Lecture 8: Kernel Linear Regression & Gaussian Processes

Week 8

Lecturer: Tianyu Wang

1 Kernel Ridge Regression

Let k be a symmetric and positive definition kernel function, and let $\mathcal{H}$ be the reproducing kernel Hilbert space associated with k. Consider a dataset $\{(x_i, y_i)\}_{i=1}^n$ and the following regression objective

$$\frac{1}{2}\sum_{j=1}^{n}(f(x_j)-y_j)^2+\frac{\lambda}{2}\|f\|_{\mathcal{H}}^2.$$

By the representer theorem, the solution to the above objective must be of the form $f(\cdot) = \sum_{i=1}^n \alpha_i k(x_i, \cdot)$ for some $\alpha_1, \alpha_2, \cdots, \alpha_n \in \mathbb{R}$. Next we will solve for α_i.

For $f(\cdot) = \sum_{i=1}^n \alpha_i k(x_i, \cdot)$, we have

$$f(x_j) = \left\langle \sum_{i=1}^{n} \alpha_i k(x_i, \cdot), k(x_j, \cdot) \right\rangle = \sum_{i=1}^{n} \alpha_i k(x_i, x_j), \quad \forall j,$$

and

$$\|f\|_{\mathcal{H}}^2 = \left\langle \sum_{i=1}^{n} \alpha_i k(x_i, \cdot), \sum_{j=1}^{n} \alpha_j k(x_j, \cdot) \right\rangle = \sum_{i,j=1}^{n} \alpha_i \alpha_j k(x_i, x_j).$$

The loss in terms of α is

$$\begin{aligned} l(\alpha) &:= \frac{1}{2}\sum_{j=1}^{n}(f(x_j)-y_j)^2+\frac{\lambda}{2}\|f\|_{\mathcal{H}}^2 \\ &= \frac{1}{2}\sum_{j=1}^{n}\left(\sum_{i=1}^{n}\alpha_i k(x_i,x_j)-y_j\right)^2+\frac{\lambda}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j k(x_i,x_j), \end{aligned}$$

for some hyperparameter $\lambda > 0$. Recall the representer theorem at this point.

In matrix form, $l(\alpha)$ can be written as

$$l(\alpha) = \frac{1}{2}\|K\alpha - \mathbf{y}\|_2^2 + \frac{\lambda}{2}\alpha^\top K\alpha,$$

where K is the matrix such that $K_{ij} = k(x_i, x_j)$, $\mathbf{y}$ is the vector of labels, and $\alpha = [\alpha_1, \alpha_2, \cdots, \alpha_n]^\top$. The matrix K is called the *Gram matrix*.

Set $\nabla l(\alpha) = 0$ gives

$$K\left(K\alpha - \mathbf{y}\right) + \lambda K\alpha = 0,$$

which gives

$$\alpha = \left(K^2 + \lambda K\right)^{-1} K\mathbf{y} = \left(K + \lambda I\right)^{-1} \mathbf{y}.$$

Compare this to ridge regression.

1.1 Random Fourier Features

A drawback of kernel linear regression is that the complexity can grow with number of data. If we use the Gaussian kernel, the Gram matrix is always full rank and the size of the Gram matrix grow with number of data points. It is expensive to invert this matrix. One can use random Fourier features to reduce the computational cost.

By the Mercer's theorem, we know that, for any symmetric positive definite kernel k, there exists a feature map ϕ, such that $\phi(x)^\top \phi(z) = k(x, z)$ for all x, z. Here we allow the vectors $\phi(x)$, $\phi(z)$ to be (countably) infinitely dimensional.

Consider the Gaussian kernel $k(x, z) = \exp\left(-\gamma \|x - z\|_2^2\right)$, where γ is a hyperparameter. For simplicity, we let $\gamma = \frac{1}{2}$. Cases for other values of γ can be derived in similar ways. We will construct a finite-dimensional computationally-friendly feature map ϕ, so that $\phi(z)^\top \phi(x)$ is an approximation for $k(z, x)$.

Let $w \sim \mathcal{N}(0, I_d)$ be a standard Gaussian random vector. For any $z \in \mathbb{R}^d$, we have

$$\begin{aligned}
\mathbb{E}\left[\exp\left(-iw^\top z\right)\right] &= \frac{1}{(2\pi)^{\frac{d}{2}}} \int_w \exp\left(-\frac{1}{2} w^\top w\right) \exp\left(-iw^\top z\right), dw \\
&= \frac{1}{(2\pi)^{\frac{d}{2}}} \int_w \exp\left(-\frac{1}{2} w^\top w - iw^\top z + \frac{1}{2} z^\top z - \frac{1}{2} z^\top z\right), dw \\
&= \exp\left(-\frac{1}{2} z^\top z\right) \frac{1}{(2\pi)^{\frac{d}{2}}} \int_w \exp\left(-\frac{1}{2}\left(w + iz\right)^\top \left(w + iz\right)\right), dw \\
&= \exp\left(-\frac{1}{2} z^\top z\right).
\end{aligned}$$

