Python机器学习-NBA总冠军预测项目

# 实训项目背景

## 背景意义

机器学习有下面几种定义： “机器学习是一门人工智能的科学，该领域的主要研究对象是人工智能，特别是如何在经验学习中改善具体算法的性能”。 “机器学习是对能通过经验自动改进的计算机算法的研究”。 “机器学习是用数据或以往的经验，以此优化计算机程序的性能标准。机器学习已经有了十分广泛的应用，例如：数据挖掘、计算机视觉、自然语言处理、生物特征识别、搜索引擎、医学诊断、检测信用卡欺诈、证券市场分析、DNA序列测序、语音和手写识别、战略游戏和机器人运用。

python语言在机器学习中的优势：首先是Python 的语法，它被描述为“优雅”和“数学类”。专家指出，Python 的语义与许多常见的数学思想具有特定的对应关系，因此在Python语言中应用这些数学思想并不需要学习曲线。Python也经常被描述为简单易学，这是其对任何应用使用（包括机器学习系统）的吸引力的重要组成部分。一些程序员将Python描述为具有有利的“复杂性/性能权衡”，并且描述了如何使用Python比其他语言更直观，因为其可访问的语法。还有人支出指出，Python还具有特别的工具，在使用机器学习系统方面非常有用。有些引用了一系列的框架和库，以及像NumPy这样的扩展，这些配件使得Python任务更容易实现。因此，编程语言本身的上下文对于这些应用使用的普及也很重要。另一个资源是scikit模块，称为“Python中的机器学习”，它可以指导专业人员在此能力中使用Python。

基于python语言的诸多优势和机器学习在未来人工智能领域的重要性，本项目以NBA总冠军预测为角度，利用python语言，让同学们自主定义分析模型，应用和学习机器学习算法，构建预测系统模型。

## NBA赛程介绍

NBA比赛分为：常规赛和季后赛三大部分。

季前赛为各球队的热身赛（在本项目中可以不考虑季前赛的情况）。常规赛是每年的10月末至次年的4月中旬NBA30支球队之间进行的轮回赛，每支球队需要参加82场比赛，30支球队一共进行1230场比赛。常规赛结束后，东、西部联盟常规赛战绩排前八位的球队进入季后赛争夺总冠军。

如：2017-18赛季

常规赛选出东西部的前八名，东西部分别决出东西部冠军，东西部冠军决出总决赛冠军。具体的对阵方式如下图：



图 1

## 项目要求

1. 获取数据并且进行数据处理，建立自己的数据模型，确定收集的历史数据年份，和预测的年份。
2. 利用机器学习的方法训练建模。
3. 模型评估。
4. 预测结果并展示。

# 实训目标：

强化学生的动手能力；加强学生对问题的思考能力通过对数据的分析和理解，加深对机器学习的认识；加强学生对数据的分析和处理能力，能够挖掘爬取自己需要的有用数据；自主学习机器学习的经典算法并能够应用；提高学生对机器学习领域建模的整体过程的理解，并且能够独立创新的构建自己的模型；加强学生的团队意识；提高学生自主解决问题、协作解决问题的能力。

# 系统功能

## 业务流程要求

1. 根据NBA历史信息，进行数据建模，从网上挖掘有用信息，整理和预处理相关数据。
2. 根据历史信息，得出季后赛的对阵情况，进行训练建模。
3. 进行模型评估。
4. 进行预测和展示结果。



图2

## 效果展示要求

1. 将数据特征提取，和数据建模的过程展示，并做详细的说明。
2. 展示所使用算法进行训练建模，并进行分析说明。
3. 展示如何对模型评估，并作出评估结果说明。
4. 预测结果展示，要求展示季后赛对阵图（类似图1），并作出项目总结和进一步展望。

# 技术路线

1. 建立自己的数据库，编写爬虫程序，挖掘和比赛预测相关的有用数据，并进行处理。

现成的数据网站倒还是有不少，这里整理一些常见的网站。

* [http://www.basketball-reference.com/](https://link.zhihu.com/?target=http%3A//www.basketball-reference.com/)

最常用的数据网站，以基础数据为主。可进行相关数据分析和查询。

* [http://hoopdata.com/default.aspx](https://link.zhihu.com/?target=http%3A//hoopdata.com/default.aspx)

基本数据之外加入了诸如被盖率，加罚率，以及距离分析之类的详细数据。

* [http://popcornmachine.net/](https://link.zhihu.com/?target=http%3A//popcornmachine.net/)

记录即比赛走势，可以着重观察某个球员在某场比赛的某个时段发挥情况，以及整场比赛的比分走势。

* [http://www.82games.com/](https://link.zhihu.com/?target=http%3A//www.82games.com/)

里面所记录的数据很有自己的特色，每一项基本数据都有具体的细分。

* [http://espn.go.com/espn/elias](https://link.zhihu.com/?target=http%3A//espn.go.com/espn/elias)

