

UNIVERZA V LJUBLJANI
FAKULTETA ZA MATEMATIKO IN FIZIKO

Finančna matematika – 1. stopnja

Sara Kovačič

Regresija z Gaussovimi procesi

Delo diplomskega seminarja

Mentor: doc. dr. Aljoša Peperko

Ljubljana, 2019

KAZALO

| | |
|---------------------------------------|---|
| 1. Uvod | 4 |
| 2. Osnove Bayesove statistike | 4 |
| 3. Gaussovi procesi | 4 |
| 4. Regresija | 5 |
| 4.1. Standardni linearni model | 5 |
| 4.2. Pogled iz prostora funkcij | 6 |
| 5. Kovariančne funkcije | 8 |
| 5.1. Stacionarne kovariančne funkcije | 8 |
| Slovar strokovnih izrazov | 9 |
| Literatura | 9 |

Regresija z Gaussovimi procesi

POVZETEK

V povzetku na kratko opiši vsebinske rezultate dela. Sem ne sodi razlaga organizacije dela – v katerem poglavju/razdelku je kaj, pač pa le opis vsebine.

Angleški naslov dela

ABSTRACT

Prevod zgornjega povzetka v angleščino.

Math. Subj. Class. (2010): navedi vsaj eno klasifikacijsko oznako – dostopne so na www.ams.org/mathscinet/msc/msc2010.html

Ključne besede: navedi nekaj ključnih pojmov, ki nastopajo v delu

Keywords: angleški prevod ključnih besed

1. UVOD

Strojno učenje je veja umetne inteligence, ki se uporablja za analizo podatkov in odkrivanje zakonitosti v podatkovnih bazah. Osnovi princip strojnega učenja je modeliranje pojavov iz podatkov, rezultat učenja pa so lahko pravila, funkcije, relacije, sistemi enačb, verjetnostna porazdelitev ipd. V diplomski nalogi se bom ukvarjala z modeliranjem z Gaussovimi procesi ali krajše GP. Model GP je neparametričen, kar pomeni, da neznanega sistema ne poskuša opisati s prilagajanjem parametrov baznih funkcij, ki sestavljajo model. Sestavljen je iz vhodno-izhodnih podatkov, ki opisujejo obnašanje opisovanega sistema in jih model uporablja za napovedovanje, in kovariančne funkcije, ki pove, v kakšni medsebojni odvisnosti so podatki. Izhod GP modela je verjetnostna porazdelitev v obliki Gaussove porazdelitve, pri čemer je srednja vrednost najbolj verjetna vrednost izhoda, varianco pa interpretiramo kot zaupanje v to napoved. Glede na karakteristiko izhoda, razlikujemo med klasifikacijskimi problemi (kjer je izhod diskreten) in regresijskimi problemi (kjer imamo zvezen izhod (največkrat je to neka funkcija)). V splošnem označimo vhod kot x in izhod kot y . Vhod predstavimo kot vektor x , saj imamo ponavadi več vhodnih spremenljivk. Izhod, pravimo tudi ciljna spremenljivka, y , je lahko zvezen (primer regresije), lahko pa diskreten (klasifikacija). Imamo množico podatkov (dataset) $D = \{(x_i, y_i) | i = 1, \dots, n\}$. S temi vhodnimi podatki x želimo ustvariti napoved za neke nove vhodne podatke x^* , ki niso bili v učni množici.

2. OSNOVE BAYESOVE STATISTIKE

Definicija 2.1. Fiksirajmo dogodek $B \in \mathcal{F}$, kjer je \mathcal{F} σ -algebra na verjetnostnem prostoru (Ω, \mathcal{F}, P) . Naj velja $P(B) > 0$. Potem je pogojna verjetnost dogodka A pri pogoju B enaka $P(A|B) = \frac{P(A \cap B)}{P(B)}$.

