Gaussian Process

Gaussian / Normal Distribution

- Univariate Form

$$X \sim N(\mu, \sigma^2)$$
$$p(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{(x-\mu)^2}{2\sigma^2})$$

- Multivariate Form

$$X \sim N(X|\mu, \Sigma)$$
$$X \in \mathbb{R}^d \quad \mu \in \mathbb{R}^d$$
$$\Sigma \in \mathbb{R}^{d\times d} \quad \Sigma_{ij} = E(x_i - \mu_i)(x_j - \mu_j) = cov(x_i, x_j)$$
$$p(X) = (\frac{1}{\sqrt{2\pi}})^d \cdot \frac{1}{|\Sigma|^{\frac{1}{2}}} \cdot \exp(-\frac{1}{2}(X-\mu)^T\Sigma^{-1}(X-\mu))$$

  - Σ is a covariance matrix, determines how x_i, x_j increase together or not
- CLT
  - When n independent random variables are summed up, their normalized sum tends a *Gaussian distributed random variable*, even if these original random variables are not Gaussian
- Any **linear combination** of Gaussian distributed random variables follow a Gaussian distribution
- **Concatenation** of Gaussian distributed random variables result in a multivariate Gaussian distributed random variable
- Given $x = \begin{pmatrix} x_a \\ x_b \end{pmatrix} \in \mathbb{R}^{a+b}$, if

$$x \sim N(\begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}, \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix})$$

  then
  1. $x_a \sim N(\mu_a, \Sigma_{aa})$, $x_b \sim N(\mu_b, \Sigma_{bb})$
  2. $P(x_a|x_b) = N(x_a|\mu_{a|b}, \Sigma_{a|b})$
     $\mu_{a|b} = \mu_a + \Sigma_{ab}\Sigma_{bb}^{-1}(x_b - \mu_b)$
     $\Sigma_{a|b} = \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}$

Different views of Linear Regression

ERM view

- the original form:

$$\min_w \sum_{i=1}^{n}(y_i - w^T\phi(x_i))^2$$

MLE view

- $y = w^T\phi(x) + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$
  - ϵ is a Gaussian noise
- $P(y|x; w, \sigma^2) = N(y|w^T\phi(x), \sigma^2)$
  - w is the parameter
  - σ^2 is a hyperparameter
- MLE form:

$$
\begin{aligned}
&\max_w \sum_{i=1}^{n} \log P(y_i|x_i) \\
\Leftrightarrow &\max_w \sum_{i=1}^{n} \log\Big(\frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{(y_i - w^T\phi(x_i))^2}{2\sigma^2})\Big) \\
\Leftrightarrow &\max_w \sum_{i=1}^{n} -\frac{1}{2}\log 2\pi\sigma^2 - \frac{(y_i - w^T\phi(x_i))^2}{2\sigma^2} \\
\Leftrightarrow &\min_w \sum_{i=1}^{n} (y_i - w^T\phi(x_i))^2
\end{aligned}
$$

MAP view

- **Maximum a Posteriori**
  - Previously, we treat w as fixed(constant) parameter. In Bayesian viewpoint, the world is uncertain, even the parameter w are random variables
  - Select the mode of the posteriori distribution as a point estimation
- Prior probability:

$$
P(w) = N(w|0, \sigma_w^2 I) \quad w \in \mathbb{R}^d
$$

  - $\mu = 0$ is a **prior belief**
  - 我们先验地认为，w的期望为0，方差越小，说明我们对w的取值在0附近这一事件越自信
- Estimation for *Ridge Regression*:

$$
P(y|x, w; \sigma^2, \sigma_w^2) = N(y|w^T\phi(x), \sigma^2)
$$

  From Bayesian Expectation Equation:

$$
P(w|y, x) = \frac{P(y|x, w)P(w|x)}{P(y|x)}
$$

  $P(y|x)$ is not dependent on w, let it be z, $P(w|x)$ is actually $P(w)$, so

$$
\begin{aligned}
P(w|y, x) &= \frac{1}{z}\cdot\prod_{i=1}^{n} P(y_i|x_i, w)P(w) \\
&= \frac{1}{z}\cdot(\frac{1}{\sqrt{2\pi}\sigma})^n \cdot \exp(-\frac{\sum_{i=1}^{n}(y_i - w^T\phi(x_i))^2}{2\sigma^2})\cdot(\frac{1}{\sqrt{2\pi}\sigma_w})^d \cdot \exp(-\frac{w^T w}{2\sigma_w^2}) \\
\max_w P(w|y, x) &\Leftrightarrow \max_w -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - w^T\phi(x_i))^2 - \frac{1}{2\sigma_w^2}w^T w \\
&\Leftrightarrow \min_w \sum_{i=1}^{n}(y_i - w^T\phi(x_i))^2 + \frac{\sigma^2}{\sigma_w^2}\|w\|^2
\end{aligned}
$$