专门记录匪夷所思的数据。

* [http://www.prosportstransactions.com/basketball/](https://link.zhihu.com/?target=http%3A//www.prosportstransactions.com/basketball/)

1. 在机器学习和数据挖掘的应用中，scikit-learn是一个功能强大的python包。在数据量不是过大的情况下，可以解决大部分问题。学习并使用scikit-learn，建立预测模型。关于scikit-learn，大家可以参考网上的相关博客，或者scikit-learn的文档<http://scikitlearn.org/stable/>。
2. 对模型进行验证和评估。

监督学习的主要任务就是用模型实现精准的预测。我们希望自己的机器学习模型在新数据（未被标注过的）上取得尽可能高的准确率。换句话说，也就是我们希望用训练数据训练得到的模型能适用于待测试的新数据。正是这样，当实际开发中训练得到一个新模型时，我们才有把握用它预测出高质量的结果。

参考相关的博客介绍：

http://[blog.csdn.net/pipisorry/article/details/52574156](http://blog.csdn.net/pipisorry/article/details/52574156)

<https://blog.csdn.net/batuwuhanpei/article/details/51884351>。或者其他网上资源。

1. 得出预测结果。

本文具体实现过程可以参考网上例子：

参考：

https://blog.csdn.net/oxuzhenyi/article/details/64441519

# 实训计划

项目实训人分组：每组要求7-8人（建议每组中至少有一名了解篮球规则，或者爱好篮球的同学）。下面是简单的分工，但是整个实验过程全体参与，因为各个阶段的成员要相互配合。

|  |  |  |
| --- | --- | --- |
| **角色** | **人数** | **任务** |
| 项目经理 | 1 | 负责协调项目成员工作 |
| 开发过程管理 | 1 | 文档管理，代码管理，封板 |
| 数据分析处理，挖掘需求数据 | 2 | 建立自己的数据模型，挖掘所需数据，提取特征值。 |
| 训练模型 | 1 | 确定训练模型，进行训练数据。 |
| 模型评估 | 1 | 进行模型评估。 |
| 系统测试 | 1-2 | 系统测试，展示结果。 |

项目进度安排如下：

|  |  |  |
| --- | --- | --- |
| **日期** | **星期** | **实训内容** |
| 第1周 | 星期一 | 介绍实训项目，准备环境 |
| 星期二 | 搭建开发环境 |
| 星期三 | 小组分工、讨论项目计划  各小组制定计划 |
| 星期四 | 各开发小组讨论项目要求，进行需求分析，测试小组参与。 |
| 星期五 | 形成整个项目流程，并进行分工协调。 |
| 第2周 | 星期一 | 各开发小组进入设计阶段  测试小组参与项目整个环节，并思考如何展示。 |
| 星期二 | 各小组进行讨论开发。 |
| 星期三 | 代码开发、测试 |
| 星期四 | 代码开发、测试 |
| 星期五 | 代码开发、测试 |
| 第3周 | 星期一 | 完成第一阶段，数据挖掘和特征值提取。整理文档，讨论下一阶段。 |
| 星期二 | 代码开发、测试 |
| 星期三 | 代码开发、测试 |
| 星期四 | 代码开发、测试 |
| 星期五 | 代码开发、测试 |
| 第4周 | 星期一 | 完成训练模型建立，和模型评估过程。对整个实验过程进行小组讨论，根据要求总结本小组的实验成果。 |
| 星期二 | 进一步修正、测试，按照展示要求进行整理。 |
| 星期三 | 作出实验成果展示，并进行小组内讨论完善。 |
| 星期四 | 小组答辩 |
| 星期五 | 结业典礼 |

在整个过程中，大家在完成自己的任务同时，要多加讨论相互帮助。

# 实训人员考核评价标准说明

实训人员考核实行百分制，主要从下面4个方面进行：

## 完成程度（总分值50）

完成程度分为五个等级：

结构上从图2的四个方面考察：

1. 一级：系统要结构完整，系统能够运行，得分45-50；
2. 二级：完成75%以上，存在少量bug，得分40；
3. 三级：完成50%以上，存在大量bug，得分35；
4. 四级：完成25%以上，存在大量bug，得分20；
5. 五级：完成25%以下，得分20分以下；

## 创新能力（总分值20）

1、从数据模型的新颖性，特征值提取的新颖和创新度等进行适当评分。（0-5）。

2、从训练模型的选取应用和分析程度进行加分，应用不同的模型分析。（0-5）。

3、评估模型，利用不同的评估方法评测，不同的角度评测进行评分。（0-5）。

4、结果展示效果，与结论展望进行评分（0-5）。

## 团队能力（总分值20）

1. 与团队内部成员互动能力以及活跃度；
2. 帮助组内其他成员或者组外成员；
3. 组内讨论；
4. 组内角色，是否服从安排；
5. 结合组完成任务情况；

## 工作态度（总分值10）

1. 对待实训态度
2. 工作认真负责程度
3. 是否违反规定纪律
4. 完成工作质量