Sedaj imejmo poskus, ki ga opravimo v dveh korakih. V prvem koraku se zgodi natančno eden od paroma (končno ali števno mnogo) nezdružljivih dogodkov H_1, H_2, \dots . V drugem koraku nas zanima dogodek A . Ker je $\{H_1, H_2, \dots\}$ popoln sistem dogodkov, je $A = A \cap (\bigcup_i H_i) = \bigcup_i (A \cap H_i)$. Zaradi števne oz. končne aditivnosti verjetnosti sledi $P(A) = P(\bigcup_i (A \cap H_i)) = \sum_i P(A \cap H_i) = \sum_i P(H_i) \cdot P(A|H_i)$, kjer smo na zadnjem koraku upoštevali definicijo pogojne verjetnosti.

Definicija 2.2. (Bayesova formula) $P(H_k|A) = \frac{P(A \cap H_k)}{P(A)} = \frac{P(H_k) \cdot P(A|H_k)}{\sum_i P(H_i) \cdot P(A|H_i)}$.

Opomba 2.3. V Bayesovi statistiki se $P(H_k|A)$ imenujejo **posteriorne verjetnosti**, verjetnosti $P(H_k)$ pa **apriorne**. Na Bayesovi formuli temelji cela veja statistike, ki ji pravimo Bayesova statistika. Ta se razlikuje od klasične inferenčne statistike po tem, da potrebuje apriorne verjetnosti za tisto, kar nas zanima. Klasična statistika ne privzema nikakršnih apriornih verjetnosti, zato pa so tudi njeni rezultati šibkejši narave.

3. GAUSSOVI PROCESI

Definicija 3.1. Slučajni proces je zaporedje slučajnih spremenljivk $(X_t)_{t \geq 0}$.

Definicija 3.2. Slučajni proces $(X_t)_{t \in T}$ je Gaussov, če je za katerokoli končno podmnožico $F \subset T$, slučajni vektor $X_F := (X_t)_{t \in F}$ (večrazsežno) normalno porazdeljen.

Opomba 3.3. Slučajni proces je Gaussov, če je za vsak vektor neodvisnih spremenljivk x , vrednost funkcije $f(x)$ porazdeljena po normalni (Gaussovi) porazdelitvi.

4. REGRESIJA

Za interpretacijo regresije z Gaussovimi procesi (GP) obstaja več načinov. Gaussove procese si lahko predstavljamo kot definiranje porazdelitev za funkcije, kjer inferenca poteka direktno v prostoru funkcij. Ta način razlage je zelo privlačen, vendar težji za razumeti, zato se najprej osredotočimo na Weight-space view (utežen prostor??).

4.1. Standarni linearni model. Za lažje razumevanje si najprej pogledajmo Bayesovo analizo standardnega linearnega modela regresije z Gaussovim šumom

$$f(x) = x^\top w, \quad y = f(x) + \epsilon,$$

kjer je x vhodni vektor, w je vektor uteži (parametrov) linearnega modela, f je funkcija in y opazovana izhodna (ciljna) vrednost. Predpostavimo, da je šum ϵ neodvisno in enako normalno porazdeljen s povprečjem 0 in varianco σ_n^2 .

$$\epsilon \sim \mathcal{N}(0, \sigma_n^2).$$

Predpostavka o šumu skupaj z modelom pripelje do verjetnostne gostote (nevem se izraziti?):

$$\begin{aligned} (1) \quad p(y|X, w) &= \prod_{i=1}^n p(y_i|x_i, w) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp\left(-\frac{(y_i - x_i^\top w)^2}{2\sigma_n^2}\right) \\ &= \frac{1}{(2\pi\sigma_n^2)^{n/2}} \exp\left(-\frac{1}{2\sigma_n^2} |y - X^\top w|^2\right) \sim \mathcal{N}(X^\top w, \sigma_n^2 I). \end{aligned}$$

