

Regresija z Gaussovimi procesi

Sara Kovačič

Mentor: doc. dr. Aljoša Peperko

- tukaj napiši vse kar boš povedala na predstavitvi

- Strojno učenje je področje umetne inteligence, ki se ukvarja z razvojem tehnik, ki računalnikom (oz. strojem) omogočajo, da se lahko učijo.
- Je metoda za kreiranje računalniških programov na podlagi podatkov (vzorcev).
- Vhodni podatki: x
- Izhodni podatki: y

Večrazsežna normalna porazdelitev

Slučajni vektor $X = [X_1, X_2, \dots, X_n]$ ima večrazsežno normalno porazdelitev s povprečjem (matematičnim upanjem) $\mu \in \mathbb{R}^n$ in kovariančno matriko $\Sigma \in \mathbb{R}^{n \times n}$, če je njena funkcija gostote enaka:

$$p(x; \mu, \Sigma) = \frac{1}{(2\pi)^{n/2} \cdot |\Sigma^{1/2}|} \exp \left(-\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right).$$

Pišemo $X \sim N(\mu, \Sigma)$.

- 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
$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$
- $$P(A) = P\left(\bigcup_i (A \cap H_i)\right) = \sum_i P(A \cap H_i) = \sum_i P(H_i) \cdot P(A|H_i)$$

Definicija

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

Enostavna linearna regresija

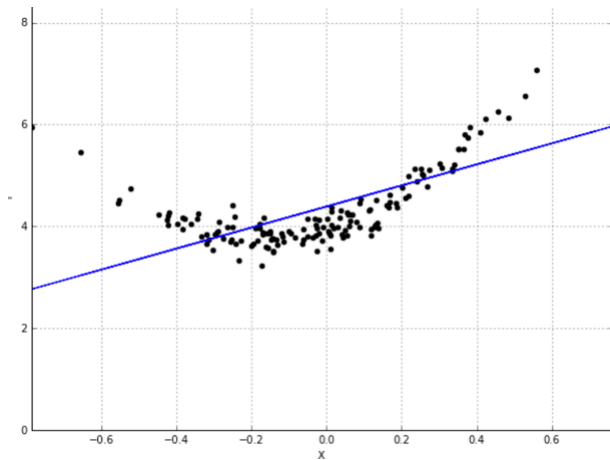
$y = f(x) + \varepsilon$, kjer je ε slučajno odstopanje od linearne zveze.

- $E(\varepsilon) = 0$.

Iščemo torej parametra β_0 in β_1 za $y = \beta_0 + \beta_1 \cdot x + \varepsilon$.

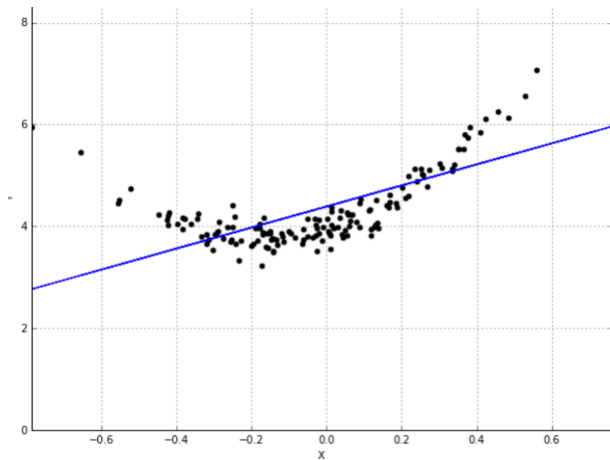
- Bayesova linearna regresija: porazdelitev nad parametri ki se posodablja ob vsakem opazovanju novih točk.
- Regresija z GP: **neparametričen** pristop, saj najde porazdelitev nad možnimi funkcijami $f(x)$, ki so skladne z opazovanimi podatki.

Problem linearne regresije



$$y = \beta_0 + \beta_1 \cdot x + \beta_2 \cdot x^2 + \varepsilon$$

Problem linearne regresije



$$y = \beta_0 + \beta_1 \cdot x + \beta_2 \cdot x^2 + \varepsilon$$

Definicija

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

Definicija

*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.*

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

Standardni linearni model

Standardni linearni model regresije z Gaussovim šumom

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

- x vhodni vektor
- w vektor uteži/parametrov
- f funkcija
- y opazovana izhodna (ciljna) vrednost

Predpostavimo:

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

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

Predpostavimo: $w \sim \mathcal{N}(0, \Sigma_p)$.

