Bayesian Non-Parametrics

Plan:

- Bayesian Lin. Regression
- Gaussian Process
- Dirichlet Process / DP mixtures
- de Finetti's Representation Theorem (dFRT).

Linear Regression:

$$y(x, w) = w_0 + w_1 x_1 + \dots + w_D x_D$$

$$y(x, w) = w_0 + \sum_{j=1}^{M-1} w_j \phi_j(x)$$

ϕ_j ← basis fn.

$$= \sum_{j=0}^{M-1} w_j \phi_j(x)$$

where $\phi_0(x) = 1$

$$= w^T \phi(x)$$

Gaussian basis fn:

$$\phi_j(x) = \exp\left\{ -\frac{(x - \mu_j)^2}{2s^2} \right\}$$

MLE: $t = y(a, w) + \epsilon$ ← zero mean, β Precision

$$p(t \mid x, w, \beta) = \mathcal{N}\left(t \mid y(x, w), \beta^{-1}\right)$$

$$= \prod_{n=1}^{N} \mathcal{N}\left(t_n \mid w^T \phi(x_n), \beta^{-1}\right)$$

$$w_{ML} = \left(\Phi^T \Phi\right)^{-1} \Phi^T t.$$

Bayesian Linear Regression:

β is known.

w - parameter to estimate

$$p(w) = \mathcal{N}(w \mid m_0, S_0)$$

Posterior = Prior × likelihood.

$$p(w \mid t) = \mathcal{N}(w \mid m_N, S_N)$$

ex.

$$m_N = S_N \left(S_0^{-1} m_0 + \beta \Phi^T t \right)$$

$$S_N^{-1} = S_0^{-1} + \beta \Phi^T \Phi .$$

$$w_{MAP} = m_N$$

$$S_0 = \alpha^{-1} I \quad \text{with } \alpha \to 0 \quad m_N \to w_{ML}$$

---

assume: $p(w|\alpha) = N(w|0, \alpha^{-1} I)$

$$m_N = \beta S_N \Phi^T t$$

$$S_N^{-1} = \alpha I + \beta \Phi^T \Phi .$$

$$\ln p(w|t) = -\frac{\beta}{2} \sum_{n=1}^{N} \{ t_n - w^T \phi(x_n) \}^2 - \frac{\alpha}{2} w^T w + \text{Const} .$$

$$p(\bar{t} \mid \mathbb{t}, \alpha, \beta) = \int p(t \mid w, \beta)\, p(w \mid \mathbb{t}, \alpha, \beta)\, dw$$

$$p(t \mid x, \mathbb{t}, \alpha, \beta) = \mathcal{N}\left(t \mid m_N^T \phi(x), \sigma_N^2(x)\right)$$

$$\sigma_N^2(x) = \frac{1}{\beta} + \phi(x)^T S_N \phi(x)$$

---

kernel view

$$y(x, m_N) = m_N^T \phi(x)$$

$$= \beta\, \phi(x)^T S_N \Phi^T \mathbb{t}$$

$$= \sum_{n=1}^{N} \beta \phi(x)^T S_N \phi(x_n)\, t_n$$

$$y(x, m_N) = \sum_{n=1}^{N} k(x, x_n)\, t_n$$

where $k(x, x') = \beta \phi(x)^T S_N \phi(x')$

↳ equivalent kernel / Smoother matrix.

$$\text{Cov}\left(y(x), y(x')\right) = \text{Cov}\left[\phi(x)^T w,\ w^T \phi(x')\right]$$

$$= \phi(x)^T S_N \phi(x')$$

$$= \beta^{-1} k(x, x')$$

Gaussian Processes:

functional view:

$$y(x) = w^T \phi(x)$$

$$p(w) = \mathcal{N}\left(w \mid 0, \alpha^{-1} I\right)$$

every $w \sim p(w) \rightarrow$ fn. $y(x)$

distribution over fn $y(x)$

$y(x)$

$y(x_1) \ldots y(x_n)$

$x_1 \ldots \ldots x_n \longrightarrow (y(x_1) \ldots y(x_n))$

$$y = \phi w$$

$$E[y] = \phi E[w] = 0$$

$$\text{Cov}[y] = E[yy^T]$$

$$= \phi E[ww^T]\phi^T$$

$$= \frac{1}{\alpha}\phi\phi^T = K$$ → Gram matrix.

