Lecture 25 - Bayesian non-parametrics

Tuesday, December 5, 2023   2:45 PM

today: Bayesian non-parametrics
- GP
- DP

Bayesian non-parametrics

non-parametric model → infinite # of parameters (or growing with # of datapoints)

e.g. • KNN classifier → boundary complexity grows with # of datapoints

• kernel density estimation $\hat{p}(x) = \frac{1}{n} \sum_{i=1}^{n} K(x, x_i)$

Bayesian non-parametric → need prior over ∞-dim parameter

→ define dist. on ∞-dim vector (stochastic process)

stochastic process

collection of random variables indexed by a (potentially infinite) index set T

$\{X(t) : t \in T\}$

examples: $T = \{1, \dots, N\}$ $X(t) = X_t$ → random vector $(X_1, \dots, X_n)$

but also $T = \mathbb{N}$ → ∞-sequence $(X_1, X_2, X_3, \dots, \quad)$

or $T = \mathbb{R}$ → "random function"

Gaussian process → random functions

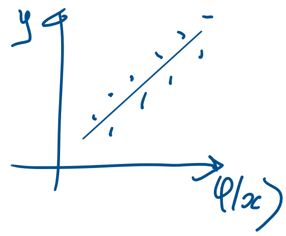
Dirichlet process → random measures/distributions

Gaussian process:

motivation from Bayesian linear regression

consider fixed location $x_1, \dots, x_n$ [conditional $Y|X$]

model of $Y|X$: $y = w^T \varphi(x) + \sigma_y \underset{\text{i.i.d. } N(0,1)}{\varepsilon}$



thus $Y|X, w \sim N(w^T\varphi(x), \sigma_y^2)$

Bayesian: prior on $w \sim N(0, \sigma_w^2 I)$ $Y_i \triangleq y(x_i)$

$\mathbb{E}[\mathbb{E}[Y(x)|W]] = \mathbb{E}[w^T\varphi(x)] = \mathbb{E}[w]^T \varphi(x) = 0$

$$\mathbb{E}[\mathbb{E}[Y(x)\mid w]] = \mathbb{E}[w^T\varphi(x)] = \underbrace{\mathbb{E}[w]}_{0}{}^T\varphi(x) = 0$$

$$\mathbb{E}[Y_i Y_j] = \mathbb{E}[(w^T\varphi(x_i)+\sigma_y\varepsilon)(w^T\varphi(x_j)+\sigma_y\varepsilon)]$$
$$= \mathbb{E}[w^T\varphi(x_i)\,w^T\varphi(x_j)] + \underbrace{\sigma_y^2\varepsilon^2]}_{} + 0$$
$$= \mathbb{E}[tr(\varphi(x_i)^T w w^T \varphi(x_j))] + \sigma_y^2$$
$$= tr(\varphi(x_i)^T \underbrace{\mathbb{E}[ww^T]}_{\sigma_w^2 I}\varphi(x_j)) + \sigma_y^2$$
$$= \sigma_w^2 \underbrace{\varphi(x_i)^T\varphi(x_j)}_{K(x_i,x_j)\to \text{similarity}} + \sigma_y^2$$

so generally, marginal on y's [function values] a priori: $\quad \Phi_N = \begin{pmatrix} -\varphi(x_1)^T- \\ \vdots \\ -\varphi(x_N)^T- \end{pmatrix}$

$$y_{1:N} \sim N(0, \underbrace{\sigma_w^2 \Phi_N \Phi_N^T}_{K \text{ (kernel matrix)}} + \sigma_y^2 I_N)$$

