Semi-supervised Learning with Graph Gaussian Processes

2019.7.4

Gaussian processes for regression

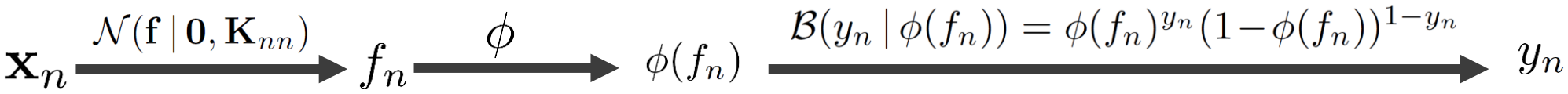
$$\mathbf{X} = \{\mathbf{x}_n\}_{n=1}^N \quad \mathbf{y} = \{y_n\}_{n=1}^N \qquad p(\mathbf{f}|\mathbf{X}) = \mathcal{N}(\mathbf{f}|\mathbf{0}, \mathbf{K}_N) \qquad [\mathbf{K}_N]_{nn'} = K(\mathbf{x}_n, \mathbf{x}_{n'})$$

$$y_i = f_i + \epsilon_i \qquad \epsilon_i \sim N(0, \sigma^2) \quad f_i = f(\mathbf{x}_i)$$

Integrating out the latent function values we obtain the marginal likelihood

$$p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{y}|\mathbf{0}, \mathbf{K}_N + \sigma^2\mathbf{I})$$

$$p(y|\mathbf{x}, \mathcal{D}, \boldsymbol{\theta}) = \mathcal{N}\left(y\middle|\mathbf{k}_\mathbf{x}^\top(\mathbf{K}_N + \sigma^2\mathbf{I})^{-1}\mathbf{y},\ K_{\mathbf{xx}} - \mathbf{k}_\mathbf{x}^\top(\mathbf{K}_N + \sigma^2\mathbf{I})^{-1}\mathbf{k}_\mathbf{x} + \sigma^2\right) \qquad [\mathbf{k}_\mathbf{x}]_n = K(\mathbf{x}_n, \mathbf{x})$$



We denote the probit inverse link function as $\phi(x) = \int_{-\infty}^{x} \mathcal{N}(a \mid 0, 1)\mathrm{d}a$

the Bernoulli distribution $\mathcal{B}(y_n \mid \phi(f_n)) = \phi(f_n)^{y_n}(1-\phi(f_n))^{1-y_n}$

The joint distribution of data and latent variables becomes $p(\mathbf{y}, \mathbf{f}) = \prod_{n=1}^{N} \mathcal{B}(y_n \mid \phi(f_n))\, \mathcal{N}(\mathbf{f} \mid \mathbf{0}, \mathbf{K}_{nn})$

Given a data set of size N with D-dimensional features $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]^\mathsf{T}$, a symmetric binary adjacency matrix $\mathbf{A} \in \{0, 1\}^{N \times N}$ that represents the relational graph of the data points and labels for a subset of the data points, $\mathcal{Y}_o = [y_1, \ldots, y_O]$, with each $y_i \in \{1, \ldots, K\}$, we seek to predict the unobserved labels of the remaining data points $\mathcal{Y}_U = [y_{O+1}, \ldots y_N]$. We denote the set of all labels as $\mathbf{Y} = \mathcal{Y}_O \cup \mathcal{Y}_U$.

Sparse Pseudo-input Gaussian processes (SPGPs)

We consider a model with likelihood given by the GP predictive distribution, and parameterized by a pseudo data set.

pseudo-inputs $\bar{\mathbf{X}} = \{\bar{\mathbf{x}}_m\}_{m=1}^M$ pseudo targets $\bar{\mathbf{f}} = \{\bar{f}_m\}_{m=1}^M$

$$p(y|\mathbf{x}, \bar{\mathbf{X}}, \bar{\mathbf{f}}) = \mathcal{N}\left(y\middle|\mathbf{k}_\mathbf{x}^\top \mathbf{K}_M^{-1}\bar{\mathbf{f}},\ K_{\mathbf{xx}} - \mathbf{k}_\mathbf{x}^\top \mathbf{K}_M^{-1}\mathbf{k}_\mathbf{x} + \sigma^2\right)$$

$$[\mathbf{K}_M]_{mm'} = K(\bar{\mathbf{x}}_m, \bar{\mathbf{x}}_{m'}) \qquad [\mathbf{k}_\mathbf{x}]_m = K(\bar{\mathbf{x}}_m, \mathbf{x})$$

The target data are generated i.i.d. given the inputs, giving the complete data likelihood

$$p(\mathbf{y}|\mathbf{X}, \bar{\mathbf{X}}, \bar{\mathbf{f}}) = \prod_{n=1}^N p(y_n|\mathbf{x}_n, \bar{\mathbf{X}}, \bar{\mathbf{f}}) = \mathcal{N}(\mathbf{y}|\mathbf{K}_{NM}\mathbf{K}_M^{-1}\bar{\mathbf{f}},\ \mathbf{\Lambda} + \sigma^2\mathbf{I})$$

$$\mathbf{\Lambda} = \mathrm{diag}(\boldsymbol{\lambda}) \qquad \lambda_n = K_{nn} - \mathbf{k}_n^\top \mathbf{K}_M^{-1}\mathbf{k}_n \qquad [\mathbf{K}_{NM}]_{nm} = K(\mathbf{x}_n, \bar{\mathbf{x}}_m)$$

$$p(\mathbf{y}|\mathbf{X}, \bar{\mathbf{X}}, \bar{\mathbf{f}}) = \prod_{n=1}^{N} p(y_n|\mathbf{x}_n, \bar{\mathbf{X}}, \bar{\mathbf{f}}) = \mathcal{N}(\mathbf{y}|\mathbf{K}_{NM}\mathbf{K}_M^{-1}\bar{\mathbf{f}},\ \mathbf{\Lambda} + \sigma^2\mathbf{I})$$

Learning in the model involves finding a suitable setting of the parameters – an appropriate **pseudo data set that explains the real data well**.
However rather than simply **maximize the likelihood** with respect to **pseudo data** it turns out that we can **integrate out the pseudo targets.**

We place a Gaussian prior on the pseudo targets:

$$p(\bar{\mathbf{f}}|\bar{\mathbf{X}}) = \mathcal{N}(\bar{\mathbf{f}}|\mathbf{0}, \mathbf{K}_M)$$

This is a very reasonable prior because we expect the pseudo data to be distributed in a very similar manner to the real data, if they are to model them well.

