Bagging og Random Forests

Peter N. Bakker

08-06-2023

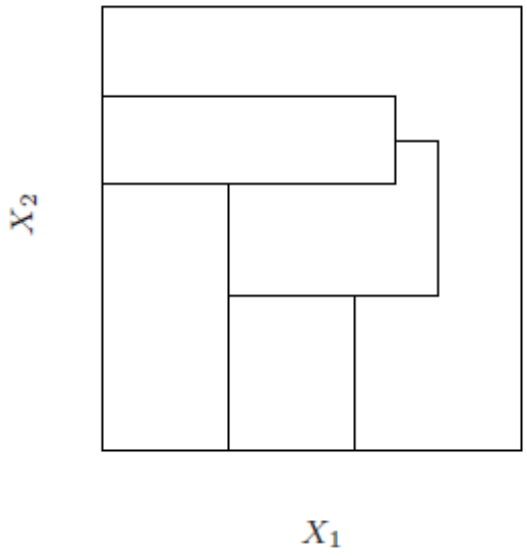
Træer

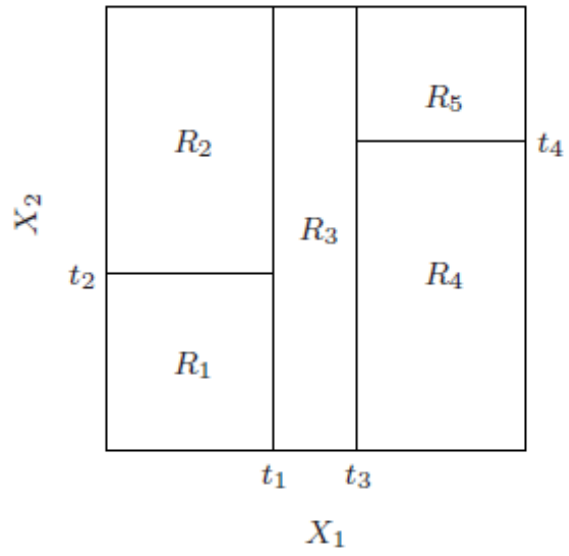
- Træer er en split-baseret (partitioning) metode til at indele datasættet i subset. De er ikke-parametriske.

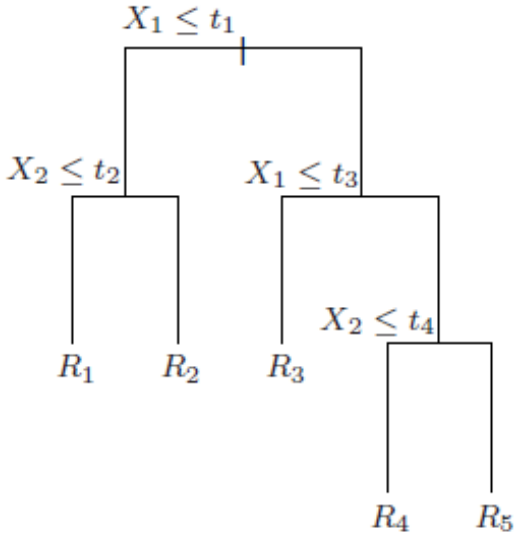
$$f(x) = \sum_{m=1}^{M} c_m I(x \in R_m)$$

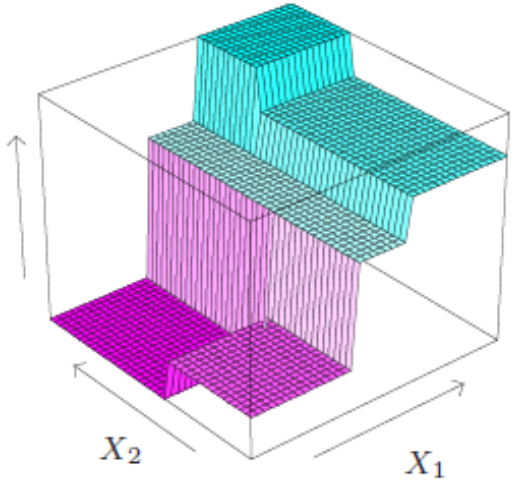
- Struktur af et træ:
  - Hver node repræsenterer et spørgsmål eller en test på en variabel.
  - Hvert split repræsenterer et mulig udfald af testen, hvor man så går videre i træet (eller rammer en leaf node).
  - Hver leafnode repræsenterer en forudsigelse (regression) eller en klassifikation.
- Træer kan nemt tilpasses og er lette at fortolke.
- Meget tilbøjelig til overfit (lav bias og høj varians).

