LATVIJAS UNIVERSITATE
ANNO 1919

Darba autore: Agrita Garnizone
a.garnizone@gmail.com
Darba vadītājs: Dr.sc.comp. Kārlis Freivalds

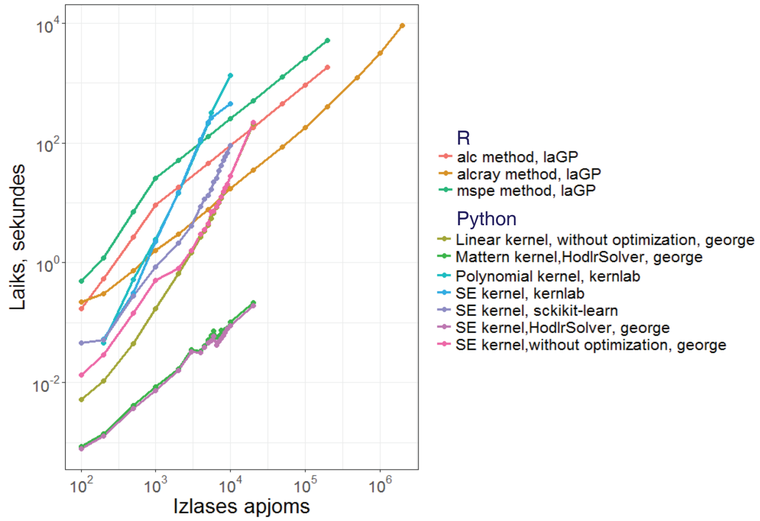
Gausa procesu regresija lielām datu kopām

Ievads

Datu struktūras atpazīšana pie dotiem empīriskiem datiem ir labi pazīstama kā uzraudzītās apmācības problēma, kur galvenais uzdevums ir spēja precīzi atpazīt attēlojums starp novērojumu raksturojošām vērtībām un mainīgo vērtību. Gausa procesi tiek plaši izmantoti gan regresijas, gan klasifikācijas, gan datu kopas dimensijas samazināšanas problēmas risināšanā.

Gausa procesiem pēdējos gados tiek pievērsta pastiprināta uzmanība to spēcīgo īpašību un prognozes precizitātes dēļ, taču viens no lielākajiem trūkumiem ir to skaitļošanas laiks lielām datu kopām. Ja datu kopas apjoms ir N, tad modeļa apmācīšanas laiks ir $O(N^3)$, bet prognozes veikšana $O(N^2)$.

Līdz ar tehnoloģisko attīstību un jaunām skaitļošanas iespējam, tiek pastiprināti meklēti risinājumi šī brīža Gausa procesa modeļu ierobežojumiem, lai šīs metodes spēcīgās īpašības, var izmantot arī lielām datu kopām.



Kursa darba mērķi

- veikt padziļinātu izpēti Gausa procesa regresijas teorētiskajā pamatojumā, pievēršot uzmanību biežāk izmantotajām kovariāciju funkcijām un tās aproksimācijas iespējām;
- aplūkot šī brīža Gausa procesa regresijas praktiskās iespējas un ierobežojumus prognozes veikšanā valodā R un Python.

Secinājumi

Tāpat kā līdz šim tas ir pierādījies mazām datu kopām, varbūtiskie modeļi spēj dot ļoti labu prognozi arī lielām datu kopām, kas ir galvenā motivācija meklēt iespējas pielāgot Gausa modeļus arī lielām datu kopām.

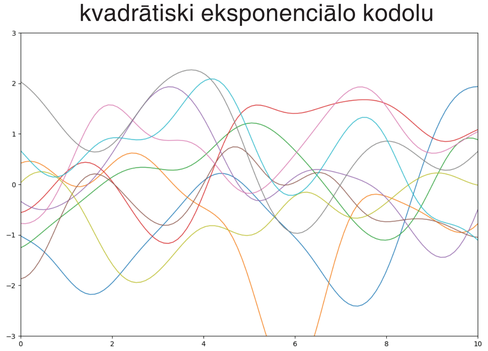
Pētījuma rezultātā novērots, ka šī brīža ierobežojumi tiek risināti divos virzienos, - vai nu dati tiek ņemti mazās daļās, veicot lokālas aproksimācijas, vai tiek veidoti kovariāciju funkcijas tuvinājumi ar lineārās algebras palīdzību.

Aplūkojot šī brīža praktiskās iespējas, tiek secināts, ka gan R, gan Python jau eksistē bibliotēkas speciāli paredzētas Gausa procesu regresijas aproksimācijām, uzlabojot skaitļošanas laiku, tomēr lielākā daļa no tām joprojām ir aktīvā izstrādes stadijā.

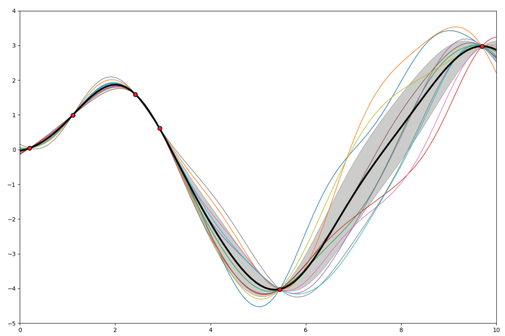
Turpmākais darbs

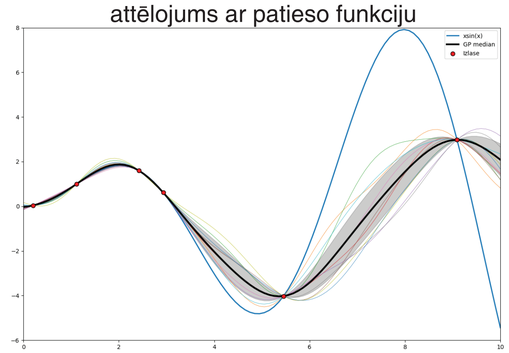
Tā kā iedziļinoties teorētiskajā pamatojumā un praktiski izmēģinot šī brīža iespējas, tēma ar pētījumu nav izsmelta, tādēļ kā loģiskais turpinājums tiks pētītas padziļināti metodes, ar kurām šobrīd tiek aproksimētas kovariāciju funkcijas, ar mērķi izveidot precīzu prognozēšanas modeli, kas piemērots lielai datu kopai.

Aprioro funkciju attēlojums, kas veidots uz kvadrātiski eksponenciālo kodolu



pievienojot dažus izlases novērojumus, ierobežojot funkciju



ja netiek veikti jauni mērījumi, tad tiek iegūts gala regresijas modelis, ar ko veikt prognozi

Aposteriorās sadalījuma funkcijas attēlojums ar patieso funkciju