V Bayesovem formalizmu moramo določiti predhodne parametre, ki izražajo naša predhodna (apriorna) prepričanja o parametrih, preden pogledamo opazanja. Predpostavimo, da so uteži porazdeljene normalno, s povprečno vrednostjo 0 in kovariančno matriko Σ_p

$$(2) \quad w \sim \mathcal{N}(0, \Sigma_p).$$

Sklepanje v Bayesovem linearnem modelu temelji na posteriorni porazdelitvi nad utežmi, izračunanimi po Bayesovem pravilu,

$$(3) \quad \text{posterior} = \frac{\text{verjetje} \cdot \text{prior}}{\text{robno verjetje}}, \quad p(w|y, X) = \frac{p(y|X, w) \cdot p(w)}{p(y|X)}.$$

Robno verjetje je neodvisno od uteži in podano kot

$$(4) \quad p(y|X) = \int p(y|X, w) p(w) dw.$$

Posterior iz enačbe 3 združi verjetje in prior in zavzame vse, kar vemo o parametrih.

Definicija 4.1. Pravimo, da je funkcija f sorazmerna g , če je $f(x) = k \cdot g(x)$ za poljubno konstanto k , neodvisno od x . Označimo $f \propto g$.

Če v enačbo 3 vstavimo enačbi 1 in 2, dobimo:

$$\begin{aligned} (5) \quad p(w|X, y) &\propto \exp\left(-\frac{1}{2 \cdot \sigma_n^2} (y - X^\top w)^\top (y - X^\top w)\right) \cdot \exp\left(-\frac{1}{2} w^\top \Sigma_p^{-1} w\right) \\ &\propto \exp\left(-\frac{1}{2} (w - w^*)^\top \left(\frac{1}{\sigma_n^2} X X^\top + \Sigma_p^{-1}\right) (w - w^*)\right), \end{aligned}$$

kjer $w^* = \sigma_n^{-2}(\sigma_n^{-2}XX^\top + \Sigma_p^{-1})^{-1}Xy$. Tako opazimo, da je posterior porazdeljen Gaussovo normalno s povprečjem w^* in kovariančno matriko A^{-1} , kjer $A = \sigma_n^{-2}XX^\top + \Sigma_p^{-1}$. Zapišemo torej:

$$(6) \quad p(w|X, y) \sim \mathcal{N}(w^*, A^{-1})$$

Da bi naredili napovedi za testni primer, povpečimo vse možne vrednosti parametra, utežene z njihovo posteriorno verjetnostjo. Tako dobimo napovedno porazdelitev za f^* pri danem x^* .

$$(7) \quad \begin{aligned} p(f^*|x^*, X, y) &= \int p(f^*|x^*, w) \cdot p(w|X, y)dw \\ &= \mathcal{N}\left(\frac{1}{\sigma_n^2}(x^*)^\top A^{-1}Xy, (x^*)^\top A^{-1}x^*\right) \end{aligned}$$

4.2. Pogled iz prostora funkcij. Popolnoma identične rezultate kot v prejšnjem podpoglavju dobimo, če na problem pogledamo direktno iz prostora funkcij. Gaussov proces je popolnoma definiran s funkcijo povprečja in kovariančno funkcijo. Funkcijo povprečja $m(x)$ in kovariančno funkcijo $k(x, x')$ realnega procesa $f(x)$ definiramo kot:

$$(8) \quad \begin{aligned} m(x) &= \mathbb{E}[f(x)], \\ k(x, x') &= \mathbb{E}[(f(x) - m(x))(f(x') - m(x')))]. \end{aligned}$$

Gaussov proces pa zapišemo kot:

$$(9) \quad f(x) \sim GP(m(x), k(x, x')).$$

Opomba 4.2. Ponavadi zaradi poenostavitve vzamemo, da je funkcija povprečja enaka 0.

Izrek 4.3. *Marginalization property* Če GP določa $(y_1, y_2) \sim \mathcal{N}(\mu, \Sigma)$, potem enolično določa tudi $y_1 \sim \mathcal{N}(\mu_1, \Sigma_{11})$, kjer je Σ_{11} pripadajoča podmatrika matrike Σ .

