MVE051/MSG810 2016 Föreläsning 1

Petter Mostad

Chalmers

November 1, 2016

Introduktion till statistik

- En vanlig definition av statistik: Sammanställningar och visualiseringar av data.
- Matematisk statistik: Ett sätt att göra vetenskapliga modeller för värkligheten som tar hänsyn till osäkerhet.
- Vetenskapliga modeller för värkligheten:
  - Beskriver en specificerad del av värkligheten.
  - Använder ofta ett matematisk ramverk.
  - Används för att *predikera* nya observationer: En modell är bättre än en annan om den predikerar bättre.
  - "All models are wrong but some models are useful" (G. Box)
- För många delar av värkligheten är det naturligt att prediktionerna görs *med osäkerhet* (Exempel "Det är 90% chans för soligt väder i morgon"). Modeller som gör prediktioner med osäkerhet kallas *stokastiska modeller* (eller statistiska modeller).
- Första delen (största delen) av vår kurs handlar om *sannolikhetsteori*, det matematiska språket för stokastiska modeller.

Vad är sannolikhet?

- Sannolikhetsteori är en väletablerad matematisk teori. Men när det gäller *tolkningen* av prediktioner från stokastiska modeller, så finns det olika åsikter, eller skolor. Exempel: Hur tolkas
  1. Sannolikheten för att denna geologiska formationen innehåller olja är 34%.
  2. Sannolikheten för lyckad hjärtoperation är ca. 96%.
- I en *subjektiv/personlig* eller *Bayesiansk* modell anser man att den stokastiska modellen modellerar en enskild persons *kunskap* om värkligheten. Dock på ett sådant sätt att två fornuftiga personer som har samma information bör få samma sannolikheter. Därmed kan man prata om sannolikheten för händelser som inte kan repeteras, som i (1) över.
- En *frekventistisk* synpunkt anser att bara händelser som på något sätt kan repeteras kan ges en sannolikhet; sannolikheten är då gränsen för *frekvensen* av händelsen när antalet repetitioner närmar sig oändligheten. I (2) kan 96% ha framkommit som en frekvens.

Hur kan man ta fram en sannolikhet?

- Vanligtvis hittar man en modell för värkligheten som man argumenterar är rimlig, och härleder sen sannolikheter från denna.
- Det enklaste exemplet:
  - Man anser att det finns ett ändligt antal möjliga utfall, och att varje av dessa utfall har samma sannolikhet.
  - Sannolikheten för en händelse som motsvarar en delmängd av de möjliga utfallen är summan av antalet utfall i delmängden delat på det totala antalet utfall.
- Exempel: Sannolikheten för att nästa kast med tärningen blir 4 är $1/6$, därför att det är 6 olika utfall, och alla borde ha samma sannolikhet.
- Exempel: Sannolikheten för att nästa kast med tärningen ger mindre än 4 är $1/2$, därför att denna händelse motsvarar 3 av 6 möjliga utfall.

Några viktiga begrepp

- Utfall (eng. outcome (eller "sample point")): En specifik möjlighet för den delen av värkligheten vi beskriver.
- Utfallsrum (eng. sample space): Samlingen av alla utfall som vi anser möjliga.
- Händelse (eng. event): En delmängd av utfallsrummet.
- Trädiagram (eng. tree diagram): Ett sätt att beskriva vissa utfallsrum.

Axiomer för sannolikhetsteori

Låt S vara ett utfallsrum. En sannolikhetsfördelning är en funktion som till alla ("mätbara") delmängder A av S tillordnar ett tal $\Pr[A]$, så att

- $\Pr[S] = 1$
- $\Pr[A] \geq 0$
- Låt $A_1, A_2, \dots, A_k, \dots$ vara händelser (delmängder av S) så att $A_i \cap A_j = \emptyset$ (de är icke-överlappande) om $i \neq j$. Då är

$$\Pr[A_1 \cup A_2 \cup \cdots \cup A_k \cup \dots] = \Pr[A_1] + \Pr[A_2] + \cdots + \Pr[A_k] + \dots$$

Några konsekvenser av axiomerna

- $\Pr[\emptyset] = 0$
- $\Pr[A'] = 1 - \Pr[A]$, där A' är komplementet till A i S (skrivs även A^c eller $S \setminus A$).
- Om A_1 och A_2 är två händelser så har vi

$$\Pr[A_1 \cup A_2] = \Pr[A_1] + \Pr[A_2] - \Pr[A_1 \cap A_2]$$

Betingade sannolikheter

- Den *betingade* sannolikheten för A_2 givet A_1, där $\Pr(A_1) > 0$, är definerad som
$$\Pr[A_2 \mid A_1] = \frac{\Pr[A_1 \cap A_2]}{\Pr[A_1]}$$
- Tolkning: OM vi vet att A_1 händer (eller: givet att A_1 händer), vad är då sannolikheten för A_2?
- Lär dig att skilja mellan $\Pr[A_1 \mid A_2]$ och $\Pr[A_2 \mid A_1]$!
- Från definitionen får vi att
$$\Pr[A_1 \cap A_2] = \Pr[A_1] \cdot \Pr[A_2 \mid A_1]$$

