Predict Allsvenskan – Bayesian modeling of football results

Gustav Sternelöv

May, 26th, 2016

Contents

1 Introduction

1.1 Project description

A very popular part of the sports industry is to predict future outcomes. In football the predictions often concerns which team that will win next game, how many goals will each team score and which team will have the most points at the end of the season. To be able to predict outcomes of this sort an estimate of how good the respective teams are in relation to each other is needed since the teams in a football league not are equally skillful. Instead, the difference in team strength is the factor with the most influence on the results.

The aim with this report is to measure the skill, or strength, of the teams in Allsvenskan, Sweden´s top football division. These measures of team skill will then be used for making predictions on the outcomes in future games. To model the strength of the teams in Allsvenskan will Bayesian techniques be used. More specifically, a so-called hierarchical Bayesian model will be estimated for measuring the team strength.

Similar types of studies with hierarchical Bayesian models has been conducted for other football leagues and these studies are highly inspirational for this study. A few examples of studies where this kind of model has been used for measuring team strength and making predicitions is presented in the next chapter.

1.2 Earlier studies

Blangiardo and Baio [2010] used match results for the 1991-92 season in the Italian Serie A to build a hierarchical Bayesian model with the aim of producing predictions of football results. The authors measured both the attacking and defensive skill of each team and used these estimates for predicting match results. Apart from the skill parameters they also estimated a parameter *home* which modeled the advantage of playing at home.
The authors found that the effect of playing at home was positive. Moreover, the league winners AC Milan had the highest estimate for the attacking parameter and was also one of the clubs with the best defensive parameter estimate. Among the teams with weakest defensive capability were the relegated side Ascoli. The predictive ability of the model were tested by simulating the results of the 1991-92 season and comparing with the actual results. In general, the authors concluded, did the hierarchical Bayesian model sem to be a rather good fit.

At his blog *Pass the ROC* Dan Weitzenfeld [2014] used the model constructed by Biangiardo and Baio as a starting point for modeling football results with 2013-14 season of the English Premier League as data. He modified the model of Biangiardo and Baio a little bit by tweaking the priors and including an intercept. Weitzenfeld also concluded that there is an significant advantage of playing at home. He also found it to be a good idea to include an intercept since the HDI (Highest density interval) for the parameter not included zero. Just as Biangiardo and Baio, Weitzenfelds estiamtes of the average attacking/defensive were clearly correlated with the teams positions in the league. Weitzenfeld checked the predictive abiltiy of the model by simulating the 2013-14 season and comparing the actual number of goals during the season with the simulated number of goals. He noted that the model worked rather well, but that shrinkage toward zero for the attack parameter had some implications on the estimates. For the high-scoring teams the model therefore overshrinked the estimates and predcited these teams to score slightly fewer goals than they actually did. Reversly, for the low-scoring teams, the model estimated them to score more goals than they actually did.

A third example of where a hierarchical Bayesian model has been used for predicting the outcomes of football games is in the paper "*Modeling Match Results in La Liga Using a Hierarchical Bayesian Poisson Model*"[Bååth, 2013]. Bååth's model is a bit different to the two eariler ones in the sense that he measures the team strength by a single skill parameter instead of with both an attack and a defense parameter. To model the home-advantage he specified two different intercepts, one for the team playing at home and one for the away team.
Bååth also found that there is a significant home advantage and that it corresponds to, on average, almost 0.5 more goals for the home team. The author then compares the expected number of goals when playing at home for all teams given the skill parameter to examine how well the teams are ranked by the parameter.

This ranking agrees well with actual ranking and realtion between the teams in the league. For testing the predictive ability of his model Bååth then examine how well the simulations agrees with the actual number of goals in game and the actual results. He finds that his model on 34 % of the times predicts the correct number of home goals and that it predicts the right match outcome 56 % of the time. He tehn also simulates results for all remaining games of the 2012-13 season of La Liga but does not compare it with the actual outcomes.

1.3 Data

In the data set are all the match results for the 2015 season of Allsvenskan and the results for the first twelve rounds of the 2016 season. For the remaining games of the 2016 are there no results since these games not have been played yet. There are 16 teams in Allsvenskan and they play each other twice in a season, on time at home and one away. After the 2015 season two teams were relegated (Åtvidabergs FF and Halmstads BK) and replaced by two teams (Östersunds FK and Jönköpings Södra) from the second highest divison, Superettan. As a result of that, the data set includes match results for 18 different teams and the number of observations vary between the teams. For the relegated sides there are 30 observations and for the new teams 12 observated results, the other teams have 42 observated match results each. The data set contains four columns which are *HomeTeam*, *AwayTeam*, *HomeGoals* and *AwayGoals* and the first six observations are shown below.

```
##        Home.Team       Away.Team Home.Goals Away.Goals
## 1    Hammarby IF       BK Häcken          2          0
## 2      Kalmar FF Helsingborgs IF          0          0
## 3 Falkenbergs FF        Gefle IF          0          2
## 4   IFK Göteborg  Åtvidabergs FF          1          0
## 5 Djurgårdens IF     IF Elfsborg          1          2
## 6 IFK Norrköping       Örebro SK          1          1
```