Opomba 4.4. Gaussov proces je definiran kot nabor slučajnih spremenljivk. S pomočjo izreka 4.3 ugotovimo, da zmanjšanje množice ne spremeni porazdelitve, kar postane uporabno, ker nam zaradi te lastnosti ni potrebno delati na neskončnem prostoru.

V nadaljevanju bomo ugotovili, da je izbira kovariančne funkcije zelo pomembna, vendar bomo zaradi boljše predstave, v tem poglavju kot primer vzeli eksponentno kovariančno funkcijo. Kovariančna funkcija določa kovarianco med paroma slučajnih spremenljivk

$$(10) \quad \text{cov}(f(x_p), f(x_q)) = k((x_p, x_q) = \exp(-\frac{1}{2}|x_p - x_q|^2).$$

Specifikacija kovariančne funkcije določa porazdelitev funkcij. Izberemo neko število vhodnih točk X_* in zapišemo pripadajočo kovariančno matriko (uporabimo enačbo 10). Potem generiramo slučajni normalni vektor z novo kovariančno matriko

$$(11) \quad f_* \sim \mathcal{N}(0, K(X_*, X_*)),$$

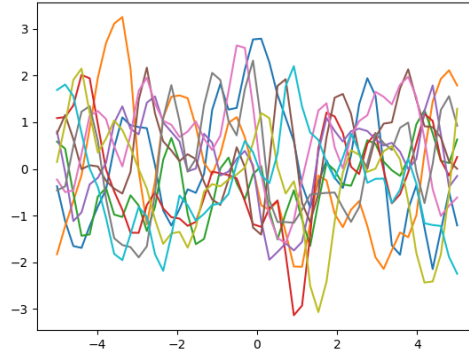
in narišemo generirane vrednosti kot funkcijo vhodnih podatkov.

Generiranje normalnih vzorcev

Za ustvarjanje vzorcev $x \sim \mathcal{N}(m, K)$, s poljubnim povprečjem m in kovarianco K uporabimo t.i. Gaussov generator (ki je na voljo v mnogih programskih okoljih) in nadaljujemo na naslednji način:

- (1) Izračunamo razcep Choleskega L pozitivno definitne simetrične kovariančne matrike $K = LL^T$, kjer je L spodnje-trikotna.
- (2) Generiramo $u \sim \mathcal{N}(0, I)$.
- (3) Izračunamo $x = m + Lu$, ki ima želeno porazdelitev s povprečjem m in kovarianco K .

SLIKA 1. Primer 10 naključnih funkcij iz GP apriori



Običajno nas ne zanimajo naključne funkcije iz apriorija, ampak želimo vključiti znanje o funkciji, ki ga zagotavljajo vhodni učni podatki. Najprej bomo obravnavali preprost poseben primer, kjer bodo opazovanja brez šuma, torej $\{(x_i, f_i) | i = 1, \dots, n\}$.

Skupna porazdelitev učnih izhodov f in testnih izhodov f_* , glede na apriori je

$$(12) \quad \begin{bmatrix} f \\ f_* \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} K(X, X) & K(X, X_*) \\ K(X_*, X) & K(X_*, X_*) \end{bmatrix}\right)$$

Če imamo n učnih točk in n_* testnih točk, je $K(X, X_*)$ $n \times n_*$ matrika kovarianc vseh parov učnih in testnih točk.

Da dobimo posteriorno porazdelitev funkcij, moramo omejiti skupno apriori porazdelitev tako, da bo vsebovala le funkcije, ki se prilegajo oz. vsebujejo opazovane vhodne točke.

Opomba 4.5. Če bi posteriori funkcije generirali kar iz apriori funkcij in potem odstranjevali tiste, ki se prilegajo vhodnim točkam, bi metoda sicer delovala, vendar bi bila računsko zelo zahtevna.

