

Datorlaboration 2

Josef Wilzen

August 25, 2024

Allmänt

Datorlaborationerna kräver att ni har R och Rstudio installerat.

- Kodmanual: [länk](#)
- **ISL**: An introduction to Statistical Learning,
 - Boken: [länk](#)
 - R-kod till labbar: [länk](#)
 - Dataset: [länk](#) och [länk](#)
- **IDM**: Introduction to Data Mining,
 - Kod till boken: [länk](#)
 - Sample chapters
- Dataset till vissa uppgifter finns här

Notera att ni inte behöver göra alla delar på alla uppgifter. Det viktiga är att ni får en förståelse för de olika principerna och modellerna som avhandlats. Dessa uppgifter ska inte lämnas in, utan är till för er övning.

Datoruppdelning

För att motverka överanpassning bör ni dela upp data till träning-, validering-, (och testmängd). Detta kan göras med `createDataPartition()` från `caret`-paketet. Argument till den funktionen som är av vikt här är `p` som anger hur stor andel av observationerna som ska användas till träningsmängden. Ni kan också använda `subset()` för att göra detta, men det blir svårare att tydligt ange de observationer som ska tilldelas till valideringsmängden. Denna uppdelning måste vara slumpmässigt. Notera att om en testmängd ska skapas måste uppdelningen ske en gång till från valideringsmängden.

Del 1: Polynomregression och stegfunktioner

1. Gå igenom Lab 7.8.1 i ISL
2. Vad är skillnaden mellan följande tre funktioner för polynomregression?

```
y ~ poly(x, 4)K
y ~ poly(x, 4, raw = TRUE)
y ~ x + I(x^2) + I(x^3) + I(x^4)
```

3. Läs in datamaterialet `lab2_data_1.csv`. Dela upp i träning och validering och använd korsvalidering för att skatta den bästa polynomregression och stegfunktionen till detta datamaterial. Vilken grad av polynom används? Hur många stegfunktioner används?

Del 2: Splines

1. Gör Lab 7.8.2 i ISL
2. Ladda in datamaterialet `lab6_data_1.csv`. Välj knutarna i 1, 2, 3, och 4. Testa nu att skatta vanliga splines (`bs()`) med `degree` 1, 2, 3 och 4. Vad blir skillnaden i resultat? Vilken ser bäst ut?
3. Använd samma datamaterial och ändra till natural splines (`ns()`). Vad blir skillnaden? Lägg till knutar i 0.5 och 4.5 och jämför.
4. Testa andra platser för knutarna och använd korsvalidering för att hitta den bästa modellen.

Del 3: GAM

1. Gör Lab 7.8.3 i ISL

Del 4: Email Spam

1. Ladda in `spambase.csv` datasetet och bekanta dig med det (se bilaga för mer info). Vi vill skapa en modell som predikterar **spam** (0 eller 1) givet de förklarande variablerna. Då våra förklarande variabler är tungsvansade kan en log-transformation fungera bra ($\log(x + 0.1)$) där vi lägger på 0.1 för att undvika $\log(0)$ problem.
2. Dela upp datamaterialet i 70% träningsdata och 30% valideringsdata.
3. Börja med att anpassa en vanlig multiple logistisk regression till datamaterialet. Vilket klassificeringsfel får du?
4. Anpassa nu en GAM-modell med kubiska natural splines med 4 frihetsgrader för varje förklarande variabel. Vilket klassificeringsfel får du nu?
5. Testa med andra ordningar av frihetsgrader för dina splines och se om du kan få till en bättre modell.
6. Testa att använda lokal regression för någon eller några variabler och se hur det påverkar dina resultat.

d