Lecture 25 - Bayesian non-parametrics

Tuesday, December 6, 2022 3:01 PM

today: · finish factor analysis
· Bayesian non-parametrics
· GP
· DP

factor analysis continuation:

$X|Z \sim N(Wz+\mu, D)$ $\quad \varepsilon \sim N(0, D)$

$p(z)$ is Gaussian $\quad \mathbb{E}[X] = \mathbb{E}[\underbrace{\mathbb{E}[X|z]}_{Wz+\mu}] = W\underbrace{\mathbb{E}[z]}_{0} + \mu = \mu$

$$cov(X,X) = cov(\underbrace{Wz+\mu+\varepsilon}_{\text{indep.}}, Wz+\mu+\varepsilon)$$
$$= cov(Wz, Wz) + cov(\varepsilon, \varepsilon)$$
$$= W \underbrace{cov(z,z)}_{I_K} W^T + D$$
$$= WW^T + D$$

equivalent model on $\boxed{X \sim N(\mu, \underbrace{WW^T}_{\text{low rank covariance assumption}} + \underbrace{D}_{\text{diagonal} \to d \text{ degrees of freedom}})}$

WW^T: $d \times d$, but of rank $\leq K$

estimate W & D & μ by MLE

↷ do EM (latent variable model)

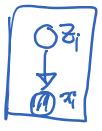
get $p(z|x)$ → Gaussian with mean $\mathbb{E}[z|x] = W^T(WW^T+D)^{-1}(x-\mu)$

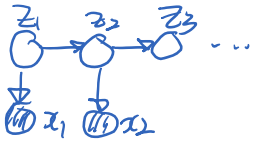
probabilistic PCA: special case of factor analysis where suppose $D = \sigma^2 I_d$

$$\lim_{\sigma \to 0} W^T(WW^T+\sigma^2 I)^{-1} = W^+ \leftarrow \text{pseudo-inverse}$$
$$= W^T \quad \text{if } W^TW = I_K$$

this suggests that PCA is limit of PPCA as $\sigma \to 0$

Kalman filter

factor analysis 

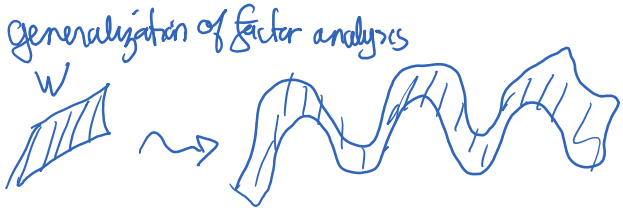


move to state space model: unroll in time (HMM style)

Kalman filter: $z_t | z_{t-1} \sim N(Az_{t-1}, B)$ ← Gaussian

→ doing "sum-product" alg. in HMM $p(z_t | x_{1:t})$
get "Kalman filtering" alg.

variational auto-encoder



$z \sim N(0, I_K)$ — diagonal noise

$x | z \sim N(\mu_w(z), \sigma_w^2(z))$

where $\mu_w(z)$ ← output of a NN ("decoder")

MLE → use EM

↳ $p(z|x)$ is intractable ⇒ approximate with variational inference

approximate $p(z|x)$ with $q_\varphi(z|x)$

$z|x \sim N(\mu_\varphi(x), \sigma_\varphi^2(x))$

output of a NN → "encoder"

in EM $\log p(x) \geq \mathbb{E}_q[\log p(z,x)] + H(q)$

$= \mathbb{E}_{q_\varphi(z|x)}[\log p_w(x|z)] - KL(q_\varphi(z|x) \| p(z))$

$N(0, I_K)$

→ allows "reparametrization trick"

$z|x \rightarrow \mu_\varphi(x) + \sigma_\varphi(x) \cdot \underset{\varepsilon}{\varepsilon}$

$\varepsilon \sim N(0,1)$

- VAE innovations:
  - share parameters phi among data points for their variational approximation q_phi(z|x)
  - re-parameterization trick to only have parameters appear in simple deterministic transformation, stochasticity is all left in N(0,1) noise variables (no parameters) => allow simple backpropagation of gradient through expectations
    - see also: https://gregorygundersen.com/blog/2018/04/29/reparameterization/
  - for more details, see: Slides on VAE by Aaron Courville - deep learning class Winter 2017

