用一种线性堆栈技术来综合多个预测模型的预测，以得出最终推荐结果。您可建立多个子系统以运行不同预测模型，然后综合这些系统的输出结果，得出最终结果。该架构非常灵活，因为您在开发新算法的同时可不断增添更多预测模型进行集成。

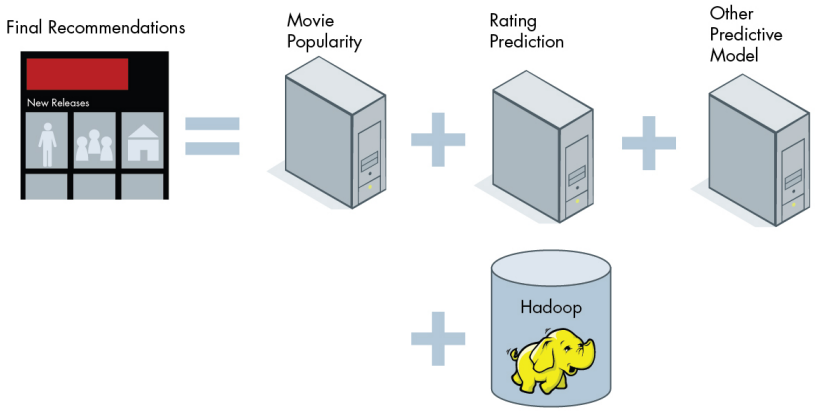


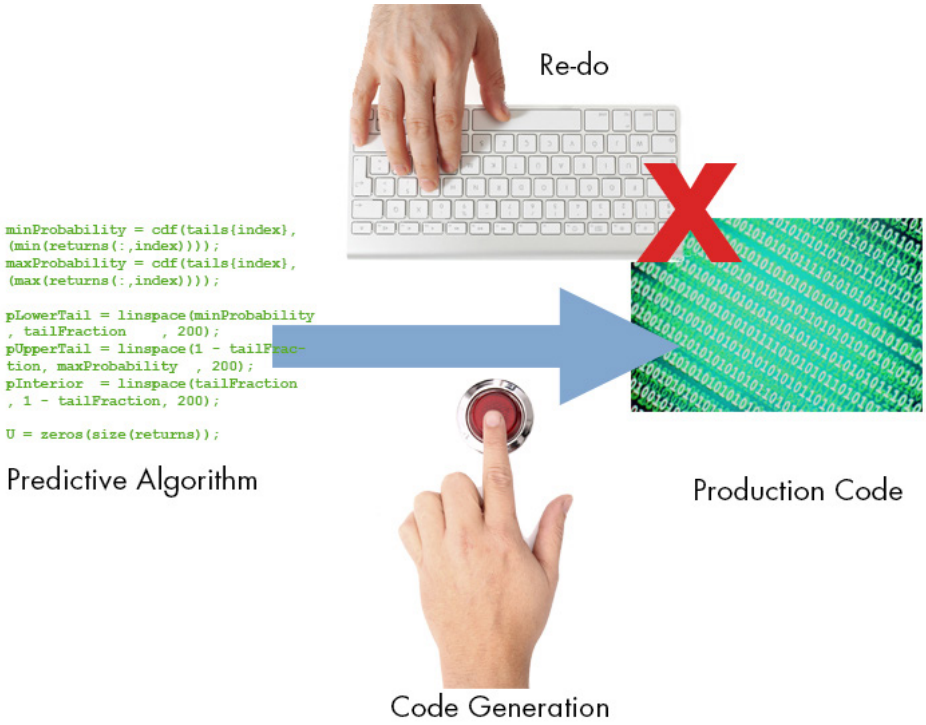
图 1：线性堆栈。

Netflix使用三层服务：脱机、近线、在线。

- 脱机处理数据—预计算批处理中耗时的步骤
- 近线处理事件—通过在活跃用户进行操作前预计算经常执行的操作并捕获结果，以此建立两个子系统间的连接
- 在线处理请求—利用脱机和近线的输出结果，即刻响应用户操作