Illustration af et træ









Growing

- Growing (vækst) af træer:
  - Start med et enkelt leaf node og gror træet herefter. Træet gros pba. at minimere SSR på tværs af variablene (Her vises for regressionstræer, men kan også overføres til klassifikation):

$$R_1(j,s) = \{X \mid X_j \leq s\} \quad \text{og} \quad R_2(j,s) = \{X \mid X_j > s\}$$

$$\min_{j,s} \left[\min_{c_1} \sum_{x_i \in R_1(j,s)} (y_i - c_1)^2 + \min_{c_2} \sum_{x_i \in R_2(j,s)} (y_i - c_2)^2 \right]$$

  - Man fortsætter med at gro træet (splitte det) indtil man rammer et stop kriterie.
  - Stopkriterier kan f.eks. være nået maksimal dybde i træet eller minimum antal datapunkter i en node. Stopkriterier er også en form for hyperparametre.

Pruning

- Pruning er en teknik til at reducere kompleksiteten af træer ved at fjerne unødvendige branches (mindsker varians).
- Cost Complexity Criterion er en metode til at evaluere og vælge den optimale pruningsstrategi.
- Kombinerer træets kompleksitet med træets prædiktive præstation.
- Formlen er givet ved:

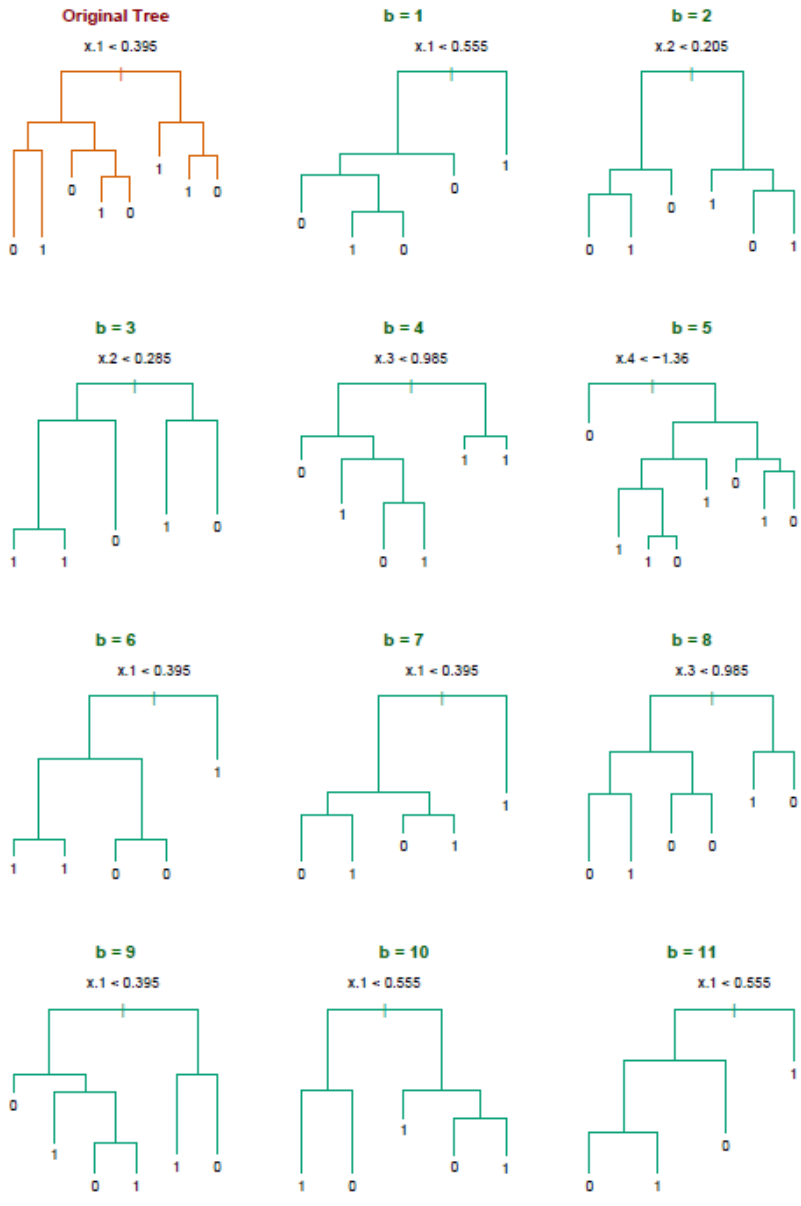
$$C_\alpha(T) = \sum_{m=1}^{|T|} N_m Q_m(T) + \alpha|T|$$