$$K_{nm} = k(x_n, x_m) = \frac{1}{\alpha}\phi(x_n)^T\phi(x_m)$$

$$k(x, x') = \exp(-\theta|x - x'|)$$

How to use this GP for regression?

$$t_n = y_n + \epsilon_n$$

$$p(t_n | y_n) = \mathcal{N}(t_n | y_n, \beta^{-1})$$

$$p(\mathbf{t} | \mathbf{y}) = \mathcal{N}(\mathbf{t} | \mathbf{y}, \beta^{-1} \mathbf{I}_N)$$

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y} | \mathbf{0}, \mathbf{K})$$

$$p(\mathbf{t}) = \int p(\mathbf{t} | \mathbf{y}) p(\mathbf{y}) \, d\mathbf{y}$$

$$= \mathcal{N}(\mathbf{t} | \mathbf{0}, \mathbf{C}) \quad \text{ex.}$$

$$C(x_n, x_m) = k(x_n, x_m) + \beta^{-1} \delta_{nm} \quad (*)$$

$$p(t_{N+1}) = \mathcal{N}(t_{N+1} | 0, C_{N+1})$$

$$C_{N+1} = \begin{pmatrix} C_N & k \\ k^T & c \end{pmatrix} \quad \text{ex.}$$

where $k \rightarrow k(x_n, x_{n+1})$

$c \rightarrow k(x_{n+1}, x_{n+1})$

$$P(t_{N+1} \mid t) = \mathcal{N}(m, \sigma^2)$$

$$m = k^T C_N^{-1} t$$

$$\sigma^2 = c - k^T C_N^{-1} k.$$

One restriction:

Covariance matrix, (*) should be positive definite.

C.P regression	basis fn linear reg.
invert $N \times N$ matrix.	invert $M \times M$ matrix.
$O(N^3)$	$O(M^3)$
one step	$M \ll N$
every step $O(N^2)$	$O(M^2)$

Main advantage of GP:

you can use inf. dimensional basis fn. → kernel.

Parametric Model

finite set of Param. θ

Given θ, Prediction is indep. of D.

$$P(x|\theta, D) = P(x|\theta)$$

- even if data is unbounded, Complexity of model is bounded.
- Not flexible.

Non-parametric Model.

- infinite dimensional θ.

$\theta \Rightarrow$ function.

→ amt of info. θ can capture grows as D grows.

Example:

1) fn approximation

Poly. regression Vs. G.P.

2) Classification

log. regression Vs. GP classif.

3) Clustering.

mixture model, kMeans Vs.

DP mixtures.

4) time series.

HMM vs. infinite HMM.

de Finetti's Representation Theorem :

dFRT

Coin tossing :

I. I. D.

Independence :

$$p(x_{n+1} \mid x_1 \ldots x_n) = p(x_{n+1})$$

only with I.D, we can learn.

Exchangeability :

{0, 1, 0, 0, 1, 0}

order does not matter.

$P(F, F, F, G, G, G) >$

$P(G, G, G, F, F, F)$

order matters.

exchangeable $\Rightarrow$ weaker than indep

Defn: A set of R.V. $\{X_n\}$ is said to be exchangeable if given the joint density $p(x_1 \ldots x_n)$ we have

$$p(x_1 \ldots x_n) = p(x_{\sigma(1)} \ldots x_{\sigma(n)})$$

σ - perm. of $(1 \ldots n)$

IID $\Rightarrow$ exchangeable.

exchangeable $\not\Rightarrow$ IID.

exchang $\rightarrow$ ID.

---

dFRT for Bernoulli R.V

Let $\{x_i\}_{i=1}^{\infty}$ be a seq. of finitely exchag. random variable.

i.e. $\forall n > 0$, each finite subsequence of $\{x_i\}_{i=1}^{n}$ is exchanged.

Then $\exists$ a random variable θ and a distrib. fn $F(\theta)$ such that

$$P\left(\lim_{n\to\infty} \frac{\sum x_i}{n} = \theta\right) = 1$$

with $\theta \sim F(\theta)$

and

$$p(x_1 \dots x_n) = \int_0^1 \prod_i \theta^{x_i} (1-\theta)^{(1-x_i)} \, dF(\theta)$$