TDAB01 SANNOLIKHETSLÄRA OCH STATISTIK

LABB 3: BAYESIANSK INFERENS

JOSE M. PEÑA, MÅNS MAGNUSSON, MATTIAS VILLANI
IDA, LINKÖPINGS UNIVERSITET

1. INSTRUKTIONER

- Laborationen ska göras **två och två**
- Labben ska vara en **PDF-rapport** med kod, analys och grafer. I rapporten ska följande ingå:
  - Båda studenternas namn och LiU-id.
  - Laborationsnummer
  - Uppgifterna ni besvarar (t ex som rubriker).
- Ett tips är att använda R markdown. Här finns en R markdownmall 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. Där hittar ni också **deadlinen**.

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 342-343 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 $p(\theta|\mathbf{y}) \propto p(\mathbf{y}|\theta)p(\theta)$. 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 344 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 en binomialfördelning med sannolikhet p.

(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. **Obs!** Tänk på att du ännu inte observerat några data ännu.

(2) Produkt A har du nu demonstrerat för 13 personer varav 8 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. 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 $E(X_1)$ och $E(X_2)$?

3.3. **Multinomialmodell med Dirichlet prior.** En generalisering av betafördelningen är Dirichletfördelningen och på samma sätt är multinomialfördelningen en generalisering av binomialfördelningen. Använd föreläsningsanteckningar om apriori och aposteriorifördelningen för multinomialfördelningen för att lösa dessa uppgifter.

3.3.1. *Analys av opinionsundersökningar.* I denna uppgift ska vi analysera väljarstödet för de olika partierna i den Svenska Riksdagen.

(1) Vi börjar med att försöka bestämma vår aprioriförderlning för de olika partierna. Vår aprioriförderlning specificerar vi som en Dirichletfördelning med parametrarna $\alpha_1, \ldots, \alpha_9$, eftersom vi betraktar nio partier (från Socialdemokraterna till Feministiskt Initiativ). Pröva dig fram och simulera från din prior. Dirichletfördelningen finns i R paketet `gtools`. Se exempel nedan.

```
> install.packages("gtools")
> library(gtools)
> set.seed(4711)
> rdirichlet(n = 3, alpha = c(1, 1.2, 0.2, 3, 2))

          [,1]     [,2]       [,3]    [,4]    [,5]
[1,] 0.304216 0.269486 0.00077692 0.18934 0.23618
[2,] 0.011199 0.062539 0.22542080 0.47626 0.22458
[3,] 0.228893 0.293479 0.01991985 0.21123 0.24648
```

Eftersom Dirichletfördelningen är multivariat görs en dragning för alla partier på en och samma gång. Gör 1000 dragningar från din aprioriförderlning och presentera din (marginella) prior för respektive parti som ett histogram. Välj parametrarna $\alpha_1, \ldots, \alpha_9$ så att utfallen från valet 2014 inte är allt för osannolikt enligt din aprioriförderlning. Använd `abline()` och argumentet `v` för att visualisera din valresultatet 2014 tillsammans med din (marginella) prior för samtliga partier.

(2) Här finns samtliga svenska opinionsundersökningar. Ladda ned eller använd en av de senaste undersökningarna. Det finns en hel del problem med opinionsundersökningar vilket gör att vi inte kan räkna baserat på institutens urvalsstorlekar direkt. Anta därför istället att den undersökning du valt har 200 personer som deltagit och räkna ut (och avrunda) hur många som svarat för respektive parti. Ange vilken undersökning du valt och dina (avrundade) antal observationer för respektive parti.

(3) Använd observationerna du fick i (2) för att beräkna aposteriorifördelningen för andelen per riksdagsparti och gör 10 000 simuleringar från din aposteriorifördelning. När du svarar på frågorna nedan ska du ta hänsyn till den så kallade 4 % spärren, dvs att ett parti måste få minst 4 % av rösterna för att sitta i Riksdagen. Får ett parti mindre än 4 % ska du sätta deras andel till 0 % för denna dragning och normalisera om proportionerna. Svara på följande frågor:

(a) Vad är sannolikheten att de rödgröna är större än alliansen ?
(b) Vad är sannolikheten att Sverigedemokraterna (SD) är större än Moderaterna (M) ?
(c) Vad är sannolikheten att Kristidemokraterna inte skulle komma in i Riksdagen, dvs de får mindre än 4 % ?
(d) Vad är sannolikheten att Miljöpartiet (MP) skulle åka ur Riksdagen ?
(e) Skapa ett sannolikhetsintervall (95 %) för Socialdemokraterna.