15h57

Bayesian non-parametrics

non-parametric model → infinite # of parameters (or growing with # of data pts.)

e.g. • kNN classifier → boundary complexity grows with # of pts

n — density on x given x_i

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

Bayesian non-parametric → need prior on infinite dim. parameter

→ define dist. on ∞-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 $(X_1, X_2, X_3, \dots)$

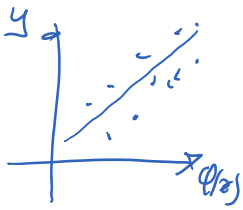
or $T = \mathbb{R}$ ⇝ "random function"

Gaussian proces → random function

Dirichlet process → random measure/distribution
useful for mixture model

Gaussian process:

motivation from Bayesian linear regression:



consider fixed location $x_1, \dots, x_N$ [conditional model of $Y|X$]

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

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

$Y_i \triangleq y(x_i)$

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

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

$$\mathbb{E}[Y_i Y_j] = \mathbb{E}[w^T \varphi(x_i) w^T \varphi(x_j) + 0 + \sigma_y^2 \varepsilon^2]$$

$$= \mathbb{E}[\mathrm{tr}(\varphi(x_i)^T \varphi(x_j) w w^T)] + \sigma_y^2 = \sigma_w^2 k(x_i, x_j) + \sigma_y^2$$

$k(x_i, x_j)$ → similarity

so generally, marginal on y's [function values] a priori:

$$\Phi_N = \begin{pmatrix} — \varphi(x_1) — \\ \vdots \\ — \varphi(x_N) — \end{pmatrix}$$

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

Gaussian prior on function outputs

Gaussian process: generalization of Gaussian to ∞-dimension

Gaussian process: generalization of Gaussian to ∞-dimension

parameterized by $\mu(x)$ $\Sigma(x,x')$

prior mean 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 mean $\begin{pmatrix}\mu(x_1)\\ \vdots \\ \mu(x_n)\end{pmatrix}$ and covariance Σ s.t. $\Sigma_{ij}=\Sigma(x_i,x_j)$ (has to be psd)

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 (uncertainty)

$$\Sigma(x,x')=C\exp\left(-\frac{\|x-x'\|^2}{2\sigma^2}\right)$$

(length scale)

(like ∞-dim φ)

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

what is posterior on $y|x$)?

simple? condition in Gaussian model?

$k_i \triangleq \Sigma(x,x_i)$ → how output a x covaries with output at x_i

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

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

note that only need to compute once

That's it?

demo: http://chifeng.scripts.mit.edu/stuff/gp-demo/

⊛ applications to Bayesian optimization

- GPC use $p(y=1|x)=\sigma(f(x))$
- hyperparameter selection → can maximize the marginal likelihood (model selection)
  - ↳ closed form expression

Dirichlet process:

⊛ used to model ∞-mixture model in Bayesian model

Bayesian mixture model (finite)

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

$x | z \sim p(x|\theta_z)$ [e.g. $N(x|\mu_z, \Sigma_z)$]

Bayesian → put a prior on π [Dirichlet $(\alpha_1, \dots \alpha_K)$]

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

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

Dirichlet process

$G \sim DP(\alpha, G_0)$ — G: random measure

G_0: dist. on Θ

α: concentration parameter

stochastic process index by measurable set of Θ

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

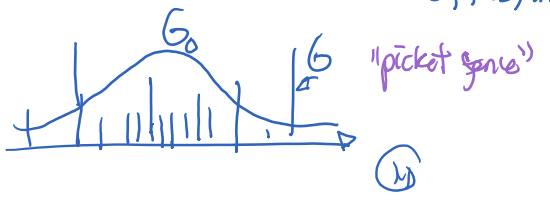
then $(G(A_1), \dots, G(A_n)) \sim \text{Dir}(\alpha G_0(A_1), \dots \alpha G(A_n))$

stick breaking construction:

turns out that $G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}$ (Dirac)

where $\pi = (\pi_1, \dots) \sim GEM(\alpha)$

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



stick-breaking construction

$\pi_1 \sim \text{Beta}(1, \gamma)$ $\quad V_i \overset{iid}{\sim} \text{Beta}(1, \gamma)$

All V_1

$\pi_2 = (1 - \pi_1) V_2$

$\pi_3 = (1 - \pi_2 - \pi_1) V_3$