Gaussian process: generalization of Gaussian to ∞-dimension

parameterized by $\mu(x)$ (prior mean) $\quad \Sigma(x,x')$ (covariance (kernel))

stochastic process $Y(x)$ where for any $x_1,\dots,x_n$ (and n)

marginal: $(Y(x_1),\dots,Y(x_n))$ follows a Gaussian with (has to be psd)

mean $\begin{pmatrix}\mu(x_1)\\ \vdots \\ \mu(x_n)\end{pmatrix}$ and covariance Σ s.t. $\Sigma_{ij} = K(x_i,x_j)$

special case of GP: Bayesian linear regression

use $\boxed{\Sigma(x,x') = \sigma_w^2\varphi(x)^T\varphi(x') + \sigma_y^2}$

but more generally, square exponential kernel

$$\Sigma(x,x') = \underset{\text{uncertainty}}{c}\exp\left(-\frac{\|x-x'\|^2}{2\underset{\text{length scale}}{\sigma^2}}\right)$$

(like ∞-dim. φ)

15h46

so Bayesian inference: suppose observed $y_1,\dots,y_n$ (for $x_1,\dots,x_n$)

simple? condition in Gaussian model? $\qquad k_i \triangleq \Sigma(x,x_i)$

simple? condition in Gaussian model?

$$Y_1, Y_2, \dots, Y_n, Y(x) \sim N\left(0, \begin{pmatrix} C_n & k \\ k^T & K \end{pmatrix}\right)$$

$k_i \triangleq \Sigma(x, x_i)$

$\hookrightarrow$ how output at x covaries with output at x_i

K: $\Sigma(x, x)$

get $Y(x) \mid Y_1, \dots, Y_n \sim N\left(0 + k^T C_n^{-1} y_{1:n},\ K - k^T C_n^{-1} k\right)$

that's it

note that only need to compute once

(*) applications to Bayesian classification

- GPC use $p(y=1 \mid x) = \sigma(f(x))$
- hyperparameter selection $\rightarrow$ can maximize the marginal likelihood

(for GP regression $\rightarrow$ closed formed formula)
$\rightarrow$ model selection

demos: https://distill.pub/2019/visual-exploration-gaussian-processes/
https://www.tmpl.fi/gp/

Dirichlet process:

* used to model ∞-mixture model in Bayesian model

Bayesian mixture model (finite)

$$z \sim \text{Mult}(\pi) \quad \pi \in \Delta_K$$

$$x \mid z \sim p(x \mid \theta_z) \quad [\text{e.g. } N(x \mid \mu_z, \Sigma_z)]$$

Bayesian $\rightarrow$ put prior on π $[\text{Dirichlet}(\alpha_1, \dots, \alpha_K)]$

and θ_z [say $\theta_z \overset{iid}{\sim} G_0$]

would like $K \rightarrow \infty$ can do θ_z & π together using DP

Dirichlet process

$G \sim DP(\alpha, G_0)$

random prob. measure

G_0: dist. on Θ

α: concentration parameter

stochastic process indexed by measurable sets of Θ

for every partition of Θ in $A_1, \dots, A_n$

then $\left(G(A_1), G(A_2), \dots, G(A_n)\right) \sim \text{Dir}\left(\alpha G_0(A_1), \dots, \alpha G_0(A_n)\right)$

stick breaking construction

turns out that $G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}$ where $\pi = (\pi_1, \pi_2, \dots) \sim \text{GEM}(\alpha)$

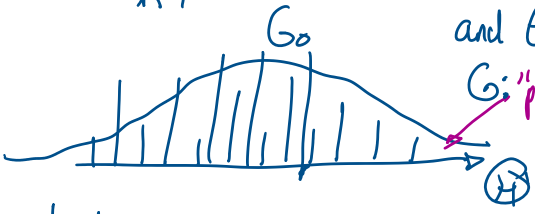
Dirac delta

G_0

and $\theta_1, \theta_2, \theta_3 \overset{iid}{\sim} G_0$

Turns out that $G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}$ where $\pi = (\pi_1, \pi_2, \dots) \sim GEM(\alpha)$

and $\theta_1, \theta_2, \theta_3 \overset{iid}{\sim} G_0$



stick-breaking construction

$\pi_1 \sim Beta(1, \alpha)$
$\quad \shortparallel$
$\quad v_1$

$\pi_2 = (1-\pi_1) v_2$

$v_i \overset{iid}{\sim} Beta(1, \alpha)$