Izrek 4.6. Naj je $\begin{bmatrix} x \\ y \end{bmatrix}$ normalno porazdeljen vektor

$$(13) \quad \begin{bmatrix} x \\ y \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} A & C \\ C^\top & B \end{bmatrix}\right)$$

potem sta **robna** porazdelitev x in **pogojna** porazdelitev x glede na y enaki

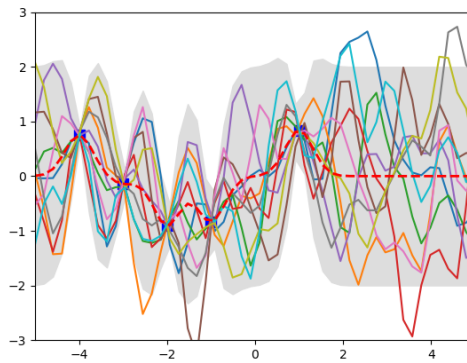
$$(14) \quad \begin{aligned} x &\sim \mathcal{N}(\mu_x, A) \\ x|y &\sim \mathcal{N}(\mu_x + CB^{-1}(y - \mu_y), A - CB^{-1}C^\top). \end{aligned}$$

Če upoštevamo izrek 4.6 dobimo porazdelitev posteriori funkcij

$$(15) \quad \begin{aligned} f_*(X_*, X, f) &\sim \mathcal{N}(K(X_*, X)K(X, X)^{-1}f, \\ &K(X_*, X_*) - K(X_*, X)K(X, X)^{-1}K(X, X_*)). \end{aligned}$$

Vrednosti funkcije f_* dobimo s pomočjo prej opisanega postopka.

SLIKA 2. Primer 10 naključnih funkcij iz GP posterior



Opomba 4.7. Slika 2 prikazuje rezultate teh izračunov na petih danih točkah označenih z modrim + simbolom. Očitno se da te izračune opraviti tudi na več dimenzionalnih podatkih, vendar je funkcije, ki jih dobimo, težje grafično prikazati.

5. KOVARIANČNE FUNKCIJE

Vloga kovariančne funkcije je pri modeliranju z Gaussovimi procesi zelo pomembna. Napovedane verjetnostne porazdelitve, ki nastopajo pri danih podatkih, so v glavnem odvisne od kovariančne funkcije in njenih hiperparametrov. Kovariančna funkcija $C(x_i, x_j)$ izraža mero podobnosti med vhodoma x_i in x_j . Za realne procese je navadno sestavljena iz dveh delov:

$$(16) \quad C(x_i, x_j) = C_f(x_i, x_j) + C_n(x_i, x_j)$$

Prvi, tj. funkcijski del $C_f(x_i, x_j)$, opisuje lastnosti neznanega sistema, ki ga želimo modelirati, drugi, šumni del $C_n(x_i, x_j)$ pa predstavlja varianco šuma. Pogosto predpostavljamo, da je šum naključen, kar pomeni, da ne pričakujemo korelacije med šumom in določenimi izhodi. Kovariančne funkcije delimo na stacionarne in nestacionarne.

5.1. Stacionarne kovariančne funkcije. Stacionarne kovariančne funkcije $C_f(x_i, x_j)$ so tiste, pri katerih je vrednost funkcije odvisna samo od relativne lege vhodnih vektorjev x_i in x_j oz. od njune medsebojne razdalje $r = |x_i - x_j|$.

Najpreprostejša oblika kovariančne funkcije je konstantna kovariančna funkcija, ki zavzame isto vrednost na celotnem območju. Opišemo jo z izrazom

$$(17) \quad C_f(r) = \frac{1}{\theta_1^2}$$

SLOVAR STROKOVNIH IZRAZOV

LITERATURA

- [1] http://valjhun.fmf.uni-lj.si/~raicm/Vaje/BPSt/Verjetnost_os.pdf