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

$$\textit{posterior} = \frac{\textit{verjetje} \cdot \textit{prior}}{\textit{robno verjetje}}, \quad p(w|y, X) = \frac{p(y|X, w) \cdot p(w)}{p(y|X)}. \quad (2)$$

Robno verjetje je neodvisno od uteži in podano kot

$$p(y|X) = \int p(y|X, w)p(w)dw. \quad (3)$$

Definicija

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

$$p(w|X, y) = \frac{p(y|X, w) \cdot p(w)}{p(y|X)}$$

$$\begin{aligned} p(w|X, y) &\propto \exp\left(-\frac{1}{2 \cdot \sigma_n^2} (y - X^T w)^T (y - X^T w)\right) \cdot \exp\left(-\frac{1}{2} w^T \Sigma_p^{-1} w\right) \\ &\propto \exp\left(-\frac{1}{2} (w - w^*)^T \left(\frac{1}{\sigma_n^2} X X^T + \Sigma_p^{-1}\right) (w - w^*)\right), \end{aligned} \quad (4)$$

$$\text{kjer } w^* = \sigma_n^{-2} (\sigma_n^{-2} X X^T + \Sigma_p^{-1})^{-1} X y.$$

$$p(w|X, y) \propto \exp(-\frac{1}{2}(w - w^*)^T(\frac{1}{\sigma_n^2}XX^T + \Sigma_p^{-1})(w - w^*))$$

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

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^* .

$$\begin{aligned} p(f^*|x^*, X, y) &= \int p(f^*|x^*, w) \cdot p(w|X, y) dw \\ &= \mathcal{N}(\frac{1}{\sigma_n^2}(x^*)^T A^{-1} X y, (x^*)^T A^{-1} x^*) \end{aligned} \tag{5}$$

Gaussov proces $f(x)$ je določen s funkcijo matematičnega upanja $m(x)$ in kovariančno funkcijo $k(x, x')$, kjer sta:

- $m(x) = E[f(x)],$
- $k(x, x') = E[(f(x) - m(x))(f(x') - m(x')))].$

Pišemo: $f(x) \sim \mathcal{GP}(m(x), k(x, x'))$.

Definicija

***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 .*

Eksponentna kovariančna funkcija

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

- Izberemo neko število vhodnih točk X_* in zapišemo pripadajočo kovariančno matriko
- Potem generiramo slučajni normalni vektor z novo kovariančno matriko

$$f_* \sim \mathcal{N}(0, K(X_*, X_*)), \quad (6)$$

- ... 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 .

Primer

```
import numpy as np
import matplotlib.pyplot as pl

# Podatki
n = 50
Xtest = np.linspace(-5, 5, n).reshape(-1,1)

# Kovariančna funkcija
def kernel(a, b, param):
    sqdist = np.sum(a**2,1).reshape(-1,1) + np.sum(b**2,1) - 2*np.dot(a, b.T)
    return np.exp(-.5 * (1/param) * sqdist)

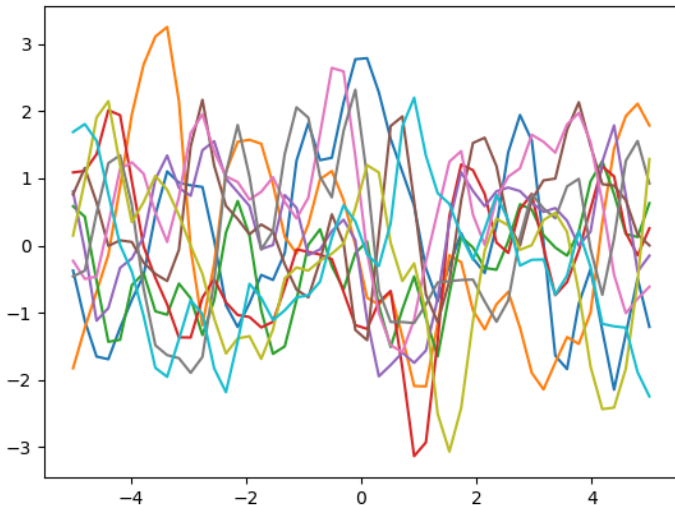
param = 0.1
K_ss = kernel(Xtest, Xtest, param)

# Cholesky
L = np.linalg.cholesky(K_ss + 1e-15*np.eye(n))

# Vzorec desetih prior funkcij
f_prior = np.dot(L, np.random.normal(size=(n,10)))

# Plot
pl.plot(Xtest, f_prior)
pl.axis([-5, 5, -3, 3])
pl.show()
```