The posterior distribution over pseudo targets

$$p(\bar{\mathbf{f}}|\mathcal{D}, \bar{\mathbf{X}}) = \mathcal{N}\left(\bar{\mathbf{f}}|\mathbf{K}_M\mathbf{Q}_M^{-1}\mathbf{K}_{MN}(\mathbf{\Lambda} + \sigma^2\mathbf{I})^{-1}\mathbf{y},\ \mathbf{K}_M\mathbf{Q}_M^{-1}\mathbf{K}_M\right) \qquad \mathbf{Q}_M = \mathbf{K}_M + \mathbf{K}_{MN}(\mathbf{\Lambda} + \sigma^2\mathbf{I})^{-1}\mathbf{K}_{NM}$$

$$p(y|\mathbf{x}, \bar{\mathbf{X}}, \bar{\mathbf{f}}) = \mathcal{N}\left(y \middle| \mathbf{k}_{\mathbf{x}}^\top \mathbf{K}_M^{-1} \bar{\mathbf{f}},\ K_{\mathbf{xx}} - \mathbf{k}_{\mathbf{x}}^\top \mathbf{K}_M^{-1} \mathbf{k}_{\mathbf{x}} + \sigma^2\right)$$

$$p(\bar{\mathbf{f}}|\mathcal{D}, \bar{\mathbf{X}}) = \mathcal{N}\left(\bar{\mathbf{f}} | \mathbf{K}_M \mathbf{Q}_M^{-1} \mathbf{K}_{MN} (\mathbf{\Lambda} + \sigma^2 \mathbf{I})^{-1} \mathbf{y},\ \mathbf{K}_M \mathbf{Q}_M^{-1} \mathbf{K}_M\right)$$

$$p(y_*|\mathbf{x}_*, \mathcal{D}, \bar{\mathbf{X}}) = \int \mathrm{d}\bar{\mathbf{f}}\ p(y_*|\mathbf{x}_*, \bar{\mathbf{X}}, \bar{\mathbf{f}})\, p(\bar{\mathbf{f}}|\mathcal{D}, \bar{\mathbf{X}}) = \mathcal{N}(y_*|\mu_*, \sigma_*^2)$$

$$\mu_* = \mathbf{k}_*^\top \mathbf{Q}_M^{-1} \mathbf{K}_{MN} (\mathbf{\Lambda} + \sigma^2 \mathbf{I})^{-1} \mathbf{y} \qquad \sigma_*^2 = K_{**} - \mathbf{k}_*^\top (\mathbf{K}_M^{-1} - \mathbf{Q}_M^{-1}) \mathbf{k}_* + \sigma^2$$

$$\mathbf{Q}_M = \mathbf{K}_M + \mathbf{K}_{MN} (\mathbf{\Lambda} + \sigma^2 \mathbf{I})^{-1} \mathbf{K}_{NM}$$

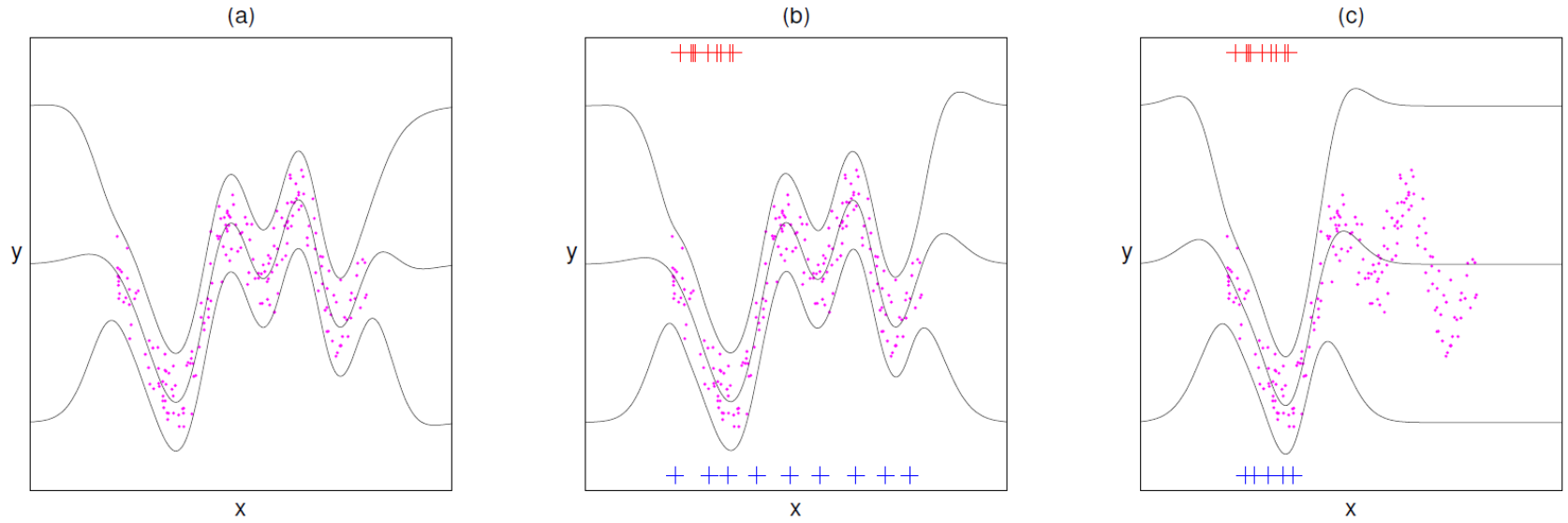
We are left with the problem of finding the pseudo-input locations $\bar{\mathbf{X}}$ and hyperparameters $\boldsymbol{\Theta} = \{\boldsymbol{\theta}, \sigma^2\}$

We can do this by computing the marginal likelihood

$$p(\mathbf{y}|\mathbf{X}, \bar{\mathbf{X}}, \boldsymbol{\Theta}) = \int \mathrm{d}\bar{\mathbf{f}}\; p(\mathbf{y}|\mathbf{X}, \bar{\mathbf{X}}, \bar{\mathbf{f}})\, p(\bar{\mathbf{f}}|\bar{\mathbf{X}}) \quad = \mathcal{N}(\mathbf{y}|\mathbf{0},\; \mathbf{K}_{NM}\mathbf{K}_M^{-1}\mathbf{K}_{MN} + \boldsymbol{\Lambda} + \sigma^2\mathbf{I})$$

The marginal likelihood can then be maximized with respect to all these parameters $\{\bar{\mathbf{X}}, \boldsymbol{\Theta}\}$ by gradient ascent.

Seeger's method of PLV $\quad p(\mathbf{y}|\mathbf{X}, \bar{\mathbf{X}}, \boldsymbol{\Theta}) = \mathcal{N}(\mathbf{y}|\mathbf{0},\; \mathbf{K}_{NM}\mathbf{K}_M^{-1}\mathbf{K}_{MN} + \sigma^2\mathbf{I})$



两种变分方法的下界及其优化求解

Variational Learning of Inducing Variables in Sparse Gaussian Processes AISTATS2009

Scalable Variational Gaussian Process Classification AISTAS2015

We wish to define a sparse method that **directly approximates the posterior GP mean and covariance functions**

$$m_{\mathbf{y}}(\mathbf{x}) = K_{\mathbf{x}n}(\sigma^2 I + K_{nn})^{-1}\mathbf{y}$$

$$k_{\mathbf{y}}(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}, \mathbf{x}') - K_{\mathbf{x}n}(\sigma^2 I + K_{nn})^{-1}K_{n\mathbf{x}'}$$

