TBDA01: Laboration 3

Måns Magnusson, Mattias Villani

12 oktober 2015

Instruktioner

- Laborationen ska göras **två och två**
- Labben ska vara en **PDF-rapport** med kod, analys och grafer. Ett tips är att använda R markdown. I rapporten ska följande ingå:
  - Båda studenternas namn och LiU-id.
  - Universitet och institution
  - Laborationsnummer
- Deadlinen för laborationen framgår av **kurshemsidan**
- Registrera dig på **webreg** med följande länk:
  - https://www.ida.liu.se/webreg3/TDAB01-2015-1/LAB1
- Samtliga uppgifter ska skickas via e-post till **mans.magnusson@liu.se**
  - Ange rubriken `Labb [labnummer] för grupp [labbgrupp] - TBDA01` i meddelandet.

Innehåll

Kapitel 1

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**.

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**.

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**.

Kapitel 2

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 strheter, utan vi betraktar dem som okända stokastiska variabler.

2.1 Bayes sats och aposteriorifördelningen

Bayes sats ges av

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

Dock kan $p(\mathbf{y}) = \int p(\mathbf{y}|\theta)p(\theta)d\theta$ 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, d.v.s.

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

Se Baron [2013, s. 342-343] för exempel på en härledning av en onormaliserad sannolikhetsfunktion.

Uppgift 1 Visualisera posteriorn

Vi ska nu visualisera en lite klurigare posterior. Antag att dina data kommer från en normalfördelning där $\sigma = 1$ (d.v.s är känd). Du är intresserad av parametern μ och vill beräkna dess posterior. Som prior väljer du student-t fördelning med $\nu = 1$.

a) Visualisera din prior exakt (d.v.s. använd `dt()`) över intervallet $[-5, 15]$.

b) Härled den proportionella (onormaliserade) posteriorn, d.v.s. $p(\theta|\mathbf{y}) \propto p(\mathbf{y}|\theta)p(\theta)$. Tänk på att de faktorer som inte innehåller μ kan förkortas bort.

c) Nedan är 7 datapunkter som du observerat. Visualisera dessa som ett histogram på intervallet $[-5, 15]$.

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

Tips! Använd argumentet `xlim` i `hist()`.

d) Visualisera den onormaliserade posteriorn genom att beräkna värden för 1000 punkter på intervallet $[-5, 15]$ med funktionen du härlett i b) ovan. Visualisera den onormaliserade posteriorn med `plot(type='l')`.

2.2 Binomial - beta

Vi ska nu studera aposteriorifördelningen för p i en binomialfördelning. Mer information finns i Baron [2013, s. 344].

Uppgift 2 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. Produkt A har du demonstrerat för 13 personer varav 8 var intresserade och produkt B har ni bara haft möjlighet att demonstrera för tre personer och av dessa var två personer intresserade. Ni kommer initialt bara kunna skapa en av dessa produkter och kommer därför behöva välja vilken produkt ni ska satsa på.

a) För att det ska gå att dra slutsatser från materialet behöver du bestämma din prior som parametrar i en Betafördelning. Vilka parametrar väljer du och varför. Simulera och visualisera din prior i ett histogram, eller om du väljer två olika priorfördelning (en för varje produkt), visualisera båda.

b) Använd konjugategenskapen mellan beta och binomialfördelningen för att räkna ut din posteriorfördelning analytiskt i respektive fall. Beräkna analytiskt den förväntade proportionen för respektive produkt. Vilken produkt har den högsta förväntade proportionen intresserade?

c) Precis som när det gäller maximum likelihood kan vi studera maximum av vår posteriorifördelning, vilket ofta kallas maximum aposteriori skattningen (MAP). Ta reda på hur maximum beräknas för betafördelningen. Vad är MAP skattningen för respektive produkt?

d) Simulera och visualisera dina två posteriorfördelning ovan i ett histogram och svara på följande frågor:

1. Vad är sannolikheten att proportionen intresserade kunder är större för produkt A än produkt B?
2. Vad är sannolikheten att $P(p > 0.5)$ för respektive produkt?

e) Storleken på er marknad är 87 andra företag ni vill nå med era produkter. Använd era två aposteriorfördelningar för respektive produkt för att simulera hur många intresserade kunder ni kan tänkas få för respektive produkt. Simulera först från er betafördeling (posterior) och sedan använder ni de simulerade värdena p för att dra en binomialfördelad variabel med $\text{Bin}(n = 87, p)$. Visualisera fördelningenöver antale intresserade kunder ni kommer ha med respektive produkt.

1. Hur stor är sannolikheten att få fler än 40 intresserade kunder med respektive produkt?
2. Vad är det förväntade antalet intresserade kunder med respektive produkt?

f) Vilken produkt skulle ni satsa på och varför?

2.3 Normal - normal

I denna uppgift ska vi arbeta med normalfördelade data. Som framgår av Baron [2013, s. 344] är en normal prior konjugat för μ i en normalmodell.

