Bayesian non-parametrics

- non-parametric model: infinite # of parameters
  (or growing with # of data points)
  - e.g., KNN classifier: boundary complexity grows with # of data points

Bayesian non-parametric: need prior over infinite parameter
- define prior as a mean vector (stochastic process)

Stochastic process:

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

\[ \{ X(t); t \in T \} \]

Examples:
- \( T = \mathbb{Z} \) \( \ldots \, \mathbb{N} \) \( \ldots \)
  \( X(t) = x_t \rightarrow \) random vector \( (x_1, \ldots, x_T) \)
- but also \( T = \mathbb{N} \) \( \rightarrow \) sequence \( (x_1, x_2, x_3, \ldots, ) \)
  - or \( T = \mathbb{R} \) \( \rightarrow \) "random function"

Gaussian process: random functions
- Dirichlet process: random measures/distributions

Gaussian process:

- notation from Bayesian linear regression

Consider fixed location \( x_1, \ldots, x_n \) [conditional \( y|x \)]

model of \( Y|x \): \( y = \mathbf{w}^T(x) + \sigma_y \varepsilon \)

thus \( Y|x, \mathbf{w} \sim \mathcal{N}(\mathbf{w}^T(x), \sigma_y^2) \)

Bayesian: prior on \( \mathbf{w} \sim \mathcal{N}(0, \sigma_w^2 I) \)

\[ y \sim \mathcal{N}(\mathbf{w}^T(x), \sigma_y^2) \]

\[ E[ \mathbf{w}^T(x) | y ] = \mathbf{y} | x \]

\[ E[ y^T(x) ] = E[ \mathbf{w}^T(x) ] = 0 \]
\[ E[\mathbf{Y}(x) | \mathbf{Y}] = E[\mathbf{w}(\mathbf{Q})] = E[\mathbf{w}^T \Phi(x)] = \mathbf{0} \]
\[ E[\mathbf{Y} | \mathbf{Y}] = E\left[ (\mathbf{w}^T \Phi(x) + \mathbf{y}) \mathbf{w}^T \Phi(x) + \mathbf{y} \mathbf{0} \right] \]
\[ = E\left[ \mathbf{w}^T \Phi(x) \right] \mathbf{w}^T \Phi(x) + \sigma_y^2 \mathbf{I} \]
\[ = \mathbf{w}^T \Phi(x) \mathbf{w} \Phi(x) + \sigma_y^2 \mathbf{I} \]
\[ = \mathbf{w}^T \Phi(x) \mathbf{w} \Phi(x) + \sigma_y^2 \mathbf{I} \]
\[ = \sigma_y^2 \mathbf{I} \]
\[ E[\mathbf{Y} | \mathbf{Y}] \xrightarrow{\text{similarity}} \]

So generally, marginal on \( y_i \)’s: function values \( a \) priori:
\[ Y_{n+1} \sim \mathcal{N}(0, \sigma_y^2 \Phi \Phi^T + \sigma_y^2 \mathbf{I}) \]

\[ K(\mathbf{x}_i, \mathbf{x}_j) \]

**Gaussian process:** Generalization of Gaussian to \( \mathbb{R}^d \)-domain

- Parameterized by \( \mu(x) \) and \( \Sigma(x, x') \)
- Mean \( \mu(x) \)
- Covariance \( \Sigma(x, x') \)

Stochastic process \( Y(\mathbf{x}) \) where for any \( \mathbf{x}, \ldots, \mathbf{x}_n \) (and \( n \))

Marginal: \( \{Y(x_1), \ldots, Y(x_n)\} \) follows a Gaussian with

- mean \( \mu(x) \)
- covariance \( \Sigma(x, x') \)

Special case of GP: Bayesian linear regression

Use \[ \Sigma(x, x') = \sigma_y^2 \Phi(x) \Phi(x') + \sigma_y^2 \]

But more generally, square exponential kernel

\[ \Sigma(x, x') = \sigma_y^2 \exp\left(-\frac{\|x - x\|^2}{2\theta^2}\right) \]

Length scale

\[ \theta \quad \text{and dim.(\theta)} \]

So Bayesian inference: Suppose observed \( y_1, \ldots, y_n \) for \( x_1, \ldots, x_n \)

Simple? condition in Gaussian model:

\[ K + \Sigma(x, x') \]

15h46
Demos: [https://distill.pub/2019/visual-exploration-gaussian-processes/](https://distill.pub/2019/visual-exploration-gaussian-processes/)
[https://www.tmpl.fi/gp/](https://www.tmpl.fi/gp/)

Simple condition in Gaussian model:

\[
Y_1, Y_2, \ldots, Y_n, Y(x) \sim N(0, (C_t^{-1})_t \mathbf{0})
\]

Let us adopt at \(x\) covariance with adapt at \(x\).

get \(Y(x) | Y_1, \ldots, Y_n \sim N(0 + k^t C_t^{-1} y(x), k - k^t C_t^{-1} k)\)

That's it, we only need to compute once.

Applications to Bayesian classification:

- GPC: use \(p(y=1|x) = \sigma(f(x))\)
- Hyperparameter selection → can maximize the marginal likelihood
  
  \[\text{for GP regression} \rightarrow \text{closed form expression}\]
  
  \[\text{for model selection}\]

\textbf{Dirichlet process:}

* used to model \(k\)-mixture model in Bayesian model.

Bayesian mixture model (finite):

\[Z \sim \text{Mult}(\pi), \pi \in \Delta K\]

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

Bayesian → put prior on \(x\) \[\text{Dirichlet } (\alpha_1, \ldots, \alpha_K)\]

and \(Z\) \[\text{say } Z_2 \text{ iid Go}\]

Would like \(k \rightarrow \infty\) can do \(Z_2 \mid \alpha\) together using DP.

\textbf{Dirichlet process}

\[\alpha \sim \text{DP}(\omega, G_0)\]

\(\omega, G_0\) = dist on \(\alpha\)

\(\alpha\) = concentration parameter


For every partition of \(G_0\) in \(A_1, \ldots, A_n\)

then \(G(A_1), G(A_2), \ldots, G(A_n) \sim \text{DP}(\alpha(G(A_1)), \ldots, \alpha(G(A_n)))\)

\textbf{Stick breaking construction}

\[\text{lets set that } G = \sum_{k=1}^{\tilde{N}} \pi_k \delta_{C_k} \quad \text{where } \pi = (\pi_1, \pi_2, \ldots) \sim \text{GEM}(\alpha)\]

\[B, \omega, \alpha, G\]
We have that
\[ u = \sum_{k=1}^{n} \pi_k \Delta x_k \]
where \( \pi = (\pi_1, \pi_2, \ldots) \) \( \in \text{GEM}(\alpha) \)
and \( \theta, \alpha_x, \alpha_y \) \( \in \mathbb{G_0} \)

tie failed face

stick-breaking construction

\[ \pi \sim \text{Beta}(1, \alpha) \]

\[ \pi_2 = (1 - \pi_1) \pi_2 \]

\[ V \sim \text{Beta}(1, \alpha) \]