

Polya tree priors

Begeleider: Aad van der Vaart

Veronderstel dat $X_1, X_2, \dots, X_n$ onafhankelijke stochastische variabelen zijn met verdelingsfunctie $F(x) = P(X_i \leq x)$. Als geen informatie over de vorm van F beschikbaar is, dan is een voor de hand liggende schatter de *empirische verdelingsfunctie*

$$\mathbb{F}_n(x) = \frac{1}{n} \sum_{i=1}^n 1_{X_i \leq x}.$$

Er zijn echter veel andere mogelijke schatters. Voorstanders van de *Bayesiaanse statistiek* geven de voorkeur aan een schatter die gedefinieerd is door een *a priori verdeling*. Dit is in dit geval een kansverdeling Π op de verzameling van alle verdelingsfuncties. Het Bayesiaanse paradigma is dan:

- Eerst wordt F bepaald volgens Π .
- Vervolgens zijn $X_1, \dots, X_n$ gegeven F voorwaardelijk verdeeld volgens F .
- Dit geeft de gezamenlijke kansverdeling van $(F, X_1, \dots, X_n)$.
- Met behulp van (een uitbreiding van) de regel van Bayes vinden we de voorwaardelijke verdeling van F gegeven $X_1, \dots, X_n$.

De verdeling in de laatste stap is de *a posteriori verdeling*. We kunnen hiervan bijvoorbeeld de verwachting nemen, als alternatief voor $\mathbb{F}_n$.

Zowel de *a priori* als de *a posteriori* verdeling zijn kansverdelingen op de verzameling van verdelingsfuncties: verdelingen van *stochastische verdelingsfuncties*. In zijn algemeenheid zijn dit ingewikkelde objecten. Polya tree priors zijn een speciaal geval, en kunnen worden geconstrueerd met het volgende algoritme.

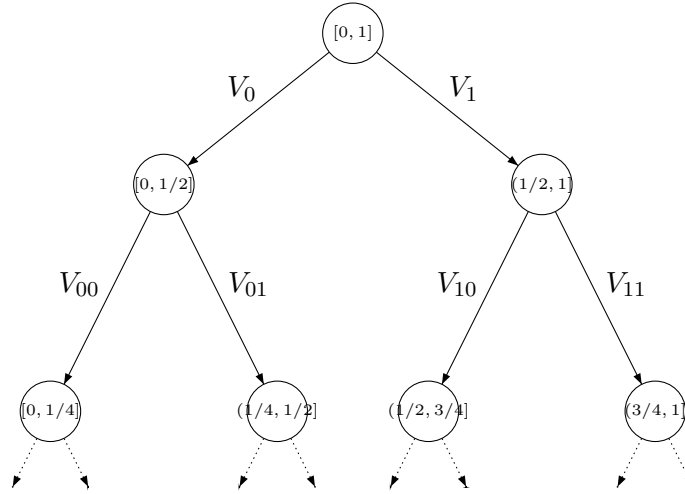
De waarde $F(x)$ van een stochastische verdelingsfunctie F is voor ieder $x \in \mathbb{R}$ een stochastische variabele met waarden in $[0, 1]$. Uiteraard is $F(-\infty) = 0$ en $F(\infty) = 1$, met kans 1. Een standaard kansverdeling op $[0, 1]$ is de $\text{Beta}(a, b)$ verdeling, gegeven door de kansdichtheid

$$x \mapsto \frac{1}{B(a, b)} x^{a-1} (1-x)^{b-1}.$$

Laat nu voor ieder rijtje $\epsilon_1 \epsilon_2 \dots \epsilon_k \in \{0, 1\}^k$ de variabele $V_{\epsilon_1 \epsilon_2 \dots \epsilon_k}$ $\text{Beta}(a, b)$ verdeeld zijn voor $a = b$ gelijk aan gegeven constanten $a_{\epsilon_1 \epsilon_2 \dots \epsilon_k}$, waarbij de variabelen $V_{\epsilon_1 \epsilon_2 \dots \epsilon_{k-1} 0}$ onafhankelijk zijn en $V_{\epsilon_1 \epsilon_2 \dots \epsilon_{k-1} 0} + V_{\epsilon_1 \epsilon_2 \dots \epsilon_{k-1} 1} = 1$. De Polya tree prior met *homogene base measure* geeft kansmassa

$$\begin{aligned} F[0, 1/2] &= V_0, & F(1/2, 1] &= V_1, \\ F[0, 1/4] &= V_0 V_{00}, & F(1/4, 1/2] &= V_0 V_{01}, & F(1/2, 3/4] &= V_1 V_{10}, & F(3/4, 1] &= V_1 V_{11}, \\ F[0, 1/8] &= V_0 V_{00} V_{000}, & & & & & \text{etc.} \end{aligned}$$

In andere woorden de V -variabelen zijn de voorwaardelijke kansen op de linker en rechter helft van een interval van gegeven een interval. De totale massa 1 wordt uitgesmeerd over



het interval $[0, 1]$ als in de boomstructuur in de figuur. De niet-homogene Polya tree prior wordt verkregen door het interval $[0, 1]$ naar $\mathbb{R}$ af te beelden met een kwantielfunctie.

Polya verdelingen zijn in de jaren 1970-80 bestudeerd, en we weten bijvoorbeeld:

- Een conditie op $a_{\epsilon_1 \epsilon_2 \dots \epsilon_k}$ die maakt dat de procedure een goed gedefinieerde random verdelingsfunctie oplevert.
- Als $a_{\epsilon_1 \epsilon_2 \dots \epsilon_k} = a_k$ en $\sum_k a_k^{-1} < \infty$, dan heeft bijna iedere realisatie van deze random verdelingsfunctie een dichtheid.
- De a posteriori verdeling heeft dezelfde vorm, maar met ‘updated’ parameters a .
- Als $a_{\epsilon_1 \epsilon_2 \dots \epsilon_k} = a_k$ en $\sum_k a_k^{-1} < \infty$, dan geeft de a posteriori verdeling asymptotisch een zwak consistente reconstructie; als $a_k^{-1} \leq 2^{-k}$, dan is de reconstructie ook sterk consistent.

Het project zou zich kunnen richten op het precies maken en bewijzen van deze beweringen. Simulaties (in R?) zouden de procedures inzichtelijk kunnen maken. Als nieuw aspect kan worden onderzocht, of splits met andere dan Beta variabelen (bijvoorbeeld met variabelen die met positieve kans $1/2$ zijn) betere statistische resultaten geeft.