Stochastic (Random) Process

- A collection of (infinitely many) random variables along on index set($\mathbb{N}$ or $\mathbb{R}$ or $\mathbb{R}^d$...)
- Infinitely many $\{(x_i, y_i)\}$ specify a distribution and function $y(x)$. Each sample of $\{(x_i, y_i)\}$ forms a *deterministic function* $y(x)$
- We can specify a **Random Process** by specifying the **joint distribution** of all random variables

Gaussian Process (GP)

- We specify the joint distribution over **any finite collection** of variables and require the joint distribution to be Gaussian
- $\{x_1, \ldots, x_n\}$ is **any** set of n points in the index set with sampled value $\{y_1, \ldots, y_n\}$, then

$$
GP \Leftrightarrow y_1, \ldots, y_n \text{ have a Multivariate Gaussian Distribution}
$$

- We need to specify $\mu_i = \text{mean}(x_i)$ and $\Sigma_{ij} = k(x_i, x_j)$ to determine a GP

$$
y(x) \sim GP(\text{mean}(x), k(\cdot, \cdot))
$$

- we usually use $\text{mean}(x)=0$ (prior belief), which means that $y=w^T\phi(x)=0$ as $w\sim N(0,\sigma_w^2 I)$
- we only need $k(x_i,x_j)=\exp(-\frac{\|x_i-x_j\|^2}{2\sigma^2})$
  - $k(x_i,x_j)$ 越大，表明x_i和x_j越接近，就越有可能同增减
- Suppose n training points $\{x_1,\dots,x_n\}$, let K be Gram matrix, $K_{ij}=k(x_i,x_j)$
  - k is a valid kernel only if $K\succeq 0$ for any $\{x_1,\dots,x_n\}$
  - Now given a new test point x^*, we can compute joint distribution of $\begin{pmatrix} y^* \\ y \end{pmatrix}\in\mathbb{R}^{n+1}$

$$P(\begin{pmatrix} y^* \\ y \end{pmatrix})=N(y|0,\begin{pmatrix} k(x^*,x^*) & k^T(x^*) \\ k(x^*) & K \end{pmatrix}),\quad k(x^*)=\begin{pmatrix} k(x^*,x_1) \\ \vdots \\ k(x^*,x_n) \end{pmatrix}$$
$$\begin{aligned} P(y^*|y)&=N(y^*|\mu^*,\Sigma^*) \\ \mu^*&=0+k^T(x^*)K^{-1}(y-0)=k^T(x^*)K^{-1}y \\ \Sigma^*&=k(x^*,x^*)-k^T(x^*)K^{-1}k(x^*) \end{aligned}$$

  So we have the dual form of Linear Regression:

$$f(x^*)=\mu^*=k^T(x^*)K^{-1}y$$

  - μ^* is called **posterori mean**
- In a more realistic setting, we can only observe $\hat{y}=y+\epsilon,\quad \epsilon\sim N(0,\sigma^2 I)$, then

$$\begin{aligned} P(\hat{y})&=N(y|0,K+\sigma^2 I) \\ \mu^*&=k^T(x^*)(K+\sigma^2 I)^{-1}y \\ \Sigma^*&=k(x^*,x^*)-k^T(x^*)(K+\sigma^2 I)^{-1}k(x^*) \end{aligned}$$

  because

$$\begin{aligned} \text{cov}(\hat{y}_i,\hat{y}_j)&=E\hat{y}_i\hat{y}_j-E\hat{y}_iE\hat{y}_j \\ &=E(y_i+\epsilon_i)(y_j+\epsilon_j) \\ &=Ey_iy_j+E\epsilon_i\epsilon_j+Ey_i\epsilon_j+Ey_j\epsilon_i \\ &=Ey_iy_j+E\epsilon_i\epsilon_j \\ &=\text{cov}(y_i,y_j)+\text{cov}(\epsilon_i,\epsilon_j) \\ &=k(x_i,x_j)+\sigma^2 1(i=j) \end{aligned}$$