Uppgift 3 Aktieanalys

Du ska bygga en handelsrobot som löpande bevakar aktieutvecklingen för olika aktier. Den dagliga avkastningen för en aktie beräknas med formeln

$$x_t = \frac{y_t - y_{t-1}}{y_{t-1}}$$

där y_t är aktievärdet vid dag t. Nedan följer den dagliga avkastningen för aktien SCA A under vecka 39.

```
[1]  0.0315 -0.0180 -0.0021 -0.0202  0.0076
```

Vi modellerar den dagliga avkastningen som oberoende normalfördelade slumpvariabler med väntevärde μ och varians σ^2. I denna uppgift gör vi det enkelt för oss, vi säger att σ^2 är känd men räknar ut den baserat på de datapunkter vi har. D.v.s. σ^2 är följande värde för SCA A:

```
[1] 0.00044654
```

a) Börja med att ange en prior för μoch ange också din osäkerhet med apriorivariansen τ^2. Simulera och visualisera var du tror, a priori, om den dagliga avkastningen för SCA A.

b) Använd dig av resultaten i Baron [2013, s. 344] för att räkna ut vår aposteriorifördelning för avkastningen efter att vi har observerat data. Vad är den förväntade avkastningen $E(\mu|Data)$, där $Data$ representerar observerad avkastning för SCA A under vecka 39?

c) Investerare är ofta intresserade av det som kallas Value-at-Risk (VaR) för en investering. VaR är ett mått på hur illa det kan gå vid en investering och defineras ofta som 1% percentilen i avkastningens fördelning. Man ser helt enkelt ett utfall i 1% percentilen som ett oturligt, men inte omöjligt, negativt utfall. Säg att er robot har investerat 100 000 kr i SCA A (eller ngn annan aktie). Vi vill nu veta vad Value-at-Risk är för denna investering per dag.
Simulera 10 000 dragningar från din posteriorfördelning för μ. Beräkna sedan den 1% percentilen för en normalfördelning med vårt simulerade μ och kända σ^2. Nu har du fått en fördelning för 1% percentilen, p_1. Använd detta för att beräkna ut fördelningen för Value-at-Risk, d.v.s.

$$\text{VaR} = |p_1 \cdot 100\ 000|$$

för varje simulerat värde för μ. Nu har du beräknat fördelningen för Value-at-Risk. Visualisera denna fördelning.

Tips! `quantile()`

d) Beräkna ett 95% sannolikhetsintervall för Value-at-Risk.

e) Din modell ovan antar att avkastningarna är oberoende över tid och att avkastningarna är normalfördelade. Diskutera om du tror att dessa antaganden är rimliga. Finns det något bra sätt att undersöka om dessa antaganden är rimliga?

2.4 Multinomial - dirichlet

En generalisering av betafördelningen är dirichletfördelningen och på samma sätt är multinomialfördelningen en generalisering av binomialfördelningen. Använd Mattias föreläsningsanteckningar om apriori och aposteriorifördelningen för multinomialfördelningen för att lösa dessa uppgifter.

Uppgift 4 Analys av opinionsundersökningar

Här finns samtliga svenska opinionsundersökningar.

a) Vi börjar med att försöka bestämma vår a priori-fördelning för de olika partierna. Vår apriorifördelning specificerar vi som en dirichletfördelning med en parameter α per parti. Pröva dig fram och simulera från din prior, dirichletfördelningen finns i R-paketet `gtools`. Eftersom dirichletfördelningen är multivariat görs en dragning för alla partier på en och samma gång. Gör 1000 dragningar från din apriorifördelning och presentera din prior för respektive parti. I din apriorifördelning bör utfallen från valet 2014 inte vara allt för osannolikt. Använd `abline()` och argumentet `v` för att visualisera din prior för respektive parti och valresultatet.

```
> install.packages("gtools")
```

```
> library(gtools)
> rdirichlet(n = 3, alpha = c(1,1.2,0.2,3,2))

          [,1]      [,2]     [,3]    [,4]     [,5]
[1,] 0.048050 0.0065094 0.095057 0.81934 0.031043
[2,] 0.179796 0.1697056 0.048327 0.32062 0.281553
[3,] 0.011985 0.0622557 0.260204 0.41314 0.252416
```

b) Ladda ned eller använd en av de senaste undersökningarna (och ange vilken du valt). Det finns en hel del problem med opinionsundersökningar (se **här** för en diskussion) 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.

c) Beräkna nu aposteriorifördelningen för andelen per riksdagsparti och gör 10 000 simuleringar från din aposteriorifördelning. När du svarar på frågorna nedan ska du ta hänsyn till de s k 4% spärren, dvs att ett parti måste få minst 4% av rösterna för att sitta i riksdagen. Svara sedan på följande frågor om det vore val idag:

1. Vad är sannolikheten att de rödgröna är större än alliansen?
2. Vad är sannolikheten att Sverigedemokraterna har en vågmästarroll, d.v.s. varken de rödgröna eller alliansen får egen majoritet?
3. Vad är sannolikheten att KD **inte** skulle komma in i Riksdagen (få mindre än 4 %)
4. Vad är sannolikheten att Fi skulle komma in i Riksdagen?
5. Vad är sannolikheten att FP skulle åka ur Riksdagen?
6. Vad är sannolikheten att SD skulle bli Sveriges största parti?
7. Skapa ett sannolikhetsintervall (95 %) för Socialdemokraterna.

Litteraturförteckning

Michael Baron. *Probability and statistics for computer scientists.* CRC Press, 2013.