This posterior GP can be also described by the predictive Gaussian

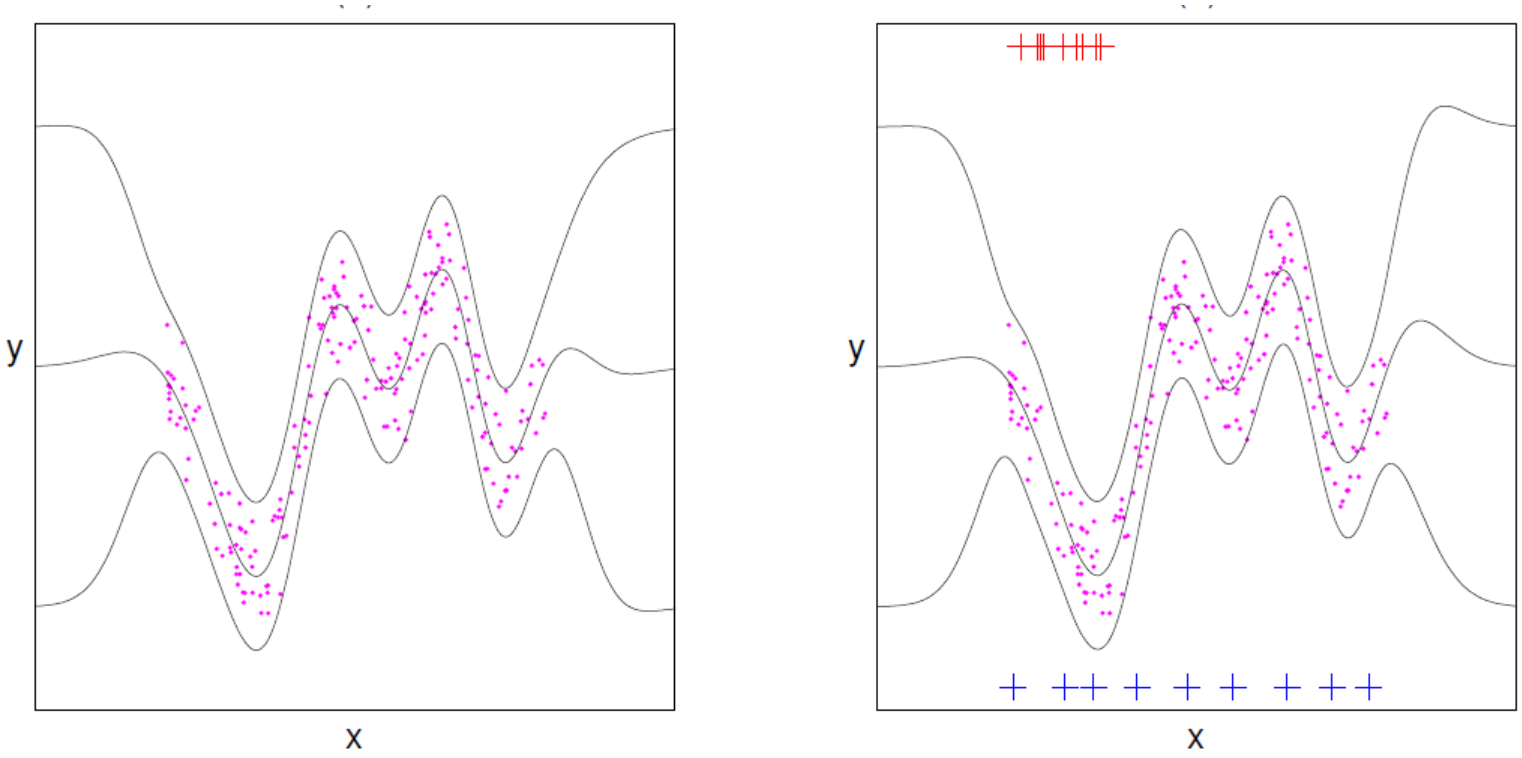
$$p(\mathbf{z}|\mathbf{y}) = \int p(\mathbf{z}|\mathbf{f})p(\mathbf{f}|\mathbf{y})d\mathbf{f}$$

We equivalently write $p(\mathbf{z}|\mathbf{y})$ as

$$p(\mathbf{z}|\mathbf{y}) = \int p(\mathbf{z}|\mathbf{f})p(\mathbf{f}|\mathbf{y})d\mathbf{f}$$

$$p(\mathbf{z}|\mathbf{y}) = \int p(\mathbf{z}|\mathbf{f}_m, \mathbf{f})p(\mathbf{f}|\mathbf{f}_m, \mathbf{y})p(\mathbf{f}_m|\mathbf{y})d\mathbf{f}d\mathbf{f}_m$$

pseudo-inputs support points active set



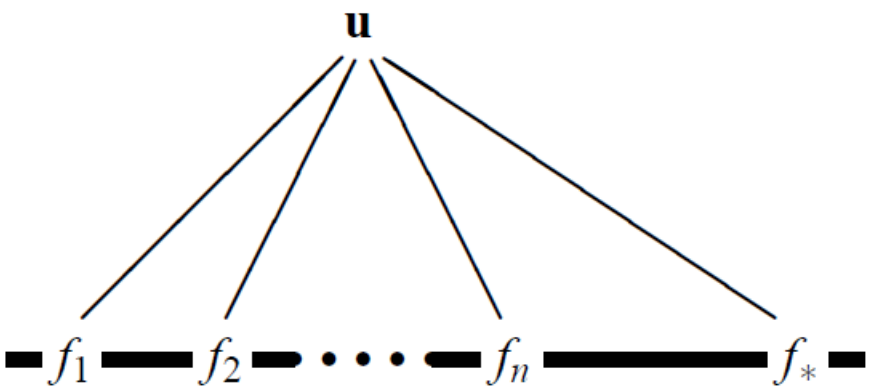
$$p(\mathbf{z}|\mathbf{y}) = \int p(\mathbf{z}|\mathbf{f}_m, \mathbf{f}) p(\mathbf{f}|\mathbf{f}_m, \mathbf{y}) p(\mathbf{f}_m|\mathbf{y}) d\mathbf{f} d\mathbf{f}_m$$

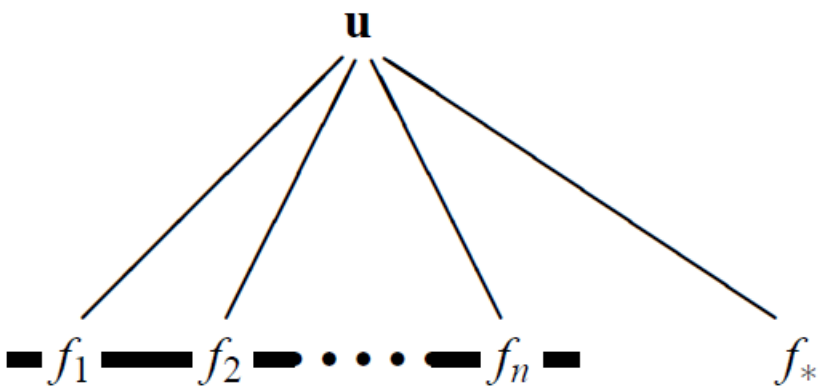
Suppose now that $\mathbf{f}_m$ is a sufficient statistic for the parameter $\mathbf{f}$ i.e. it holds

$$p(\mathbf{z}|\mathbf{f}_m, \mathbf{f}) = p(\mathbf{z}|\mathbf{f}_m)$$

A Unifying View of Sparse Approximate Gaussian Process Regression(JMLR2005)