- Første led kan ses som SSR ved træerne og $\alpha|T|$ er et mål for kompleksitet. (Minder lidt om regularisering)
- Ved at variere α kan vi opnå en trade-off mellem træets størrelse og præstationsniveauet (hyperparameter).

Bagging

- Selvom vi beskærer et træ er de enkeltvis stadig meget sårbare for overfit (høj varians).
- Bagging (bootstrap aggregation) er en teknik til at forbedre træmodellers præstation.
- Ideen bag bagging:
  - Generer flere træmodeller ved at trække tilfældige stikprøver med tilbagelægning fra det oprindelige datasæt (Bootstraping), **med alle features**.
  - Træn hvert træ på en stikprøve og få en prædiktion ved at kombinere forudsigelserne fra alle træer (i klassifikation flertal ved regression gennemsnit af forudsigelserne). (Aggregation).
- Bagging reducerer varians og kan føre til mere stabil og robust prædiktion.
- Antallet af træer er en hyperparameter.

Bagging illustration (En skov af træer)



Random Forest

- Random Forest er en variant af bagging, der yderligere introducerer variabilitet i konstruktionen af træerne.
- I stedet for at bruge alle features til at dele en node, vælges kun et tilfældigt udvalg af features.
- Random Forest kombinerer fordelene ved bagging og øger variabiliteteten yderligere for at reducere overfitting og forbedre prædiktionsevnen. Og dermed mindskes korrelationen mellem træerne i skoven fordi de er tilfældige (kan være både eg, bøg og palmer).
- Variansen ved B i.d. bootstraps: $\rho\sigma^2 + \frac{1-\rho}{B}\sigma^2$.
- Her kan RF modsat bagging gøre mere ved $\rho\sigma^2$ leddet.
- Det viser sig ofte at RF foretrækkes fremfor ren bagging.

Out-of-Bag Sampling

- Random Forests anvender out-of-bag (OOB) sampling til at estimere modellens præstation uden behov for et validerings-split under træning.
- OOB-samples er de datapunkter $z_i = (x_i, y_i)$, der ikke er inkluderet i bootstrap-samplet for hver træmodel.
- For hvert z_i prædikterer modellen ved kun at ensemble de træer, hvor z_i ikke indgik i deres bootstrap-sample.
  Støjvariable:
- Ulempen ved RF er at den tilfældigt vælger sine variable. Hvis kun nogle få variable er relevante vil mange af de anvendte variable aggere som støj.
- Derfor kan inkludering af disse "støjvariable" øge variansen for ens model.

Dataeksempel: IGF1 (Inslulin-like Growth Factor 1)

- Vi ser på et datasæt af børn og unge mellem 6-20 år af begge køn og deres $\sqrt{igf1}$.

Dataeksempel: IGF1 (Inslulin-like Growth Factor 1)

- Jeg fitter en RF model med alder og køn som prædiktorer. Datasættet har i alt 810 observationer med hhv. 357 drenge og 453 piger.
- Jeg ligger 600 i træning og 210 i test (ingen behov for validerings-split). Antallet af træer er lig 500.
- Resultater:

```
##      |      Out-of-bag   |
## Tree |      MSE  %Var(y) |
##  100 | 1.616e+04    58.88 |
##  200 | 1.609e+04    58.60 |
##  300 | 1.588e+04    57.87 |
##  400 | 1.607e+04    58.53 |
##  500 | 1.604e+04    58.43 |

## MSE for boys: 7.385033

## MSE for girls: 9.276394
```

Dataeksempel: IGF1 (Inslulin-like Growth Factor 1)