Stats Challenge

Miguel Pereira

17/04/2020

Problem

You have been handed control of a football team which plays every year in a league. Home and away games mostly alternate, and the location of the next game can be modelled as depending only on the location of the previous game and some random factors.

Given that the team wins 75% of its home matches and loses 80% of its away matches, and given the records of past seasons, how would you determine the probability that the next game will be home given that the current game is away?

Proposed solution

Let X be a random variable corresponding to the location of the match where it can take one of two possible values: home or away. The goal is to calculate of the following game being home given that the current games was away. If assume time steps indexed by t, this corresponds to calculating $P(X_{t+1} = Home|X_t = Away)$.

The locations of the matches are a Markov chain since the probability of an event occurring depends solely on the previous event and not on the events before that. Formally, this means that the probability of $X = x$ at time $t + 1$ is expressed as follows:
$P(X_{t+1} = x|X_1 = x_1 + X_2 = x_2 + \ldots + X_t = x_t) = P(X_{t+1} = x|X_t = x_t)$, where $x \in \{Home, Away\}$

In this particular case, we know that at time t the game was away and want to calculate the probability of the next game being home: $P(X_{t+1} = Home|X_t = Away)$.

In addition, we have information about the conditional probabilities of winning or losing a game given its the location. If we consider R to be a random variable corresponding to the result of a given game (win or loss), we know the following:

1. $P(R_t = Win|X_t = Home) = 0.75$
2. $P(R_t = Lose|X_t = Away) = 0.80$

and given that there are only two possible outcomes to a game (win or loss), we also know:

3. $P(R_t = Lose|X_t = Home) = 0.25$
4. $P(R_t = Win|X_t = Away) = 0.20$

We also have access to the records of previous games/past seasons. If we assume that these records contain the both the results and the location, we can treat this as a simple Markov chain and calculate the starting and transition probabilities based on the records from past seasons:

- Starting probabilities (at t_0) - using the records of multiple seasons, the starting probabilities are computed by counting the number of seasons the team started the season playing home and away and dividing that by the number of seasons and so that $P(Home_0) + P(Away_0) = 1$. If we only have access to one season, we can approximate this probability to be 0.5, $P(Home_0) = P(Away_0) = 0.5$.
- Transition probabilities - In this simple scenario, we would count the number of times a game away was followed by a game at home and divide this by the total number of transitions that occurred from an away game in the past. For examle, assuming the sequence: $home, away, home, home, away, away, away$, $\hat{P}(X_{t+1} = Home|X_t = Away) = \frac{1}{3}$.

This can be generalised: let n_{ij} be the number of times that the process moved from state i to state k with $i, k \in \{Home, Away\}$, the probability is given by:

$$\hat{\mathbf{P}}_{ij} = \frac{n_{ij}}{\sum_{k=1}^{m} n_{ik}},$$

where m is the number of states ($m = 2$ in this case).

A more complex scenario occurs when we do not know the location of the games in the past seasons and only know the results of each game. This is an example of a hidden Markov model and the goal here is again to estimate the transition probability matrix without having observed the sequence of events we are interested in. However, we know the emission probabilities, which simplifies the problem.

In order to estimate our probability of interest, $P(X_{t+1} = Home|X_t = Away)$, we would have to look at the records from previous seasons and assume a starting probability for a game being home and away as well as define a prior transition matrix. I will illustrate this with an example.

Let's assume the following set of observations: $win, win, win, lose, lose, win, win$, which corresponds to the following set of transitions: $(win, win), (win, win), (win, lose), (lose, lose), (lose, win), (win, win)$.

We have a prior belief that the games usually alternate, so we can assume that the transition matrix has equal probabilites of transitions across states:

Prior transition matrix

	Home	Away
Home	0.5	0.5
Away	0.5	0.5

Also, we can assume that, at the start of each season, the team is equally likely to play at home or away: $P(Home_0) = P(Away_0) = 0.5$.

Now, we need to calculate the probability of a transition from Away to Home given our observations. For that we need to consider the joint probability of each of the 4 possible observable outcomes and the hidden transition Away-Home:

1. $P((Win, Win), (Away, Home)) = P((Win, Win)|(Away, Home)) \cdot P((Away, Home)) = P(Win|Away) \cdot P(Win|Home) \cdot P(Home|Away) \cdot P(Away)$
2. $P((Win, Lose), (Away, Home)) = P((Win, Lose)|(Away, Home)) \cdot P((Away, Home)) = P(Win|Away) \cdot P(Lose|Home) \cdot P(Home|Away) \cdot P(Away)$
3. $P((Lose, Lose), (Away, Home)) = P((Lose, Lose)|(Away, Home)) \cdot P((Away, Home)) = P(Lose|Away) \cdot P(Lose|Home) \cdot P(Home|Away) \cdot P(Away)$
4. $P((Lose, Win), (Away, Home)) = P((Lose, Win)|(Away, Home)) \cdot P((Away, Home)) = P(Lose|Away) \cdot P(Win|Home) \cdot P(Home|Away) \cdot P(Away)$

We also need consider the most likely hidden state given the results. E.g. given that the team usually wins when it plays at home and usually loses when it plays away, if it wins two games consecutively, the most likely (expected) hidden states are Home-Home.

Probabilities of observing each state transition given an Away-Home transition

Observed states	Joint probability	Likelihood	Expected transition
win,win	0.038	0.141	Home-Home
win,win	0.038	0.141	Home-Home
win,lose	0.012	0.150	Home-Away
lose,lose	0.012	0.160	Away-Away
lose,win	0.150	0.150	Away-Home
win,win	0.038	0.141	Home-Home

The final probability $P(X_{t+1} = Home|X_t = Away)$ is given by the ratio of the sum of the joint probabilities over the sum of the probability of the most likely transition at each step.

In this example, and taking into account our prior beliefs, the final probability would be proportional to:

$$P(X_{t+1} = Home|X_t = Away) = \frac{0.038+0.038+0.012+0.012+0.15+0.038}{0.141+0.141+0.15+0.16+0.15+0.141} = 0.326$$

This is not the actual probability of transition as we would need to calculate the other probabilities of the different transitions and normalise in order for their sum to be 1.