1. 背景介绍：

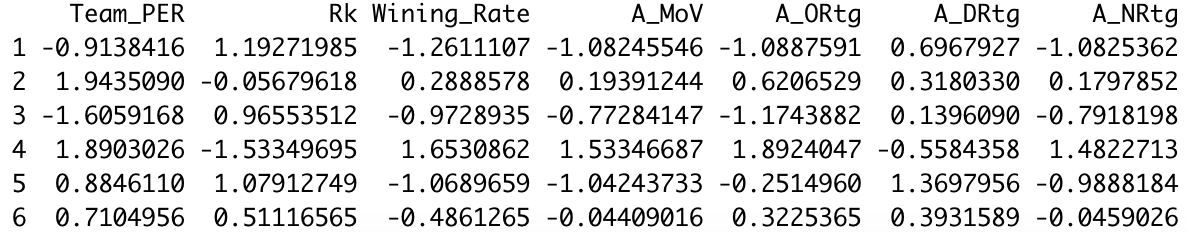
支持向量机是一种运用凸优化技术对数据集进行二分类的算法。最早起源于20世纪60年代，在70年代得到大量的研究。寻找合适的核函数在40年代末就已经被研究过。提高SVM的效率，使其应用在大规模数据集上，是统计学家研究的重点。支持向量机是针对二分类问题设计的，多分类任务是基于二分类问题的推广。支持向量回归和支持向量机相似用于回归的算法，所以我们在连续型数据集上使用SVR算法。

1. 模型构成：

* 核函数：将低维数据投影到高维数据的函数
* 超平面：在SVR中我们将它定义为一条线（蓝色），它将帮助我们预测连续值或目标值
* 间隔：支持向量机会在超平面两边找到两个边界（红色），支持向量可以在间隔上或者外边，这两个边界是用来分开数据的。
* 支持向量：最接近两个间隔的两个（或多个）观测点成为支持向量。最小化支持向量间的距离是算法的核心。在上图中，绿色的点是观测样本，我们的目标是最大化绿色点之间的距离，即最大化间隔。

三、 模型建立：

1. 首先根据数据的样子和先验知识确定核函数，将无法分割的数据投影到高维空间进行分割。最先尝试径向基函数，即高斯函数进行投射，因为高斯函数适用范围更广，更适合小数据集，且高斯核函数的抗干扰能力强，由于我们数据集存在极端值。
2. 将数据进行标准化，以防止出现因单位不同（有百分数和分数）而产生导致数据之间差距过大而模型不够稳健。
3. 将标准化的2011-12年数据作为训练集投入模型f(x,w,b)，如下：



使得预测值f(x)与y之间误差最小，未知参数为w和b。当：

时不计算损失，即允许y与f(x)之间有的偏差，当

时计算偏差。即构建一个以f(x)为中心，2的间隔带，落入此间隔带的训练样本是被正确预测的样本。

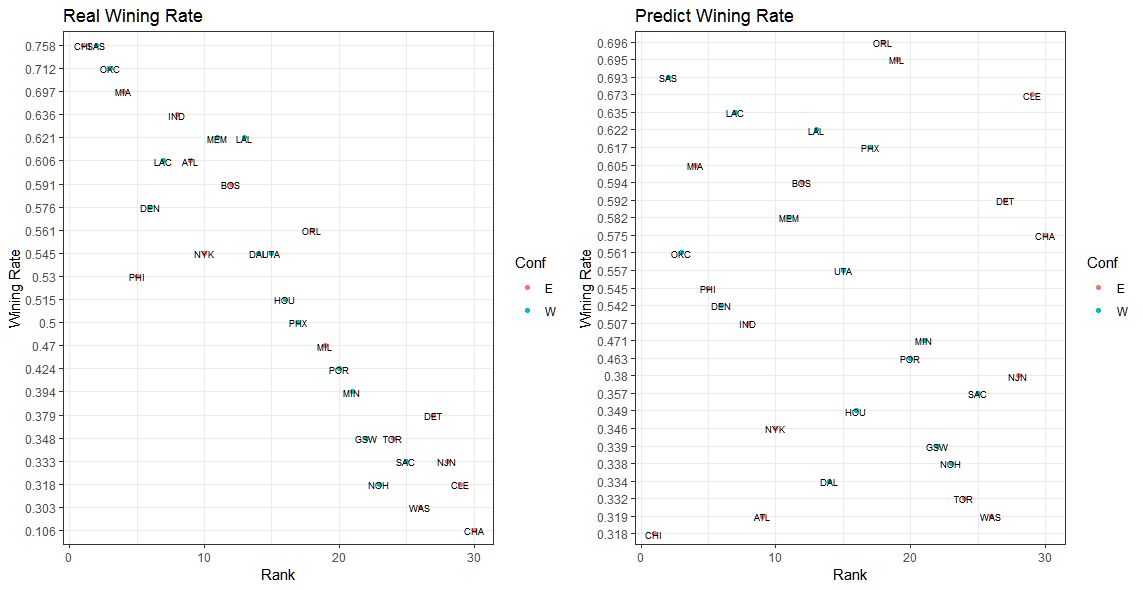
1. 构建超平面方程：

,

其中

1. 一共输入30个队伍的数据，返回29个支持向量。

根据该模型对2012-13赛季的数据进行预测，得到如下图的预测结果。



根据上述模型得到的预测结果，可以看到和真实的胜率是有一定差距的，根据支持向量模型回归的预测结果点的分散程度更大，没有规律，但是实际上胜率和球队名次是反比关系，名次越高，胜率越低。

|  |  |  |  |
| --- | --- | --- | --- |
| Team | Real | Team | Predict |
| SAS | 0.758 | ORL | 0.696002 |
| CHI | 0.758 | **MIL** | 0.6936601 |
| OKC | 0.712 | SAS | 0.6928794 |
| MIA | 0.697 | CLE | 0.6561883 |
| IND | 0.636 | LAC | 0.6476011 |
| MEM | 0.621 | LAL | 0.6218393 |
| LAL | 0.621 | PHX | 0.6173114 |
| LAC | 0.606 | MIA | 0.6051331 |
| ATL | 0.606 | MEM | 0.6027911 |
| BOS | 0.591 | BOS | 0.5917057 |
| DEN | 0.576 | DET | 0.5831185 |
| ORL | 0.561 | CHA | 0.5751557 |
| UTA | 0.545 | DEN | 0.5609477 |
| NYK | 0.545 | UTA | 0.5607916 |
| DAL | 0.545 | **OKC** | 0.5607916 |
| PHI | 0.53 | PHI | 0.5450222 |
| HOU | 0.515 | IND | 0.5028665 |
| PHX | 0.5 | POR | 0.4655509 |
| MIL | 0.47 | MIN | 0.4558707 |
| POR | 0.424 | NJN | 0.3877971 |
| MIN | 0.394 | HOU | 0.3532919 |
| DET | 0.379 | SAC | 0.3489202 |
| TOR | 0.348 | NYK | 0.3445485 |
| GSW | 0.348 | NOH | 0.3376786 |
| SAC | 0.333 | **DAL** | 0.3339315 |
| NJN | 0.333 | TOR | 0.332214 |
| NOH | 0.318 | GSW | 0.323939 |
| CLE | 0.318 | WAS | 0.3187866 |
| WAS | 0.303 | **ATL** | 0.3187866 |
| CHA | 0.106 | **CHI** | 0.3119168 |

红色加中的队伍是预测差距超过20%的队伍。预测结果和真实结果相差的百分比不是很大，但是在胜率最高和最低的极端值队伍中，支持向量机模型的表现不够好。会出现将真实值很高的队伍预测出低的结果。