Splines og GAM

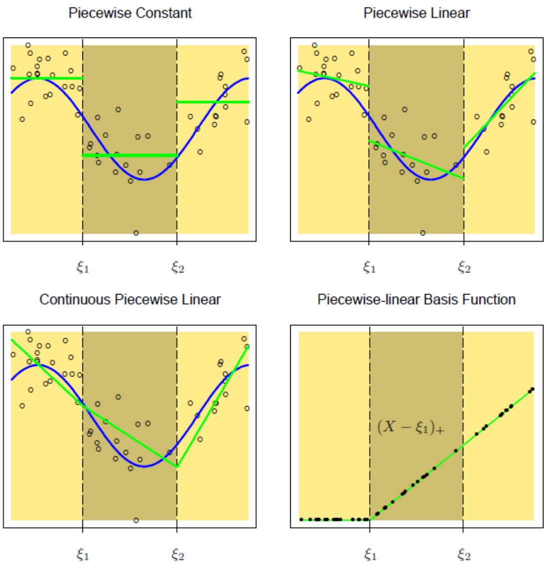
Peter N. Bakker

08-06-2023

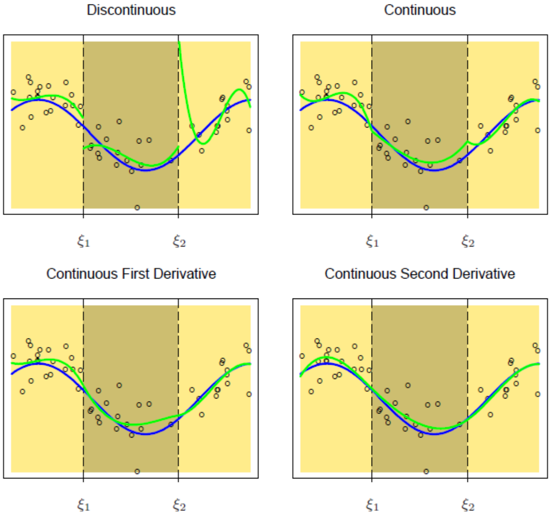
Basisfunktioner og stykkevis polynomier.

- En spline består af en lineær kombination af en række basisfunktioner: $f(X) = \sum_{m=1}^{M} \beta_m h_m(X)$
- En basis funktion er en funktion, der er defineret ved hjælp af separate segmenter inden for forskellige intervaller på X-aksen ved knudepunkter ξ_i.
- Disse segmenter er sammensat for at danne funktion på hele domænet af X (antager at X er en dimensionel her).
- Disse basisfunktioner kan komme i mange forskellige varianter. Og vil i splines typisk være i form af lavere ordens polynomier (dvs. kontinuert i typisk første og anden orden).

Basisfunktioner illustration



Stykkevis polynomier illustration



Kubiske Splines

- Kubiske splines er en type af stykkevis polynomiale funktioner.
- Hvert interval mellem knudepunkterne er tilpasset med et kubisk polynomie.
- Funktionen er glat, da der er kontinuitet i både værdier og første og anden afledede.
- **Naturlige splines:** Lineære i yder intervallerne (dvs. $f''(\xi_1) = f''(x_k) = 0$).
- Dermed for en **Naturlig kubisk spline** har vi basisfunktionerne:
  - $N_1(X) = 1$
  - $N_2(X) = X$
  - $N_{k+2}(X) = d_k(X) - d_{K-1}(X)$
  - $d_k(X) = \frac{(X-\xi_k)_+^3 - (X-\xi_K)_+^3}{\xi_K - \xi_k}$
- Denne linearitet i yderpunkterne frigiver 4 frihedsgrader, som så måske kan bruges bedre ved at indsætte flere knudepunkter ξ_i i de indre intervaller. Prisen er dog også større bias i de ydre intervaller (men den handel tager vi gerne, da vi ofte ikke har meget information her alligvel).

Smoothing Splines

En smoothing spline er en funktion $f(x)$, der opfylder følgende:

$$\text{RSS}(f, \lambda) = \sum_{i=1}^{N} (y_i - f(x_i))^2 + \lambda \int \left(f''(t)\right)^2 dt$$

- λ er en strafparameter, der kontrollerer *smoothness*.
- Den første del af målfunktionen er residual sum of squares (RSS) og måler den lokale pasning af punkterne.
- Den anden del er en glatningsstraf, der straffer store ændringer i anden afledede og fremmer glatte løsninger.

Bemærk: Ved at minimere RSS med en strafparameter λ, får vi en glatningsspline, hvor løsningen er lineær i $\mathbf{y}$.

$$\hat{\theta} = (X^T X + \lambda K)^{-1} X^T y$$

(Generaliseret ridge regression).

GAM

- GAM er en udvidelse af MLR ved at tilføje ikke-lineære funktioner af X_i.

$$E(Y|X_1, \ldots, X_p) = \alpha + f_1(X_1) + \ldots + f_p(X_p)$$

- Meget lig regression splines, men uden parametrisk form på f_j.
- En GAM må gerne både have parametrisk og ikke-parametrisk funktionsled f.eks. $g(\mu) = X^T\beta + \alpha_k + f(Z)$, *semiparametrisk*.
- GAM fittes ved *backfitting*.
- GAM kan stadig bibeholde den "fortolkning" vi har fra lineære modeller.

Backfitting-algoritme

Input: Data $(x_i, y_i), i = 1, 2, ..., N$, antal prædiktorer p

1. Initialisér $\hat{\alpha} = \frac{1}{N}\sum_{i=1}^{N} y_i$, $\hat{f}_j \equiv 0$, for alle i og j
2. Gentag indtil konvergens:
   - For $j = 1, 2, ..., p$:
     2.1 Opdater responsvariablen: $\hat{f}_j = y_i - \hat{\alpha} - \sum_{k \neq j} \hat{f}_k(x_{ik})$
     2.2 Opdater glat funktion $\hat{f}_j$ ved at løse:

$$\hat{f}_j(x_{ij}) = S_j\left(y_i^* - \frac{1}{N}\sum_{i=1}^{N} \hat{f}_j(x_{ij})\right)$$

3. Output: Den endelige model $\hat{f}(x) = \hat{\alpha} + \sum_{j=1}^{p} \hat{f}_j(x_j)$

- S_j er en naturlig kubisk spline (som regel).
- Backfitting ignorerer effekten af de andre variable i hvert step. Det kan giver problemer, hvis X_j er nogenlunde korrelerede.
- Kan også laves for klassifikation ved "Addititive Logistic Regression Model". F.eks. binært Y, logit link $g(\mu) = \log(\frac{\mu}{1-\mu})$.

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 smoothing spline for hvert køn. Jeg fitter også 95% konfidensinterval.
- MSE for piger er 8.58 og 8.45 for drenge.