避免双重实现

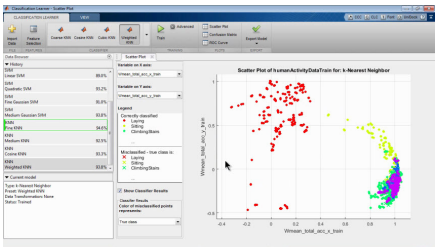
Netflix使用Hadoop运行此脱机程序，且必须在MapReduce中重写小数据集所用的算法。Netflix将该示例称为“双重实现问题”（请参阅幻灯片20），其不仅限于脱机程序。重要的是以专用工具开发、验证机器学习算法，并用其他语言重实现算法，以使其在大系统上得以扩展。该过程非常耗时，因此会限制可扩展性和灵活性。Netflix建议，尽量共享开发系统与生产系统间的组件可解决此难点。



由于*MATLAB*可以通过各种部署选项，快速将代码直接部署到生产系统中，所以可以解决此难点。

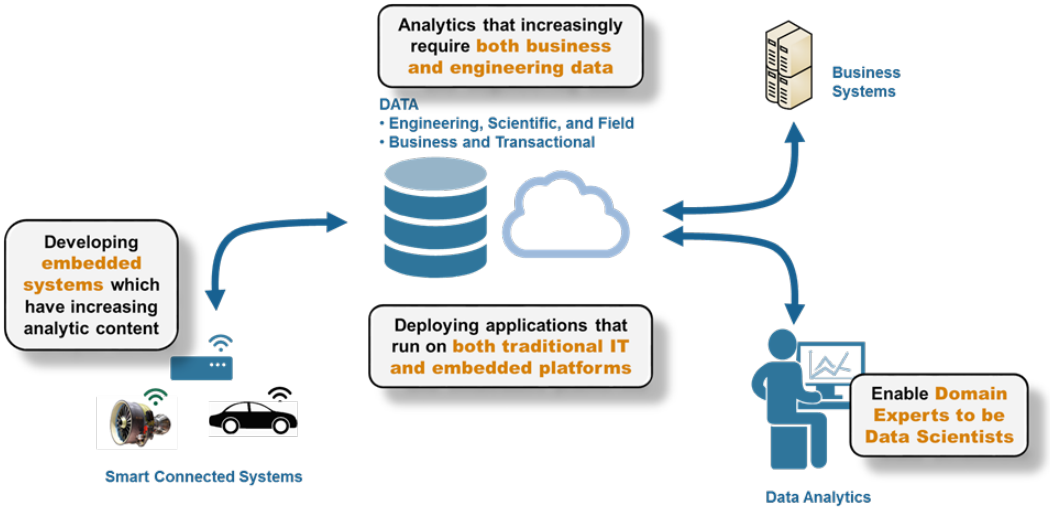
例如，在MATLAB中开发的算法可直接编译到各种部署目标中。

- 您可在MATLAB中更新算法，所作更新可立即部署到生产系统中
- 如果您切换了生产系统，只需使用适合新环境的其他部署选项即可。在生产系统中比较并验证您在MATLAB中开发的模型十分简便，因为他们所用的源代码相同



Machine Learning Made Easy

34:34

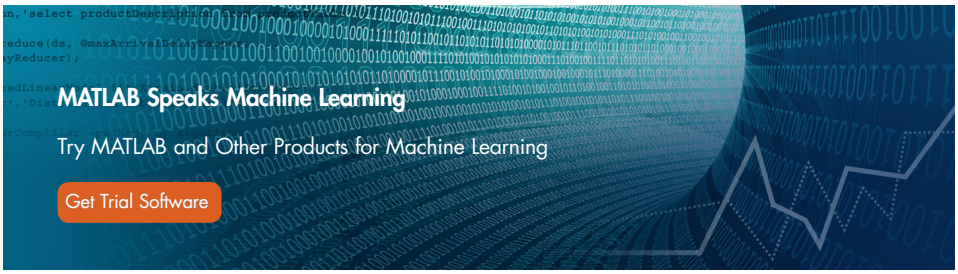


MathWorks

想要了解更多关于MATLAB功能的信息，请参阅以下资源。

- *MATLAB MapReduce与Hadoop*
- *MATLAB Production Server*
- 机器学习

使用MATLAB进行机器学习的示例



关于MathWorks

MathWorks是科学计算软件领域世界领先的开发商。全球的工程师和科学家们都依赖于MathWorks公司所提供的产品，来加快发明、创新及开发的步伐。

它所推出的MATLAB是一种用于算法开发、数据分析、可视化和数值计算的程序设计环境，称为“科学计算的语言”。该公司针对数据分析和数据处理等专业化任务开发了近100种附加产品。

请访问*cn.mathworks.com*，了解更多信息。

93027v00 07/16