- 也就是说，每次采样一些x(服从高斯分布)及其对应的y，得到一个方程$y(x)$，那么多次采样后，方程$y(x)$也服从高斯分布

Recall Ridge Regression

- recall the MAP view of Ridge Regression, we have $y_i=w^T\phi(x_i)+\epsilon_i$, and vectorize this formula:

$$y=w^T\Phi+\epsilon\quad w\sim N(0,\sigma_w^2 I)\quad \epsilon\sim N(0,\sigma^2 I)$$

  For y_i and y_j:

$$\begin{aligned} \text{cov}(y_i,y_j)&=Ey_iy_j-Ey_iEy_j \\ &=E(w^T\phi(x_i)+\epsilon_i)(w^T\phi(x_j)+\epsilon_j) \\ &=E(\phi^T(x_i)ww^T\phi(x_j))+E\epsilon_i\epsilon_j \\ &=\phi^T(x_i)Eww^T\phi(x_j)+\sigma^2 1(i=j) \\ &=\phi^T(x_i)\text{cov}(ww^T)\phi(x_j)+\sigma^2 1(i=j) \\ &=\sigma_w^2\phi^T(x_i)\phi(x_j)+\sigma^2 1(i=j) \\ &=\sigma_w^2 k(x_i,x_j)+\sigma^2 1(i=j) \end{aligned}$$

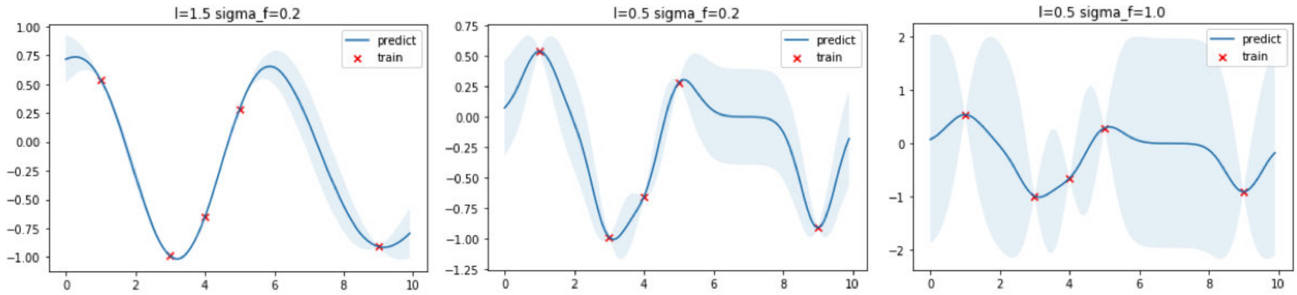
  So we have:

$$y\sim N(0,\Sigma)\quad \Sigma=\sigma_w^2 K+\sigma^2 I$$

- For a new test point $x^*\in\mathbb{R}^{n'}$

$$\mu^* = \sigma_w^2 k^T(x^*)(\sigma_w^2 K + \sigma^2 I)^{-1} y$$
$$= k^T(x^*)(K + \frac{\sigma^2}{\sigma_w^2} I)^{-1} y \quad (\text{let } \frac{\sigma^2}{\sigma_w^2} \text{ be } \lambda)$$

- **贝叶斯视角下的线性回归本质是高斯过程**
- GP connects ERM of Ridge Regression, MAP of Linear Regression, and dual solution of Ridge Regression
- **GP for Regression(GPR)**: typically take *RBF* kernel $k(x, x') = \exp(-\frac{\|x-x'\|^2}{2l^2})$
  - l is called *length scale* or *temperature*
  - $\mu(x_i)$ not necessarily equals y_i (because ϵ_i and $\text{cov}(y_i, y_j)$)
  - when $l \to 0$ and $\sigma = 0$, $\mu(x_i) = y_i$



Bayesian Optimization (BO)

- For *Black box* functions (cannot use gradient descent)
  1. Randomly sample n points $x_1, \ldots, x_n$ and corresponding $y_1, \ldots, y_n$
  2. fit a GP
  3. use some acquisition function $a(x)$ to select next point to evaluate
     - possible $a(x)$: *lower confidence bound* $\mu(x) - \kappa\sigma(x)$
  4. Use all x, y to fit a new GP
  5. Repeat until reaching a budget
- For hyperparameter optimization, the input is possible value of hyperparameter, and the output is the validation error