Slika: Primer 10 naključnih funkcij iz GP apriori



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

$$\begin{bmatrix} f \\ f_* \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} K(X, X) & K(X, X_*) \\ K(X_*, X) & K(X_*, X_*) \end{bmatrix}\right) \quad (7)$$

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

Izrek

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

$$\begin{bmatrix} x \\ y \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} A & C \\ C^\top & B \end{bmatrix}\right) \quad (8)$$

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

$$\begin{aligned} x &\sim \mathcal{N}(\mu_x, A) \\ x|y &\sim \mathcal{N}(\mu_x + CB^{-1}(y - \mu_y), A - CB^{-1}C^\top). \end{aligned} \quad (9)$$

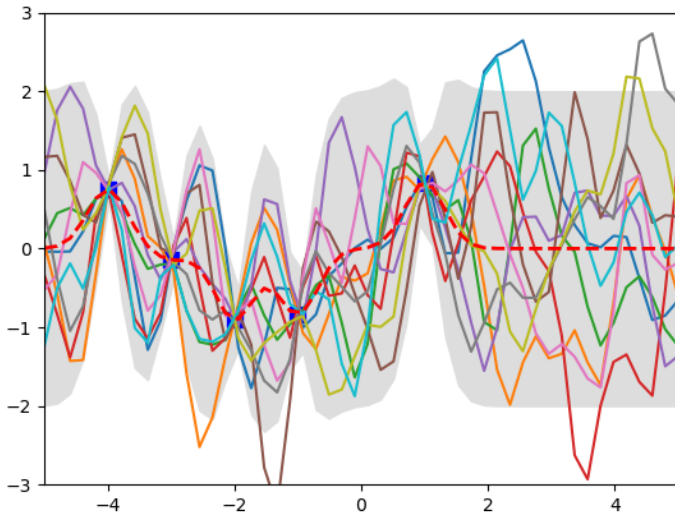
Če upoštevamo izrek dobimo porazdelitev posteriori funkcij

$$f_*|(X_*, X, f) \sim \mathcal{N}(K(X_*, X)K(X, X)^{-1}f, \\ K(X_*, X_*) - K(X_*, X)K(X, X)^{-1}K(X, X_*)).$$

(10)

Vrednosti funkcije f_* dobimo s pomočjo prej opisanega postopka.

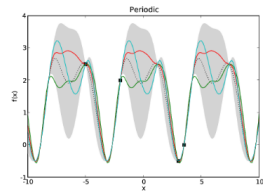
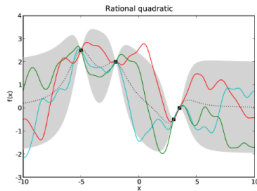
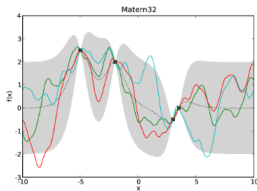
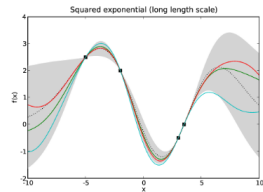
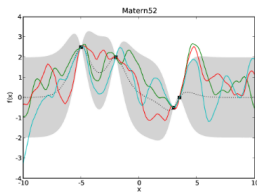
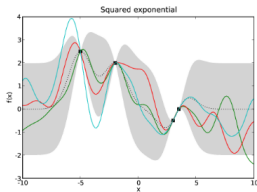
Slika: Primer 10 naključnih funkcij iz GP posterior



Vrednost kovariančne funkcije $k(x_i, x_j)$ izraža korelacijo med posameznima izhodoma $f(x_i)$ in $f(x_j)$ modela, obravnavana kot dve medsebojno povezani naključni spremenljivki.

- $k(x_i, x_j) = E[(f(x_i) - m(x_i))(f(x_j) - m(x_j))].$

Primerjava različnih kovarianc



- **Predhodno znanje** (ang. *prior*) odraža mnenje o preslikavi med vhodi in izhodi. Običajno predpostavlja gladkost.
- **Posteriorno znanje** (ang. *posterior*) dobimo, ko v model vključimo še končno število vhodno-izhodnih parov (x_i, y_i) . Dobimo posteriorno porazdelitev za predikcijo modela.
- **Vhod** v GP model: posamezne vrednosti neodvisnih spremenljivk, zbrane v vhodnem vektorju x
- **Izhod** iz GP modela: verjetnostna porazdelitev izhodne vrednosti $f(x)$ pri danem vhodnem vektorju

- Izbira kovariančne funkcije je ključnega pomena za delovanje GP modela.
- Empirični del diplomske naloge