$$q(\mathbf{z}) = \int p(\mathbf{z}|\mathbf{f}_m) p(\mathbf{f}|\mathbf{f}_m) \phi(\mathbf{f}_m) d\mathbf{f} d\mathbf{f}_m = \int p(\mathbf{z}|\mathbf{f}_m) \phi(\mathbf{f}_m) d\mathbf{f}_m = \int q(\mathbf{z}, \mathbf{f}_m) d\mathbf{f}_m$$





$$p(\mathbf{z}|\mathbf{y}) = \int p(\mathbf{z}|\mathbf{f}_m, \mathbf{f}) p(\mathbf{f}|\mathbf{f}_m, \mathbf{y}) p(\mathbf{f}_m|\mathbf{y}) d\mathbf{f} d\mathbf{f}_m$$

$$q(\mathbf{z}) = \int p(\mathbf{z}|\mathbf{f}_m) p(\mathbf{f}|\mathbf{f}_m) \phi(\mathbf{f}_m) d\mathbf{f} d\mathbf{f}_m = \int p(\mathbf{z}|\mathbf{f}_m) \phi(\mathbf{f}_m) d\mathbf{f}_m = \int q(\mathbf{z}, \mathbf{f}_m) d\mathbf{f}_m$$

$$m_{\mathbf{y}}^q(\mathbf{x}) = K_{\mathbf{x}m} K_{mm}^{-1} \boldsymbol{\mu} \qquad B = K_{mm}^{-1} A K_{mm}^{-1}$$

$$k_{\mathbf{y}}^q(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}, \mathbf{x}') - K_{\mathbf{x}m} K_{mm}^{-1} K_{m\mathbf{x}'} + K_{\mathbf{x}m} B K_{m\mathbf{x}'}$$

A principled procedure to specify ϕ and the inducing inputs X_m is to form the variational distribution $q(\mathbf{f})$ and the exact posterior $p(\mathbf{f}|\mathbf{y})$ on the training function values $\mathbf{f}$ and then minimize the distance(KL).

Equivalently, we can minimize a distance between the augmented true posterior $p(\mathbf{f}, \mathbf{f}_m|\mathbf{y})$ and the augmented variational posterior $q(\mathbf{f}, \mathbf{f}_m)$

$$q(\mathbf{f}, \mathbf{f}_m) = p(\mathbf{f}|\mathbf{f}_m) \phi(\mathbf{f}_m)$$

The augmented true posterior is associated with the augmented joint model

$$p(\mathbf{y}, \mathbf{f}, \mathbf{f}_m) = p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{f}_m)p(\mathbf{f}_m)$$

which is equivalent to the initial model $p(\mathbf{y}, \mathbf{f}) = p(\mathbf{f}|\mathbf{y})p(\mathbf{f})$

Notice that the conditional prior $p(\mathbf{f}|\mathbf{f}_m)$ and the marginal prior $p(\mathbf{f}_m)$ depend on the specific values of the inducing inputs X_m However, this dependence never affects the posterior $p(\mathbf{f}|\mathbf{y})$ or the marginal likelihood $p(\mathbf{y})$

The variational approach

$$p(\mathbf{f}, \mathbf{u}|\mathbf{y}) = \frac{p(\mathbf{f}|\mathbf{u})p(\mathbf{u})p(\mathbf{y}|\mathbf{f})}{p(\mathbf{y})} \qquad p(\mathbf{f}, \mathbf{y}) = \int \mathrm{d}\mathbf{u} p(\mathbf{f}, \mathbf{u}, \mathbf{y})$$

$$\mathrm{KL}(q(\mathbf{f},\mathbf{u})||p(\mathbf{f},\mathbf{u}|\mathbf{y})) = \int \mathrm{d}\mathbf{u}\,\mathrm{d}\mathbf{f} q(\mathbf{f},\mathbf{u}) \log \frac{q(\mathbf{f},\mathbf{u})}{p(\mathbf{f},\mathbf{u}|\mathbf{y})} \quad = \log p(\mathbf{y}) + \int \mathrm{d}\mathbf{u}\,\mathrm{d}\mathbf{f} q(\mathbf{f},\mathbf{u}) \log \frac{q(\mathbf{f},\mathbf{u})}{p(\mathbf{f},\mathbf{u},\mathbf{y})}$$

$$\log p(\mathbf{y}) = \int \mathrm{d}\mathbf{u}\,\mathrm{d}\mathbf{f} q(\mathbf{f},\mathbf{u}) \log \frac{p(\mathbf{f},\mathbf{u},\mathbf{y})}{q(\mathbf{f},\mathbf{u})} + \mathrm{KL}(q(\mathbf{f},\mathbf{u})||p(\mathbf{f},\mathbf{u}|\mathbf{y}))$$

$$\ln p(\boldsymbol{X}) = \mathcal{L}(q) + \mathrm{KL}(q \parallel p)$$

$\mathcal{F}(q(\mathbf{u})) = \int \mathrm{d}\mathbf{u}\,\mathrm{d}\mathbf{f} q(\mathbf{f},\mathbf{u}) \log \frac{p(\mathbf{f},\mathbf{u},\mathbf{y})}{q(\mathbf{f},\mathbf{u})}$ is the evidence lower bound (ELBO).

$$\log p(\mathbf{y}) = \log \int \mathrm{d}\mathbf{u}\,\mathrm{d}\mathbf{f} p(\mathbf{y},\mathbf{f},\mathbf{u}) = \log \int \mathrm{d}\mathbf{u}\,\mathrm{d}\mathbf{f} q(\mathbf{f},\mathbf{u}) \frac{p(\mathbf{y},\mathbf{f},\mathbf{u})}{q(\mathbf{f},\mathbf{u})} \geq \int \mathrm{d}\mathbf{u}\,\mathrm{d}\mathbf{f} q(\mathbf{f},\mathbf{u}) \log \frac{p(\mathbf{y},\mathbf{f},\mathbf{u})}{q(\mathbf{f},\mathbf{u})}$$

Jensen's inequality

$$\log p(\mathbf{y} \mid \mathbf{u}) \geq \mathbb{E}_{p(\mathbf{f} \mid \mathbf{u})} \left[\log p(\mathbf{y} \mid \mathbf{f})\right]$$

$$\log p(y|u) = \log \int p(y|f)p(f|u)df \geq \int p(f|u)\log p(y|f)df = \mathbb{E}_{p(f|u)}[\log p(y|f)]$$

