Empirical Bayes Estimation in Baseball Statistics

Uvod

Tekst *Understanding Empirical Bayes Estimation (using baseball statistics)* sa sajtavarianceexplained.or predstavlja objašnjenje Bajesove procjene preko statistika iz bejzbola. Autor je David Robinson i koristi podatke o efikasnosti bejzbol igrača (tačnije, broj pogodaka i broj pokušaja) kako bi prikazao važnost i primjenu ovog statističkog pristupa, koji balansira između pojedinačnih rezultata i grupnih prosjeka.

Opis problema

Cilj je procijeniti stvarnu stopu uspjeha (*batting average*) za svakog igrača na osnovu broja pogodaka (hits) i broja pokušaja (at-bats). Međutim, mali broj pokušaja može dovesti do velikih varijacija i nepouzdanih procjena. Primjer-igrači koji su imali samo nekoliko pokušaja i stopu uspjeha od 100%, što je vjerovatno posljedica male veličine uzorka. Problem je jednostavno predstavljen kroz primjer matematičkog gledišta na statistiku i efikasnosti igrača u bejzbolu. Time hoću da kažem da je matematički bolja statistika 4 uspješna pokušaja od 10 ukupnih, nego 300 uspješnih od 1000 ukupnih. Međutim ako iste brojeve posmatramo kroz efikasnost igrača u bejzbolu, npr. publika i navijači će mngo više cijeniti igrača čiji je uspjeh 300 od 1000 jer je njegova efikasnost "primjetnija", odnosno više je uticala na rezultat i on postaje zapaženiji igrač iako mu je statistika lošija. Sam tekst nije suštinski o bejzbolu, nego je samo napisan na osnovu tog primjera kako bi ilustrativno bilo lakše razumjeti temu i problem.

Frekventistički pristup

Standardni frekventistički pristup računa postotak uspješnosti (p) kao:

$$\hat{p} = \frac{\text{hits}}{\text{at-bats}}$$

Međutim, kod malih uzoraka ova procjena je jako varijabilna. Na primjer, igrač koji ima 2 pogodaka iz 2 pokušaja ima $\hat{p} = 1.00$, što ne odražava njegovu stvarnu sposobnost.

Bajesovski pristup

Bajesovska statistika nudi alternativu – umjesto da vjerujemo isključivo posmatranim podacima, uzimamo u obzir i prethodno znanje ili pretpostavke.

Beta raspodjela kao prior

Za binarne ishode (uspjeh/neuspjeh), odgovarajući prior je **Beta raspodjela**:

$$\text{Beta}(\alpha, \beta)$$

Kombinovanjem sa podacima (broj uspjeha i broj neuspjeha), dobijamo posteriornu raspodjelu:

$$\text{Beta}(\alpha + \text{hits}, \beta + \text{at-bats} - \text{hits})$$

Procjena vjerovatnoće (srednja vrijednost posteriorske raspodjele) je:

$$\hat{p}_{\text{bayes}} = \frac{\alpha + \text{hits}}{\alpha + \beta + \text{at-bats}}$$

Empirijski Bayes

Umjesto da ručno biramo α i β, **Empirijski Bayes** pristup koristi podatke iz populacije da ih procijeni. Pretpostavlja se da svi igrači dolaze iz zajedničke Beta raspodjele. Na osnovu ukupnog broja pogodaka i pokušaja u ligi, možemo procijeniti vrijednosti α i β metodom maksimalne vjerovatnoće (MLE).

Prednosti Empirijskog Bayesa

- **Stabilizacija procjena**: Igrači sa malo pokušaja imaju njihove procjene „povučenije“ ka prosjeku lige.
- **Balans između individualnih i grupnih informacija**: Igrači sa više podataka dobijaju više „povjerenja“ u njihove stvarne vrijednosti, dok oni sa malo podataka više zavise od grupnog prosjeka.
- **Jednostavna interpretacija**: Empirijski Bayes koristi klasične alate (npr. Beta raspodjelu), ali ih oslanja na podatke, umjesto subjektivne pretpostavke.

Primjeri iz članka

U članku je prikazano kako Empirijski Bayes značajno poboljšava rangiranje igrača. Igrač koji je imao 3/3 pogodaka više nije na vrhu liste jer je njegova procjena korigovana prema prosjeku, dok su pouzdanije procjene za igrače sa mnogo pokušaja zadržane.

Vizuelizacije

Autor koristi vizualizacije (npr. histogrami procijenjenih vrijednosti, poređenja prije/poslije korekcije) da pokaže razliku između frekventističkog i Empirijskog Bayes pristupa. Posebno se ističu „povlačenja“ ekstremnih vrijednosti ka prosjeku.

Zaključak

Empirijski Bayes pruža snažan i intuitivan način da se stabilizuju statističke procjene u prisustvu varijabilnosti i malih uzoraka. Na primjeru baseball statistike, pokazano je kako ovaj pristup daje uravnoteženije i pouzdanije rangiranje igrača. Ova tehnika je korisna u mnogim oblastima, uključujući medicinu, obrazovanje i eksperimentalnu analizu, gdje imamo slične izazove.

Link do teksta: http://varianceexplained.org/r/empirical_bayes_baseball/