Supervised Learning : : CHEAT SHEET

Bayesian Linear Model

Bayesian Linear Model:

$$y^{(i)} = f\left(\mathbf{x}^{(i)}\right) + \epsilon^{(i)} = \boldsymbol{\theta}^T\mathbf{x}^{(i)} + \epsilon^{(i)}, \quad \text{for } i \in \{1, \dots, n\}$$

where $\epsilon^{(i)} \sim \mathcal{N}(0, \sigma^2)$.
Parameter vector $\boldsymbol{\theta}$ is stochastic and follows a distribution.

Gaussian variant:

- Prior distribution: $\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}, \tau^2 \boldsymbol{I}_p)$
- Posterior distribution: $\boldsymbol{\theta} \mid \mathbf{X}, \mathbf{y} \sim \mathcal{N}(\sigma^{-2}\boldsymbol{K}^{-1}\mathbf{X}^\top\mathbf{y}, \boldsymbol{K}^{-1})$ with $\boldsymbol{K} := \sigma^{-2}\mathbf{X}^\top\mathbf{X} + \frac{1}{\tau^2}\boldsymbol{I}_p$
- Predictive distribution of $y_* = \boldsymbol{\theta}^\top\mathbf{x}_*$ for a new observations $\mathbf{x}_*$:

$$y_* \mid \mathbf{X}, \mathbf{y}, \mathbf{x}_* \sim \mathcal{N}(\sigma^{-2}\mathbf{y}^\top\mathbf{X}\mathbf{A}^{-1}\mathbf{x}_*, \mathbf{x}_*^\top\mathbf{A}^{-1}\mathbf{x}_*)$$

Gaussian Processes

Weight-Space View	Function-Space View
Parameterize functions	Work on functions directly
Example: $f(\mathbf{x} \mid \boldsymbol{\theta}) = \boldsymbol{\theta}^\top\mathbf{x}$	
Define distributions on $\boldsymbol{\theta}$	Define distributions on f
Inference in parameter space Θ	Inference in function space $\mathcal{H}$

Gaussian Processes: A function $f(\mathbf{x})$ is generated by a GP $\mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$ if for **any finite** set of inputs $\left\{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}\right\}$, the associated vector of function values $\boldsymbol{f} = \left(f(\mathbf{x}^{(1)}), \dots, f(\mathbf{x}^{(n)})\right)$ has a Gaussian distribution

$$\boldsymbol{f} = \left[f\left(\mathbf{x}^{(1)}\right), \dots, f\left(\mathbf{x}^{(n)}\right)\right] \sim \mathcal{N}(\boldsymbol{m}, \boldsymbol{K}),$$

with

$$\mathbf{m} := \left(m\left(\mathbf{x}^{(i)}\right)\right)_i, \quad \mathbf{K} := \left(k\left(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}\right)\right)_{i,j},$$

where $m(\mathbf{x})$ is the mean function and $k(\mathbf{x}, \mathbf{x}')$ is the covariance function.

Types of **covariance functions**:

- $k(.,.)$ is stationary if it is as a function of $\boldsymbol{d} = \boldsymbol{x} - \boldsymbol{x}'$, $\rightsquigarrow k(\boldsymbol{d})$
- $k(.,.)$ is isotropic if it is a function of $r = \|\boldsymbol{x} - \boldsymbol{x}'\|$, $\rightsquigarrow k(r)$
- $k(.,.)$ is a dot product covariance function if k is a function of $\boldsymbol{x}^T\boldsymbol{x}'$

Commonly used covariance functions:

Name	$k(\boldsymbol{x}, \boldsymbol{x}')$
constant	σ_0^2
linear	$\sigma_0^2 + \boldsymbol{x}^T\boldsymbol{x}'$
polynomial	$(\sigma_0^2 + \boldsymbol{x}^T\boldsymbol{x}')^p$
squared exponential	$\exp(-\frac{\|\boldsymbol{x}-\boldsymbol{x}'\|^2}{2\ell^2})$
Matérn	$\frac{1}{2^\nu\Gamma(\nu)}\left(\frac{\sqrt{2\nu}}{\ell}\|\boldsymbol{x}-\boldsymbol{x}'\|\right)^\nu K_\nu\left(\frac{\sqrt{2\nu}}{\ell}\|\boldsymbol{x}-\boldsymbol{x}'\|\right)$
exponential	$\exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{x}'\|}{\ell}\right)$

Gaussian Processes Prediction

Posterior Process

Assuming a zero-mean GP prior $\mathcal{GP}(\mathbf{0}, k(\mathbf{x}, \mathbf{x}'))$. For $f_* = f(\mathbf{x}_*)$ on single unobserved test point $\mathbf{x}_*$

$$f_* \mid \mathbf{x}_*, \mathbf{X}, \boldsymbol{f} \sim \mathcal{N}(\boldsymbol{k}_*^T\mathbf{K}^{-1}\boldsymbol{f}, \boldsymbol{k}_{**} - \boldsymbol{k}_*^T\mathbf{K}^{-1}\boldsymbol{k}_*),$$

where, $\mathbf{K} = \left(k\left(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}\right)\right)_{i,j}$, $\boldsymbol{k}_* = \left[k\left(\mathbf{x}_*, \mathbf{x}^{(1)}\right), ..., k\left(\mathbf{x}_*, \mathbf{x}^{(n)}\right)\right]$ and $\boldsymbol{k}_{**} = k(\mathbf{x}_*, \mathbf{x}_*)$.