This is in general intractable for the non-conjugate case. We nevertheless persist, recalling the standard variational equation:

$$\log p(\mathbf{y}) \geq \mathbb{E}_{q(\mathbf{u})} \left[\log p(\mathbf{y} \mid \mathbf{u})\right] - \mathrm{KL}\left[q(\mathbf{u}) || p(\mathbf{u})\right]$$

$$p(y, f, u) = p(y|f)p(f|u)p(u)$$

$$p(y, f, u) = p(y|f)p(f|u)p(u)$$

$$\log p(y) = \log \iint p(y, f, u) df du = \log \iint q(u) \frac{p(y, f, u)}{q(u)} df du$$

$$\log p(y) \geq \iint q(u) \frac{p(y|f)p(f|u)p(u)}{q(u)} df du = \iint q(u)p(y|f)p(f|u) df du + \int q(u) \frac{p(u)}{q(u)} du$$

$$= \int q(u)p(y|u)du - \mathrm{KL}[q(u)||p(u)] \geq \mathbb{E}_{q(u)}[\log p(y|u)] - \mathrm{KL}[q(u)||p(u)]$$

$$\log p(\mathbf{y}) = \log \int \mathrm{d}\mathbf{u}\, \mathrm{d}\mathbf{f} p(\mathbf{y}, \mathbf{f}, \mathbf{u}) = \log \int \mathrm{d}\mathbf{u}\, \mathrm{d}\mathbf{f} q(\mathbf{f}, \mathbf{u}) \frac{p(\mathbf{y}, \mathbf{f}, \mathbf{u})}{q(\mathbf{f}, \mathbf{u})} \geq \int \mathrm{d}\mathbf{u}\, \mathrm{d}\mathbf{f} q(\mathbf{f}, \mathbf{u}) \log \frac{p(\mathbf{y}, \mathbf{f}, \mathbf{u})}{q(\mathbf{f}, \mathbf{u})}$$

两种下界的最大化求解(其一)

$$\mathcal{F}(q(\mathbf{u})) = \int \mathrm{d}\mathbf{u}\,\mathrm{d}\mathbf{f} q(\mathbf{f},\mathbf{u}) \log \frac{p(\mathbf{f},\mathbf{u},\mathbf{y})}{q(\mathbf{f},\mathbf{u})}$$

$$q(\mathbf{f},\mathbf{u}) = p(\mathbf{f}|\mathbf{u})q(\mathbf{u})$$

$$\mathcal{F}(q(\mathbf{u})) = \int \mathrm{d}\mathbf{u}\,\mathrm{d}\mathbf{f} p(\mathbf{f}|\mathbf{u})q(\mathbf{u}) \log \frac{p(\mathbf{f}|\mathbf{u})p(\mathbf{u})p(\mathbf{y}|\mathbf{f})}{p(\mathbf{f}|\mathbf{u})q(\mathbf{u})} \qquad p(\mathbf{f},\mathbf{u}|\mathbf{y}) = \frac{p(\mathbf{f}|\mathbf{u})p(\mathbf{u})p(\mathbf{y}|\mathbf{f})}{p(\mathbf{y})}$$

$$= \int \mathrm{d}\mathbf{u}\,\mathrm{d}\mathbf{f} p(\mathbf{f}|\mathbf{u})q(\mathbf{u}) \log \frac{p(\mathbf{u})p(\mathbf{y}|\mathbf{f})}{q(\mathbf{u})}$$

$$= \int \mathrm{d}\mathbf{u}\,\mathrm{d}\mathbf{f} p(\mathbf{f}|\mathbf{u})q(\mathbf{u}) \log \frac{p(\mathbf{u})}{q(\mathbf{u})} + \int \mathrm{d}\mathbf{u}\,\mathrm{d}\mathbf{f} p(\mathbf{f}|\mathbf{u})q(\mathbf{u}) \log p(\mathbf{y}|\mathbf{f})$$

Consider the derivative of the ELBO w.r.t $q(\mathbf{u})$ with the addition of the Lagrange multiplier,

$$\frac{\mathrm{d}}{\mathrm{d}q(\mathbf{u})}\mathcal{F} + \lambda = \int \mathrm{d}\mathbf{f} p(\mathbf{f}|\mathbf{u}) \left[\log p(\mathbf{u}) - \log q(\mathbf{u}) - 1\right] + \int \mathrm{d}\mathbf{f} p(\mathbf{f}|\mathbf{u}) \log p(\mathbf{y}|\mathbf{f}) + \lambda$$

$$q(\mathbf{u}) = \frac{p(\mathbf{u})}{Z} \exp\left(\int \mathrm{d}\mathbf{f} p(\mathbf{f}|\mathbf{u}) \log p(\mathbf{y}|\mathbf{f})\right)$$

$$q(\mathbf{u}) = \frac{p(\mathbf{u})}{Z} \exp\left(\int \mathrm{d}\mathbf{f} p(\mathbf{f}|\mathbf{u}) \log p(\mathbf{y}|\mathbf{f})\right) \qquad = \int \mathrm{d}\mathbf{u}\,\mathrm{d}\mathbf{f} p(\mathbf{f}|\mathbf{u}) q(\mathbf{u}) \log \frac{p(\mathbf{u})p(\mathbf{y}|\mathbf{f})}{q(\mathbf{u})}$$

$$H(\mathbf{y}, \mathbf{u}) = \exp\left(\int \mathrm{d}\mathbf{f} p(\mathbf{f}|\mathbf{u}) \log p(\mathbf{y}|\mathbf{f})\right) \quad q(\mathbf{u}) = p(\mathbf{u}) H(\mathbf{y}, \mathbf{u})/Z$$

$$\mathcal{F}(q(\mathbf{u})) = \int \mathrm{d}\mathbf{u}\,\mathrm{d}\mathbf{f} p(\mathbf{f}|\mathbf{u}) q(\mathbf{u}) \log \frac{p(\mathbf{u})p(\mathbf{y}|\mathbf{f})}{p(\mathbf{u})H(\mathbf{y}, \mathbf{u})/Z}$$

$$= \int \mathrm{d}\mathbf{u}\,\mathrm{d}\mathbf{f} p(\mathbf{f}|\mathbf{u}) q(\mathbf{u}) \left[\log p(\mathbf{y}|\mathbf{f}) - \log H(\mathbf{y}, \mathbf{u}) + \log Z\right]$$

$$= \int \mathrm{d}\mathbf{u} q(\mathbf{u}) \left[\log Z \underbrace{- \log H(\mathbf{y}, \mathbf{u}) + \int \mathrm{d}\mathbf{f} p(\mathbf{f}|\mathbf{u}) \log p(\mathbf{y}|\mathbf{f})}_{=0}\right] = \log Z$$

$$\mathrm{M} = \int \mathrm{d}\mathbf{f} p(\mathbf{f}|\mathbf{u}) \log p(\mathbf{y}|\mathbf{f}) = \int \mathrm{d}\mathbf{f} \mathcal{N}(\mathbf{f}; K_{\mathbf{fu}} K_{\mathbf{uu}}^{-1} \mathbf{u}, K_{\mathbf{ff}} - K_{\mathbf{fu}} K_{\mathbf{uu}}^{-1} K_{\mathbf{uf}}) \log[\mathcal{N}(\mathbf{y}; \mathbf{f}, \sigma^2 I)]$$

$$= -\frac{1}{2\sigma^2} \mathrm{Tr}(\mathbf{B}) + \log[\mathcal{N}(\mathbf{y}; \mathbf{A}, \sigma^2 I)] \qquad \mathbf{A} = K_{\mathbf{fu}} K_{\mathbf{uu}}^{-1} \mathbf{u} \quad \mathbf{B} = K_{\mathbf{ff}} - K_{\mathbf{fu}} K_{\mathbf{uu}}^{-1} K_{\mathbf{uf}}$$

