Intelligent systems

Seminar Assignment 1

Assignment is to be done in pairs. Presentations will take place during the lab practice sessions in the week of December 5-11.

The files "nba0809.txt" in"nba0910.txt", which can be found on the course web page, contain data from NBA (National Basketball Association) basketball games from the 2008-2009 and 2009-2010 regular seasons. Each row represents one game.

Each game has the following attributes (standard basketball box-score statistics):

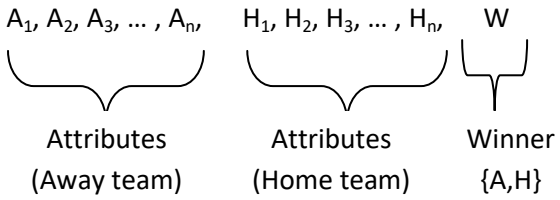
DATE	YYYYMMDD format
AWAY_NAME	three-letter away team name abbreviation
HOME_NAME	three-letter home team name abbreviation
AWAY_2PA	away team 2pt attempts
AWAY_2PM	away team 2pt made
AWAY_3PA	away team 3pt attempts
AWAY_3PM	away team 3pt made
AWAY_FTA	away team free-throw attempts
AWAY_FTM	away team free-throw made
AWAY_TO	away team turnovers
AWAY_ORB	away team offensive rebounds
AWAY_DRB	away team defensive rebounds
HOME_2PA	home team 2pt attempts
HOME_2PM	home team 2pt made
HOME_3PA	home team 3pt attempts
HOME_3PM	home team 3pt made
HOME_FTA	home team free-throw attempts
HOME_FTM	home team free-throw made
HOME_TO	home team turnovers
HOME_ORB	home team offensive rebounds
HOME_DRB	home team defensive rebounds
AWAY_PTS_Q1	points scored by away team up to the end of the 1st quarter
HOME_PTS_Q1	points scored by home team up to the end of the 1st quarter
AWAY_PTS_Q2	points scored by away team up to the end of the 2nd quarter
HOME_PTS_Q2	points scored by home team up to the end of the 2nd quarter
AWAY_PTS_Q3	points scored by away team up to the end of the 3rd quarter
HOME_PTS_Q3	points scored by home team up to the end of the 3rd quarter
AWAY_PTS_Q4	points scored by away team up to the end of the 4th quarter
HOME_PTS_Q4	points scored by home team up to the end of the 4th quarter
AWAY_PTS_FINAL	away team total points scored
HOME_PTS_FINAL	home team total points scored

Assigment

The main objective is to apply machine learning methods to quantitative analysis in sports. We are primarilly interested in forecasting the outcome of a game before the game starts.

Source data are game statistics as recorded during the actual games. As such, the data available for a particular game should not be used in forecasting the outcome of that game, becase they were not available before the game started. Therefore, we have to transform the data by calculating relevant attributes for a particular game from past games.

A simple approach would be to represent each game in one row and summarize the away and home teams with their average game statistics up to that game:



Where $A_1, A_2, A_3, \dots, A_n$ ($H_1, H_2, H_3, \dots, H_n$) are attributes calculated fom past games. For example, average points scored, etc... When interested in forecasting not just the winner but also the final points difference, we can replace W with the observed points difference for that game.

The assignment is research oriented and there are many possible ways of approaching, both in terms of model selection and data transformation. You are free to construct new attributes from the data. For example, it is well-known that field-goal shooting percentages are a good indicator of how good a team is ([2pt + 3pt made] divided by [2pt attempted + 3pt attempted]). Here are some more ideas for additional attributes: number of consecutive won/lost matches, number of non-consecutive won/lost matches in previous weeks, number of wins/losts in previous directed matches, motivation because of fans support when playing as home team, etc. **The only limitation is that for each game you can use only data from games that precede it in time (that is, older games).**

Specific tasks:

1. Data summary and visualisation
   Summarize the data with summary statistics and plots. Such exploratory analysis will help you identify potential characteristics that you can use when constructing new attributes.

2. Prediction: Classification
   Train several different types of classification models for predicting the probability of the home team winning for a given pair of teams.

3. Prediction: Regression
   Train several different types of classification models for predicting the finall points score differential for a given pair of teams.

4. Model evaluation
   Compare the chosen models in terms of predictive accuracy and comprehensiblity of results. Present the best classification and regression model.

Higher grades (9 in 10)

To be eligible for higher grades (9,10) you additionaly have to:

- predict the final winner, given the difference in points scored at the end of the 2nd quarter (half-time) or at the end of the 3rd quarter,
- predict the final points score differential, given the difference in points scored at the end of the 2nd quarter (half-time) or at the end of the 3rd quarter,
- compare how much less (more) difficult this task is compared to predicting the winner before the start of the game.

Grading

The final score will be based on the predictive accuracy of selected models and attributes, your exposition and justification of the chosen approach, and your interpretation of the final results.