For multiple unobserved test points $\boldsymbol{f}_* = \left[f\left(\mathbf{x}_*^{(1)}\right), ..., f\left(\mathbf{x}_*^{(m)}\right)\right]$:

$$\boldsymbol{f}_* \mid \mathbf{X}_*, \mathbf{X}, \boldsymbol{f} \sim \mathcal{N}(\mathbf{K}_*^T\mathbf{K}^{-1}\boldsymbol{f}, \mathbf{K}_{**} - \mathbf{K}_*^T\mathbf{K}^{-1}\mathbf{K}_*).$$

with $\mathbf{K}_* = \left(k\left(\mathbf{x}^{(i)}, \mathbf{x}_*^{(j)}\right)\right)_{i,j}$, $\mathbf{K}_{**} = \left(k\left(\mathbf{x}_*^{(i)}, \mathbf{x}_*^{(j)}\right)\right)_{i,j}$.

Predictive mean when assuming a non-zero mean GP prior $\mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$ with mean $m(\mathbf{x})$:

$$m(\mathbf{X}_*) + \mathbf{K}_*\mathbf{K}^{-1}(\boldsymbol{y} - m(\mathbf{X}))$$

Predictive variance remains unchanged.

Noisy Posterior Process

Assuming a zero-mean GP prior $\mathcal{GP}(\mathbf{0}, k(\mathbf{x}, \mathbf{x}'))$:

$$\boldsymbol{f}_* \mid \mathbf{X}_*, \mathbf{X}, \boldsymbol{y} \sim \mathcal{N}(\boldsymbol{m}_{\text{post}}, \boldsymbol{K}_{\text{post}}).$$

with nugget σ^2 and

$$\boldsymbol{m}_{\text{post}} = \mathbf{K}_*^T\left(\mathbf{K} + \sigma^2 \cdot \boldsymbol{I}\right)^{-1}\boldsymbol{y}$$
$$\boldsymbol{K}_{\text{post}} = \mathbf{K}_{**} - \mathbf{K}_*^T\left(\mathbf{K} + \sigma^2 \cdot \boldsymbol{I}\right)^{-1}\mathbf{K}_*,$$

Predictive mean when assuming a non-zero mean GP prior $\mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$ with mean $m(\mathbf{x})$:

$$m(\mathbf{X}_*) + \mathbf{K}_*(\mathbf{K} + \sigma^2\boldsymbol{I})^{-1}(\boldsymbol{y} - m(\mathbf{X}))$$

Predictive variance remains unchanged.

Train a Gaussian Processes

We can learn the numerical hyperparameters of a selected covariance function directly during GP training.
Let us assume

$$y = f(\mathbf{x}) + \epsilon, \ \epsilon \sim \mathcal{N}\left(0, \sigma^2\right),$$

where $f(\mathbf{x}) \sim \mathcal{GP}(\mathbf{0}, k(\mathbf{x}, \mathbf{x}'|\boldsymbol{\theta}))$.
Observing $\boldsymbol{y} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{K} + \sigma^2\boldsymbol{I}\right)$, the marginal log-likelihood (or evidence) is

$$\log p(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\theta}) = \log\left[(2\pi)^{-n/2}|\boldsymbol{K}_y|^{-1/2}\exp\left(-\frac{1}{2}\boldsymbol{y}^\top\boldsymbol{K}_y^{-1}\boldsymbol{y}\right)\right]$$
$$= -\frac{1}{2}\boldsymbol{y}^T\boldsymbol{K}_y^{-1}\boldsymbol{y} - \frac{1}{2}\log|\boldsymbol{K}_y| - \frac{n}{2}\log 2\pi.$$

with $\boldsymbol{K}_y := \boldsymbol{K} + \sigma^2\boldsymbol{I}$ and $\boldsymbol{\theta}$ denoting the hyperparameters (the parameters of the covariance function).

The three terms of the marginal likelihood have interpretable roles, considering that the model becomes less flexible as the length-scale increases:

- the data fit $-\frac{1}{2}\boldsymbol{y}^T\boldsymbol{K}_y^{-1}\boldsymbol{y}$, which tends to decrease if the length scale increases
- the complexity penalty $-\frac{1}{2}\log|\boldsymbol{K}_y|$, which depends on the covariance function only and which increases with the length-scale, because the model gets less complex with growing length-scale
- a normalization constant $-\frac{n}{2}\log 2\pi$