2 Method

2.1 The model

The number of goals scored in a game or by each team can be shown to be Poisson distributed. This is proven in a lot of articles, for instance by Baio and Biangiardo in the article mentioned earlier in the report.

$$Goals \sim Poisson(\lambda)$$

The expected number of goals for the home and away team are then modeled as being Poisson distributed in the following way.

$$y_{gj}|\theta_{gj} \sim Possion(\theta_{gj}$$

The number of goals scored by the home team is denoted as... and for the away teams as... g is the gth game in the season, or as in this example the sample of games since the data set contains more than a full season

The θ parameters represents the average number of goals scored in a game by the home team and the away team. For modeling these parameters is a log-linear random effct model proven to be a suitable choice (for example, Baio and Biangiardo). Then, the expressions for the respective log θ parameters are specified as follows.

$$log\theta_{gH} = Intercept_{Home} + Skill_{Home} - Skill_{Away}$$
$$log\theta_{gA} = Intercept_{Away} + Skill_{Away} - Skill_{Home}$$

For the log θ:s are two different intercepts estimated, one for the home team and one for the away team. In that way is the advantage of playing at home included in the model. The intercepts shall be interpreted as the log average number of goals in a game when both teams are equally good.
An self-evident assumption to make is that the teams in Allsvenskan are on different levels where some teams are more skillful and some teams are less skillful. However, this variable of actual skill of the teams is a latent variable since it not is possible to observe it directly. Instead earlier match results are collected for estimating the skill parameter for each team. The prediction for the number of goals a team will score is then given by the intercept plus the skill of the team minus the skill of the opponents.
The priors for the intercepts are set to be very vague in all examples mentioned in the *earlier studies* chapter. Perhaps they not have to be that vague, but on the other side do I not feel sure enough to set a prior which not is considered as vague.

$$Intercept_{Home} \sim Normal(0, 4^2)$$

$$Intercept_{Away} \sim Normal(0, 4^2)$$

The prior for the skill parameter is normally distributed with the hyper-priors μ_{Teams} and σ^2_{Teams}) as parameters.

$$Skill_{1,\dots,T} \sim Normal(\mu_{Teams}, \sigma^2_{Teams})$$

The hyper-priors for the skill parameter are also chosen to be vague:

$$\mu_{Teams} \sim Normal(0, 4^2)$$

$$\sigma_{Teams} \sim Normal(0, 4^2)$$

The decision to set the mean for the μ_{Teams} to zero may cause a bit of overshrinkness for the skill estimates. However, as mentioned earlier, this was mainly a problem for the teams at the very top of the Premier League who were a lot better than the other teams during that season. Allsvenskan on the other hand is well-known for being a very even league with a lot of teams competing for the title instead for just a few. If this fact lowers the risk of overshrinkage is hard to tell, but it might be interesting to compare and investigate if it seem o have any effect.

Anchors the first skill parameter at zero

2.2 Gibbs sampling

2.3 Expected number of goals

How to do for analysing the skill measures (beräkningar i Bååts raport för att få förväntat antal mål på hemmaplan t ex.)

An interesting result might be the estimate of how many goals a team is expected to score. This can be computed by...
One estimate for when the teams are playing at home and one for when they are playing away.

2.4 Highest density intervals (HDI)

3 Software

4 Results

5 Other things

Could also write (more) about:
How they set their priors(very vague)

Do not believe that the priors have to be that vague. Do not have to be an expert to know how many goals a home or away team can be expected to score. Neither is it unreasonable to belive that we know, at least for Allsvenskan, that the variance between the teams in skill is fairly low.
The problem with overshrinking
Setting the prior mean the zero means that the parameter estimate will be shrunken towards zero. This may cause overshrinkage for the skill parameters, with the effect that the skill estimate for the top-teams and the bottom-teams are shrunken to much towards zero.

Look at problems, fixes with priors etc in the earlier studies. For example the implications of setting attacking/defensive/mu parameters to zero.
It should absolutely be possible to set priors that not are so vague. Some knowledge about footboll should say that there is no need for them to be that vague.