TDAB01 SANNOLIKHETSLÄRA OCH STATISTIK

LABB 3: BAYESIANSK INFERENS

JOLANTA PIELASZKIEWICZ, MARYNA PRUS, JOSE M. PEÑA, MÅNS MAGNUSSON, MATTIAS VILLANI
IDA, LINKÖPINGS UNIVERSITET

1. INSTRUKTIONER

- Labbrapport ska lämnas i PDF-format med alla nödvändiga koder, analys och grafer.
- I rapporten ska följande ingå:
  - Båda studenternas namn och LiU-id.
  - Laborationsnummer
- Ange tydligt vilken uppgift ni löser, t. ex. genom att dela upp rapporten i avsnitt.
- Ett tips är att använda R markdown. Här finns en R markdown mall att utgå ifrån. Antingen kan ni skapa PDF direkt från denna mall i R-Studio (med TeX) eller så skapar ni ett Word/HTML dokument som sedan skrivs ut till PDF.
- Laborationsrapporten skickas in via LISAM.
- Information om deadline finns i Lisam/Inlämning.

2. INTRODUKTION TILL R

R är ett programmeringspråk för statistisk programmering som påminner mycket om Matlab. R bygger på öppen källkod och kan laddas ned här. R-Studio är en mycket populär IDE för R (som också påminner mycket om Matlab). Denna IDE finns att tillgå här. I R-Studio finns funktionalitet för literate programming med R markdown implementerat för att kombinera R kod med markdownsyntax. På detta sätt är det enkelt att generera rapporter med både text, grafik och kod. Det är R:s motsvarighet till Python Notebook.

För en ingång till R från andra språk kan onlineboken *Advanced R* rekommenderas som finns här. Kapitlen *Data structures*, *Subsetting* och *Functions* bör ge en snabb introduktion.

Även boken *The art of R programming* av Norman Matloff kan vara till hjälp som referenslitteratur. Boken finns här.

2.1. **Videomaterial.**

- För en introduktion till syntaxen i R se Google developers R videomaterial här.
- Mer (detaljerat) videomaterial av Roger Peng finns att tillgå här.
- För att visualisera med basgrafiken finns följande introduktionsvideo.
- För mer komplicerad grafik rekommenderas `ggplot2` paketet. En introduktionsvideo finns här.
- En introduktion till R markdown finns här.

2.2. **Cheatsheets.**

- *R reference card v.2* av Matt Baggot med vanliga funktioner i R finns att tillgå här.
- *R markdown cheatsheet* av R-Studio med tips för R markdown finns att tillgå här.

3. LABORATION

I denna laboration kommer vi gå djupare in på Bayesianska metoder. När vi arbetar med Bayesianska metoder betraktar vi inte längre våra okända parametrar som konstanta, utan vi betraktar dem som okända stokastiska variabler.

3.1. **Bayes sats och aposteriorifördelningen.** Bayes sats ges av

$$f(\theta|\mathbf{y}) = \frac{f(\mathbf{y}|\theta)f(\theta)}{f(\mathbf{y})}.$$

Dock kan $f(\mathbf{y}) = \int f(\mathbf{y}|\theta)f(\theta)d\theta$ ofta var kluriga att beräkna. Då vi är intresserade av en given parameter θ kan vi i många fall ”kasta” bort de delar som inte innehåller vår parameter av intresse, dvs

$$f(\theta|\mathbf{y}) \propto f(\mathbf{y}|\theta)f(\theta)$$

vilket är en onormaliserad aposteriorifördelning. Se sidor 353-354 (342-343 i 2:e upplagan) i Baron för exempel på en härledning av en onormaliserad sannolikhetsfunktion.

3.1.1. *Visualisera posteriorn.* Vi ska nu visualisera en lite klurigare posterior. Antag att dina data kommer från en normalfördelning där $\sigma = 1$, dvs känd. Du är intresserad av parametern μ för denna normalfördelning och vill beräkna posteriorn för μ. Som prior för μ använder vi en t-fördelning med $\nu = 1$.

(1) Visualisera din prior exakt över intervallet [-5,15]. Använd `dt()`.
(2) Nedan är sju datapunkter som du observerat. Visualisera dessa som ett histogram på intervallet [-5,15]. **Tips!** Använd argumentet `xlim` i `hist()`.

```
[1] 11.3710  9.4353 10.3631 10.6329 10.4043  9.8939 11.5115
```

(3) Skapa en funktion för log-likelihooden för μ som du kallar `normal_log_likelihood(mu, data)`. Anta att $\sigma = 1$. Visualisera log-likelihooden för μ över intervallet [-5,15], precis som med priorn i uppgift (1).

```
> llik <- normal_log_likelihood(5, data)
> round(llik, 1)

[1] -114.6
```

(4) Härled (analytiskt, steg för steg) den proportionella (onormaliserade) posteriorn för μ, dvs $f(\theta|\mathbf{y}) \propto f(\mathbf{y}|\theta)f(\theta)$. Tänk på att de faktorer som inte innehåller μ kan förkortas bort.
(5) Visualisera den onormaliserade posteriorn på samma sätt som priorn och likelihooden ovan.

3.2. **Binomialmodell med beta prior.** Vi ska nu studera aposteriorifördelningen för sannolikheten p i en binomialfördelning. Mer information finns i sida 357 (344 i 2:e upplagan) i Baron.

3.2.1. *Produkt A eller B ?.* Du har precis skapat en startup med två produktidéer, A och B. Du har skapat prototyper för de två produkterna och demonstrerat dem för ett antal personer. Du är intresserad av att veta hur många personer som kan vara intresserade av dessa produkter, dvs antal intresserade personer kan modelleras av två binomialfördelningar med parametrar p_A och p_B.

(1) För att det ska gå att dra slutsatser från materialet behöver du bestämma din prior för p som en betafördelning. Vilka parametrar väljer du för betafördelningen och varför ? Visualisera din prior, eller om du väljer två olika priorfördelning (en för varje produkt), visualisera båda dessa priorfördelningar. **Obs!** Tänk på att du ännu inte observerat några data ännu.

(2) Produkt A har du nu demonstrerat för 13 personer varav åtta var intresserade och produkt B har du bara haft möjlighet att demonstrera för tre personer och av dessa var två personer intresserade. Du kommer initialt bara kunna skapa en av dessa produkter och kommer därför behöva välja vilken produkt du ska satsa på.

Använd konjugategenskapen mellan beta och binomialfördelningen för att räkna ut din posteriorfördelning analytiskt i respektive fall. **Obs!** Du har bara en dragning från As binomial och en dragning från Bs binomial, ej 13 och tre dragningar. Likadant för B. Beräkna den förväntade proportionen för respektive produkt (med hjälp av det förväntade värdet för en betafördelning). Vilken produkt har den högsta förväntade proportionen intresserade ?

(3) Storleken på din marknad är 87 andra personer du vill nå med dina produkter. Använd dina två aposteriorfördelningar för respektive produkt för att simulera hur många intresserade kunder due kan tänkas få för respektive produkt. Simulera först från din posteriorbetafördeling och använd sedan de simulerade värdena p_i för produkt i för att dra en binomialfördelad variabel med $X_i \sim Binomial(n = 87, p_i)$. Visualisera fördelningen över antalet intresserade kunder ni kommer ha med respektive produkt.

(a) Hur stor är sannolikheten att du får fler än 40 intresserade kunder med respektive produkt ?

(b) Vad är det förväntade antalet intresserade kunder av respektive produkt, dvs $E(X_A)$ och $E(X_B)$?