$$q(\mathbf{u}) = \mathcal{N}(K_{mn}(K_{nm}K_{mm}^{-1}K_{mn} + \sigma^2 I)\mathbf{y}, K_{mm} - K_{mn}(K_{nm}K_{mm}^{-1}K_{mn} + \sigma^2 I)K_{nm})$$
$$= \mathcal{N}(\sigma^{-2}K_{mm}\Sigma K_{mn}\mathbf{y}, K_{mm}\Sigma K_{mm}),$$

$$\Sigma = (K_{mm} + \sigma^{-2}K_{mn}K_{nm})^{-1}$$

The lower bound on the marginal likelihood is:

$$\mathcal{F}(q(\mathbf{u})) = \underbrace{\log \mathcal{N}(\mathbf{y}|0, \sigma^2 I + K_{nm}K_{mm}^{-1}K_{mn})}_{\text{DTC log marginal likelihood}} \underbrace{-\frac{1}{2\sigma^2}\text{Tr}(K_{nn} - K_{nm}K_{mm}^{-1}K_{mn})}_{\text{regulariser - avoid overfitting}}$$

$$F_V(X_m) = \log\left[N(\mathbf{y}|\mathbf{0}, \sigma^2 I + Q_{nn})\right] - \frac{1}{2\sigma^2}Tr(\widetilde{K})$$

$$Q_{nn} = K_{nm}K_{mm}^{-1}K_{mn}$$

$$\widetilde{K} = \text{Cov}(\mathbf{f}|\mathbf{f}_m) = K_{nn} - K_{nm}K_{mm}^{-1}K_{mn}$$

两种下界的最大化求解(其二)

$$\log p(y|u) = \log \int p(y|f)p(f|u)df \geq \int p(f|u)\log p(y|f)df = \mathbb{E}_{p(f|u)}[\log p(y|f)]$$

$$\log p(y) = \log \iint p(y,f,u)dfdu = \log \iint q(u)\frac{p(y,f,u)}{q(u)}dfdu$$

$$\log p(y) \geq \iint q(u)\frac{p(y|f)p(f|u)p(u)}{q(u)}dfdu = \iint q(u)p(y|f)p(f|u)dfdu + \int q(u)\frac{p(u)}{q(u)}du$$

$$= \int q(u)p(y|u)du - \mathrm{KL}[q(u)||p(u)] \geq \mathbb{E}_{q(u)}[\log p(y|u)] - \mathrm{KL}[q(u)||p(u)]$$

$$\log p(\mathbf{y} \mid \mathbf{u}) \geq \mathbb{E}_{p(\mathbf{f} \mid \mathbf{u})}\left[\log p(\mathbf{y} \mid \mathbf{f})\right]$$

$$\log p(\mathbf{y}) \geq \mathbb{E}_{q(\mathbf{u})}\left[\log p(\mathbf{y} \mid \mathbf{u})\right] - \mathrm{KL}\left[q(\mathbf{u})||p(\mathbf{u})\right]$$

$$\log p(\mathbf{y}) \geq \mathbb{E}_{q(\mathbf{u})}\left[\log p(\mathbf{y} \mid \mathbf{u})\right] - \mathrm{KL}\left[q(\mathbf{u})||p(\mathbf{u})\right] \geq \mathbb{E}_{q(\mathbf{u})}\left[\mathbb{E}_{p(\mathbf{f}|\mathbf{u})}\left[\log p(\mathbf{y}|\mathbf{f})\right]\right] - \mathrm{KL}\left[q(\mathbf{u})||p(\mathbf{u})\right]$$

$$= \mathbb{E}_{q(\mathbf{f})}\left[\log p(\mathbf{y} \mid \mathbf{f})\right] - \mathrm{KL}\left[q(\mathbf{u})||p(\mathbf{u})\right] \qquad q(\mathbf{f}) := \int p(\mathbf{f} \mid \mathbf{u}) q(\mathbf{u}) \mathrm{d}\mathbf{u}$$

$$q(\mathbf{u}) = \mathcal{N}(\mathbf{u} \mid \mathbf{m}, \mathbf{S}) \qquad q(\mathbf{f}) = \mathcal{N}(\mathbf{f} \mid \mathbf{A}\mathbf{m}, \mathbf{K}_{nn} + \mathbf{A}(\mathbf{S} - \mathbf{K}_{mm})\mathbf{A}^\top) \qquad \mathbf{A} = \mathbf{K}_{nm}\mathbf{K}_{mm}^{-1}$$

Since in the classification case the likelihood factors as $p(\mathbf{y} \mid \mathbf{f}) = \prod_{i=1}^{N} p(y_i \mid f_i)$

$$\log p(\mathbf{y}) \geq \sum_{n=1}^{N} \mathbb{E}_{q(f_n)}\left[\log p(y_n \mid f_n)\right] - \mathrm{KL}\left[q(\mathbf{u})||p(\mathbf{u})\right]$$

We are left with some one dimensional integrals of the log-likelihood, which can be computed by e.g. Gauss-Hermite quadrature.

$$\mathbf{S} = \mathbf{L}\mathbf{L}^\top$$

$$\frac{\partial}{\partial \mu}\mathbb{E}_{\mathcal{N}(x|\mu,\sigma^2)}\left[f(x)\right] = \mathbb{E}_{\mathcal{N}(x|\mu,\sigma^2)}\left[\frac{\partial}{\partial x}f(x)\right]$$
$$\frac{\partial}{\partial \sigma^2}\mathbb{E}_{\mathcal{N}(x|\mu,\sigma^2)}\left[f(x)\right] = \frac{1}{2}\mathbb{E}_{\mathcal{N}(x|\mu,\sigma^2)}\left[\frac{\partial^2}{\partial x^2}f(x)\right]$$

These derivatives also have to be computed by quadrature methods, after which derivatives with respect to $\mathbf{m}, \mathbf{L}, \mathbf{Z}$

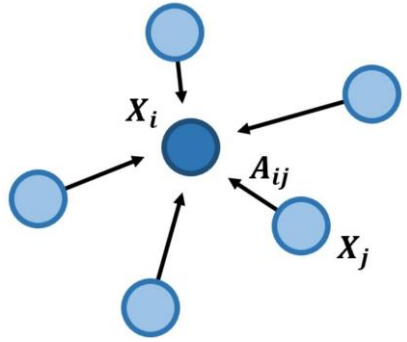
基于Graph的SVGP方法对于Graph结构的利用

图卷积的核心思想是利用边的信息对节点进行聚合，从而生成新的节点表示。

加权平均法

$$aggregate(X_i) = AX = \Sigma_{j=1}^{N} A_{ij} X_j$$

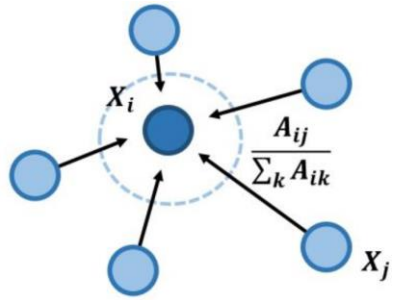
$$aggregate(X_i) = (A+I)X = \Sigma_{j=1}^{N} A_{ij} X_j + 1 \times X_i$$