Viktiga räkneregler

- En viktig räkneregel (om $0 < \Pr[A_1] < 1$) blir

$$\begin{aligned}\Pr[A_2] &= \Pr[(A_2 \cap A_1) \cup (A_2 \cap A_1')] = \Pr[A_2 \cap A_1] + \Pr[A_2 \cap A_1'] \\ &= \Pr[A_2 \mid A_1] \cdot \Pr[A_1] + \Pr[A_2 \mid A_1'] \cdot \Pr[A_1']\end{aligned}$$

- Mera generellt, om $S = B_1 \cup B_2 \cup \cdots \cup B_k$ och $B_1, \ldots B_k$ är icke-överlappande (så $B_i \cap B_j = \emptyset$ om $i \neq j$) och $\Pr[B_i] > 0$ så har vi

$$\begin{aligned}&\Pr[A] \\ =\ &\Pr[(A \cap B_1) \cup (A \cap B_2) \cup \cdots \cup (A \cap B_k)] \\ =\ &\Pr[A \cap B_1] + \Pr[A \cap B_2] + \cdots + \Pr[A \cap B_k] \\ =\ &\Pr[A \mid B_1] \Pr[B_1] + \Pr[A \mid B_2] \cdot \Pr[B_2] + \cdots + \Pr[A \mid B_k] \Pr[B_k]\end{aligned}$$

Flera räkneregler

- $$\Pr[A \mid B \cap C] = \frac{\Pr[A \cap B \cap C]}{\Pr[B \cap C]} = \frac{\Pr[A \cap B \mid C]}{\Pr[B \mid C]}$$
- $$\Pr[A \cap B \mid C] = \Pr[A \mid B \cap C] \Pr[B \mid C]$$
- $$\begin{aligned} \Pr[A_1 \cap A_2 \cap A_3] &= \Pr[A_1] \Pr[A_2 \cap A_3 \mid A_1] \\ &= \Pr[A_1] \Pr[A_2 \mid A_1] \Pr[A_3 \mid A_1 \cap A_2] \end{aligned}$$
- ...

Oberoende händelser

- Två händelser A_1 och A_2 kallas oberoende om

$$\Pr[A_1 \cap A_2] = \Pr[A_1] \cdot \Pr[A_2]$$

- Vi får att A_1 och A_2 (med $\Pr[A_1] > 0$) är oberoende om och endast om

$$\Pr[A_2 \mid A_1] = \Pr[A_2]$$

- $A_1, A_2, \ldots, A_n$ är oberoende händelser om, för alla urval $A_{(1)}, A_{(2)}, \ldots, A_{(k)}$ av dessa gäller

$$\Pr[A_{(1)} \cap A_{(2)} \cap \cdots \cap A_{(k)}] = \Pr[A_{(1)}] \Pr[A_{(2)}] \ldots \Pr[A_{(k)}]$$

Bayes teorem

- För händelser A och B med $\Pr[A] > 0$ och $\Pr[B] > 0$ får vi direkt att

$$\Pr[A \mid B] = \frac{\Pr[B \mid A]\Pr[A]}{\Pr[B]} = \frac{\Pr[B \mid A]\Pr[A]}{\Pr[B \mid A]\Pr[A] + \Pr[B \mid A']\Pr[A']}$$

  Detta kallas Bayes formel.

- Mera generellt har vi, om $S = A_1 \cup A_2 \cup \cdots \cup A_k$, $A_1, \ldots A_k$ är icke-överlappande och $\Pr[A_j] > 0$, $\Pr[B] > 0$, så har vi

$$\Pr[A_j \mid B] = \frac{\Pr[B \mid A_j]\Pr[A_j]}{\Pr[B]} = \frac{\Pr[B \mid A_j]\Pr[A_j]}{\sum_{i=1}^{k}\Pr[B \mid A_i]\Pr[A_i]}$$

- Delar vi Bayes former för $\Pr[A \mid B]$ på formeln för $\Pr[A' \mid B]$ får vi

$$\frac{\Pr[A \mid B]}{\Pr[A' \mid B]} = \frac{\Pr[B \mid A]}{\Pr[B \mid A']} \cdot \frac{\Pr[A]}{\Pr[A']}$$