

Projekt nr. 1:

Analyse af resultater fra Premier League 2000–2006.

Datasættet, som er stillet til rådighed af Thorbjørn L. Dahl, omfatter resultaterne af alle fodboldkampene i den engelske superliga Premier League i de seks sæsoner 2000–2001, 2001–2002, ..., 2005–2006.

Premier League består løbende af 20 klubber (tre af dem udskiftes efter hver sæson), som i løbet af hver sæson spiller alle mod alle både på ude- og hjemmebane, således at der i løbet af en sæson spilles i alt $2 \times \binom{20}{2} = 380$ kampe. På datafilen `07-1.txt` er der (ud over en overskriftsline med variabelnavne) en linie for hver af de $6 \times 380 = 2280$ kampe. Oplysningerne for hver kamp står, adskilt af kommaer, i følgende rækkefølge:

1. Overflødig oplysning (“E0” betyder at divisionen er Premier League)
2. Dato på formen dd/mm/åå
3. Hjemmeholdets navn (se liste på næste side)
4. Udeholdets navn (se liste på næste side)
5. Antal mål scoret af hjemmeholdet
6. Antal mål scoret af udeholdet

Herefter kommer en lang række oplysninger om dommer, halvløbsresultater, antal skudforsøg, forlænget spilletid, tilskuertal, bookmaker odds osv. osv., se filen `datadok.txt` for yderligere oplysninger. Vi får formodentlig kun brug for de oplysninger der er listet ovenfor.

Thorbjørn Dahls hensigt med fremskaffe dette datasæt er at undersøge i hvor høj grad “sports betting” — altså væddemål om udfald af sportsbegivenheder, i Danmark bedst kendt i form af tipning — kan være et seriøst alternativ til andre former for investering. Her er der naturligvis ikke tale om almindelig tipning eller andre spil der udbydes af Dansk Tipstjeneste (eller hvad det hedder idag), der er alt for høje afgifter lagt på. Idag kan man spille på nettet hos udbydere, der kun tager et forholdsvis beskedent gebyr. Ideen skulle være, at man ved at holde sig omhyggeligt a jour med hvad der sker i klubberne (sygdom eller skader, konflikter mellem træner og spillere eller mellem spillere indbyrdes, risiko for nedrykning osv.) kan styre sine indsatser på en sådan måde, at man i gennemsnit får et højt og ikke alt for risikobehæftet afkast.

For overhovedet at kunne regne på dette må man naturligvis have en statistisk model, estimeret ud fra historiske data, som man kan bruge til

estimation af sandsynlighederne for udfaldene af de kampe man spiller på. Denne model — evt. med mindre korrektioner baseret på detailviden om den seneste udvikling — udgør grundlaget for de beslutninger man skal tage, og det er den vi skal beskæftige os med i dette projekt.

Projektet går altså ud på at finde den bedst mulige statistiske model til beskrivelse af udfaldene af kampene i den engelske superliga, og undersøge i hvor høj grad den kan bruges til forudsigelse af udfaldet af den enkelte kamp.

Navne på de 30 klubber, som har spillet i Premiers League mindst en af de seks sæsoner:

Arsenal
Aston_Villa
Birmingham
Blackburn
Bolton
Bradford
Charlton
Chelsea
Coventry
Crystal_Palace
Derby
Everton
Fulham
Ipswich
Leeds
Leicester
Liverpool
Man_City
Man_United
Middlesbrough
Newcastle
Norwich
Portsmouth
Southampton
Sunderland
Tottenham
West_Brom
West_Ham
Wigan
Wolves

Projektbesvarelsen afleveres mandag d. 22. oktober.