归一化的加权平均法

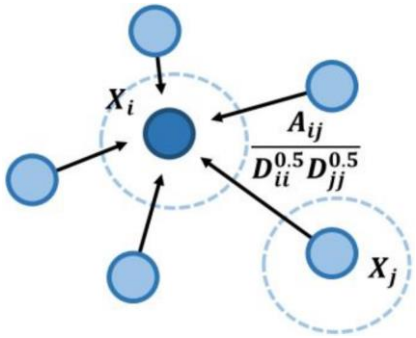
$$\begin{aligned} aggregate(X_i) &= D^{-1}(A+I)X \\ &= \Sigma_{k=1}^{N} D_{ik}^{-1} \Sigma_{j=1}^{N} A_{ij} X_j + \Sigma_{k=1}^{N} D_{ik}^{-1} \times X_i \\ &= \Sigma_{j=1}^{N} D_{ii}^{-1} A_{ij} X_j + D_{ii}^{-1} \times X_i \\ &= \Sigma_{j=1}^{N} \frac{A_{ij}}{D_{ii}} X_j + \frac{1}{D_{ii}} X_i \\ &= \Sigma_{j=1}^{N} \frac{A_{ij}}{\Sigma_{k=1}^{N} A_{ik}} X_j + \frac{1}{D_{ii}} X_i \end{aligned}$$

离群较远或者度较小的节点在聚合后特征较小



对称归一化的加权平均法

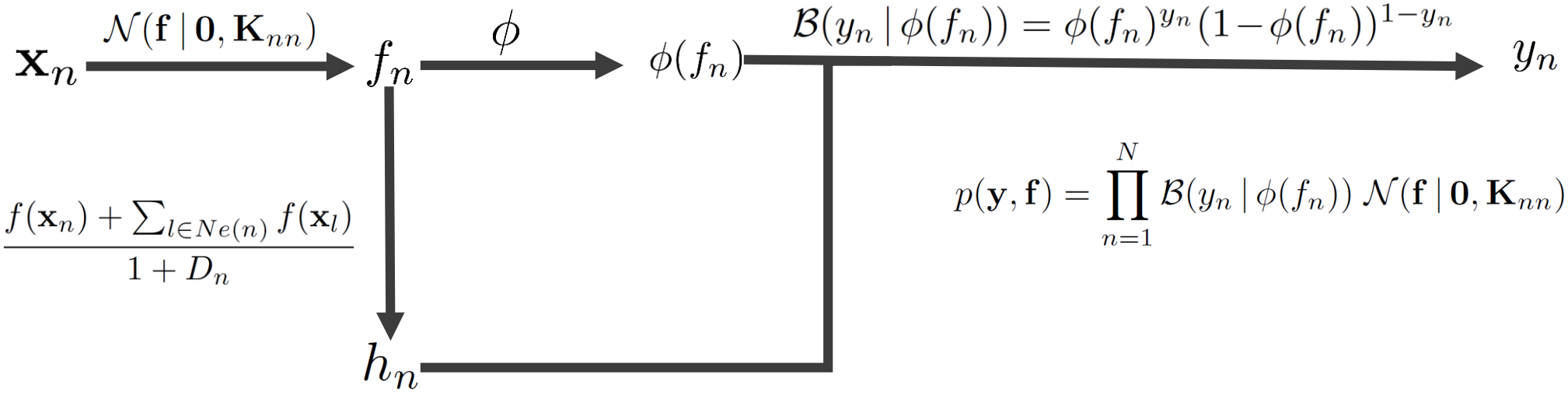
$$
\begin{aligned}
aggregate(X_i) &= D^{-0.5}\tilde{A}D^{-0.5}X \\
&= \Sigma_{k=1}^{N} D_{ik}^{-0.5} \Sigma_{j=1}^{N} \tilde{A}_{ij} X_j \Sigma_{l=1}^{N} D_{il}^{-0.5} \\
&= \Sigma_{j=1}^{N} D_{ii}^{-0.5} A_{ij} X_j D_{jj}^{-0.5} \\
&= \Sigma_{j=1}^{N} \frac{1}{D_{ii}^{0.5}} A_{ij} \frac{1}{D_{jj}^{0.5}} X_j
\end{aligned}
$$



$$\mathbf{X}_n \xrightarrow{\mathcal{N}(\mathbf{f} \mid \mathbf{0}, \mathbf{K}_{nn})} f_n \xrightarrow{\phi} \phi(f_n) \xrightarrow{\mathcal{B}(y_n \mid \phi(f_n)) = \phi(f_n)^{y_n}(1-\phi(f_n))^{1-y_n}} y_n$$

$$h_n = \frac{f(\mathbf{x}_n) + \sum_{l \in Ne(n)} f(\mathbf{x}_l)}{1 + D_n}$$

$$p(\mathbf{y}, \mathbf{f}) = \prod_{n=1}^{N} \mathcal{B}(y_n \mid \phi(f_n)) \, \mathcal{N}(\mathbf{f} \mid \mathbf{0}, \mathbf{K}_{nn})$$



$$p_\theta(\mathbf{Y}, \mathbf{h} | \mathbf{X}, \mathbf{A}) = p_\theta(\mathbf{h} | \mathbf{X}, \mathbf{A}) \prod_{n=1}^{N} p(y_n | h_n) \qquad p_\theta(\mathbf{h} | \mathbf{X}, \mathbf{A}) = \mathcal{N}(\mathbf{0}, \mathbf{P}\mathbf{K}_{\mathbf{X}\mathbf{X}}\mathbf{P}^\top)$$

$$\mathbf{P}\Phi_{\mathbf{X}}\Phi_{\mathbf{X}}^\top\mathbf{P}^\top$$

$$\widehat{\Phi}_{\mathbf{X}} = \mathbf{P}\Phi_{\mathbf{X}} = (\mathbf{I} + \mathbf{D})^{-1}\mathbf{D}\Phi_{\mathbf{X}} + (\mathbf{I} + \mathbf{D})^{-1}(\mathbf{I} - \mathbf{L})\Phi_{\mathbf{X}}$$

距离度量

小明（160,60000）；小王（160,59000）；小李（170，60000）

$$\mathbf{x} = (x_1, x_2, \ldots, x_N)^T \ , \ \mathbf{y} = (y_1, y_2, \ldots, y_N)^T \qquad D_M = \sqrt{(\mathbf{x} - \mathbf{y})^T \cdot S^{-1} \cdot (\mathbf{x} - \mathbf{y})}$$

$$D_M = \sqrt{\sum_{i=1}^{N} \frac{(x_i - y_i)^2}{s_i^2}} \qquad Cov(X, Y) = E[(X - E(X)) \cdot (Y - E(Y))]$$

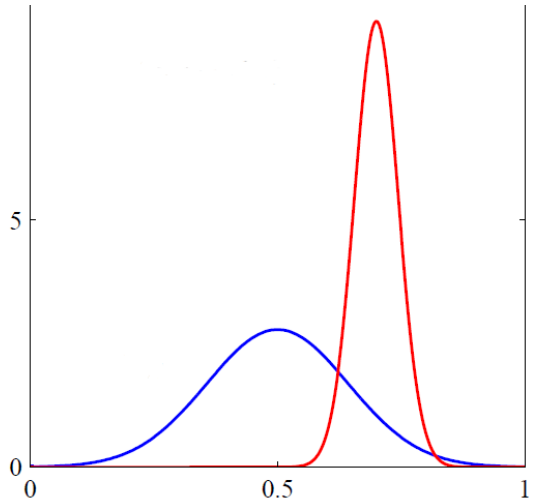
马氏距离是一种有效的计算两个未知样本集的相似度的方法。与欧氏距离不同的是它考虑到各种特性之间的联系（例如：一条关于身高的信息会带来一条关于体重的信息，因为两者是有关联的）并且是尺度无关的（scale-invariant），即独立于测量尺度。

(多)分类(K=7)问题中,模型预测出的每个类别的可能性得分分别服从某高斯分布

假设模型对图中两个节点的隐函数值推断如下：

μ	12	5	7	1	-3	17	5
σ	2	23	4	7	52	2	36

μ	5	7	9	0	17	55	4
σ	11	9	5	3	25	14	1



$$\mathbf{X}_n \xrightarrow{\mathcal{N}(\mathbf{f}\,|\,\mathbf{0},\mathbf{K}_{nn})} f_n \longrightarrow \phi(f_n) \xrightarrow{\mathcal{B}(y_n\,|\,\phi(f_n)) = \phi(f_n)^{y_n}(1-\phi(f_n))^{1-y_n}} y_n$$

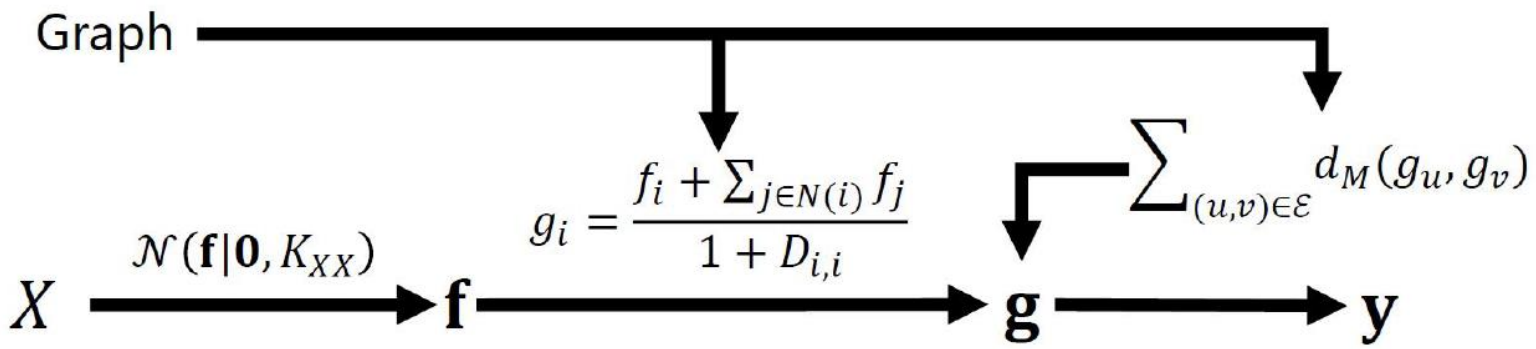
$$\phi(x) = \int_{-\infty}^{x} \mathcal{N}(a\,|\,0,1)\mathrm{d}a$$

Symmetric Mahalanobis distance for defining distance between two nodes in the graph.

$$d_M(u,v) = (\boldsymbol{\mu}_u - \boldsymbol{\mu}_v)^T(\boldsymbol{\Sigma}_u^{-1} + \boldsymbol{\Sigma}_v^{-1})(\boldsymbol{\mu}_u - \boldsymbol{\mu}_v)$$

Enforcing neighboring nodes to be close to each other

$$L_{\text{smooth}} = \sum_{(u,v)\in\mathcal{E}} (\boldsymbol{\mu}_u - \boldsymbol{\mu}_v)^T(\boldsymbol{\Sigma}_u^{-1} + \boldsymbol{\Sigma}_v^{-1})(\boldsymbol{\mu}_u - \boldsymbol{\mu}_v)$$



Rates of Convergence for Sparse Variational Gaussian Process Regression ICML2019

The computational cost seems to be linear in N, the true complexity of the algorithm depends on how M must increase to ensure a certain quality of approximation.

We show that with high probability the KL divergence can be made arbitrarily small by growing M more slowly than N.

Our results show that as datasets grow, Gaussian process posteriors **can truly be approximated cheaply,** and **provide a concrete rule** for how to increase M in continual learning scenarios.

实验

	Type	N_{nodes}	N_{edges}	$N_{label\_cat.}$	$D_{features}$	Label Rate
Cora	Citation	2,708	5,429	7	1,433	0.052
Citeseer	Citation	3,327	4,732	6	3,703	0.036
Pubmed	Citation	19,717	44,338	3	500	0.003

表 1: cora 10 次实验及其均值,GGP(NIPS18) VS GGP-M(改进)

方法	1	2	3	4	5	6	7	8	9	10	平均值
GGP	81.1	80.8	80.9	80.9	80.9	80.9	80.8	81.0	81.0	81.2	80.9
GGP-M	82.2	81.8	82.7	82.1	82.2	82.2	82.1	82.7	82.6	82.2	**82.3**
GGP(x+500)	84.6	84.7	85.1	84.6	84.8	85.0	84.8	84.8	84.8	84.7	84.8
GGP-M(x+500)	85.4	84.9	86.1	84.9	85.2	85.7	85.2	84.6	84.7	85.8	**85.3**

表 2: citeseer 10 次实验及其均值,GGP(NIPS18) VS GGP-M(改进)

方法	1	2	3	4	5	6	7	8	9	10	平均值
GGP	69.2	69.1	69.1	69.1	68.6	69.5	68.5	69.4	68.8	69.0	69.0
GGP-M	69.5	69.5	70.3	70.2	69.7	70.5	69.7	70.3	70.4	70.6	**70.1**

表 3: pubmed 4×10 次实验及其均值,GGP(NIPS18) VS GGP-M(改进)

方法	1	2	3	4	5	6	7	8	9	10	平均值
GGP(n=10000)	77.0	77.0	77.2	77.1	76.4	76.9	76.8	77.0	77.1	76.5	**77.1**
GGP-M(n=10000)	76.1	76.7	76.0	76.4	76.6	76.9	76.6	77.4	76.9	76.5	76.6
GGP(dell-1)	76.2	76.3	76.6	76.7	76.7	76.7	75.9	76.1	76.5	76.3	76.4
GGP(dell-2)	76.4	75.5	76.4	77.2	76.4	76.5	76.7	76.2	76.1	76.8	76.4
GGP(lenovo-1)	77.1	76.0	76.5	76.9	76.2	76.6	76.8	76.6	76.0	76.7	76.5
GGP(lenovo-2)	76.5	75.9	76.2	75.9	76.3	76.2	76.7	76.0	76.6	76.5	76.3
GGP-M	77.1	76.6	76.8	77.2	76.4	76.3	77.2	76.0	77.0	77.5	**76.8**