Therefore, for any $x, y \in \mathbb{R}^d$, it holds that

$$\mathbb{E}\left[\exp\left(-iw^\top (x - y)\right)\right] = \exp\left(-\frac{\|x - y\|_2^2}{2}\right).$$

By the Euler's formula, we have

$$\mathbb{E}\left[\cos\left(w^\top (x - y)\right) + i \sin\left(-w^\top (x - y)\right)\right] = \exp\left(-\frac{\|x - y\|_2^2}{2}\right).$$

Since the imaginary part is zero, we have

$$\mathbb{E}\left[\cos\left(w^\top(x-y)\right)\right] = \exp\left(-\frac{\|x-y\|_2^2}{2}\right).$$

We can now construct the random features. Draw $i.i.d.$ random variables $w_1, w_2, \cdots, w_p \sim \mathcal{N}(0, I_d)$ and $i.i.d.$ random variables $b_1, b_2, \cdots, b_p \sim \text{Uniform}(0, 2\pi)$. For any $x \in \mathbb{R}^d$, let

$$\phi(x) = \frac{1}{\sqrt{p}}\left[\cos(w_1^\top x + b_1), \cos(w_2^\top x + b_2), \cdots, \cos(w_p^\top x + b_p)\right]^\top.$$

The vector $\phi(x)$ is called *random Fourier features*.

Then we have, for any $x, y \in \mathbb{R}^d$,

$$\phi(x)^\top \phi(y) = \frac{1}{p}\sum_{i=1}^{p} \cos(w_i^\top x + b_i)\cos(w_i^\top y + b_i).$$

Note that, for any i,

$$\cos(w_i^\top x + b_i)\cos(w_i^\top y + b_i) = \frac{1}{2}\cos(w_i^\top(x-y)) + \frac{1}{2}\cos(w_i^\top(x+y) + 2b_i).$$

Since

$$\mathbb{E}_{b_i}\left[\cos(w_i^\top(x+y) + 2b_i)\right] = 0,$$

we have

$$\mathbb{E}\left[\cos(w_i^\top x + b_i)\cos(w_i^\top y + b_i)\right] = \frac{1}{2}\mathbb{E}\left[\cos(w_i^\top(x-y))\right] = \exp\left(-\frac{1}{2}\|x-y\|_2^2\right).$$

This means

$$\mathbb{E}\left[\phi(x)^\top\phi(y)\right] = \exp\left(-\frac{1}{2}\|x-y\|_2^2\right).$$

2 Gaussian Process Regression

A Gaussian process is a distribution over functions. For any $\mathcal{X}$, let $m : \mathcal{X} \mapsto \mathbb{R}$ be a (continuous) function over $\mathcal{X}$, and let $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ be a symmetric positive definite kernel. A function f defined over $\mathcal{X}$ is said to follow the Gaussian process $\mathcal{GP}(m(\cdot), k(\cdot,\cdot))$ if for any $n \in \mathbb{N}_+$ and $x_1, x_2, \cdots, x_n \in \mathcal{X}$,

$$[f(x_1), f(x_2), \cdots, f(x_n)]^\top \sim \mathcal{N}(\mathbf{m}, K),$$

where $\mathbf{m} = [m(x_1), m(x_2), \cdots, m(x_n)]^\top$ and $K_{ij} = k(x_i, x_j)$ is the Gram matrix.

We can use Gaussian processes to fit/learn functions. Given data $\{(x_i, y_i)\}_{i=1}^n$, let $\mathbf{y} = [y_1, y_2, \cdots, y_n]^\top$, let K be the Gram matrix. Consider a new point x, and let $\boldsymbol{k} = [k(x_1, x), k(x_2, x), \cdots, k(x_n, x)]^\top$. Let y be the value associated with x.

Consider the following inference model. The prior is $f \sim \mathcal{GP}(0, k(\cdot, \cdot))$ and the data is corrupted with noise: $y = f + \epsilon$ with $\epsilon \sim \mathcal{N}(0, \sigma^2)$ being an independent Gaussian noise. Then from the definition of Gaussian processes, the joint distribution under this prior is

$$\begin{pmatrix} \mathbf{y} \\ f(x) \end{pmatrix} \sim \mathcal{N}\left(0, \begin{matrix} K + \sigma^2 I & \boldsymbol{k} \\ \boldsymbol{k}^\top & k(x,x) \end{matrix}\right)$$

Marginalization gives

$$f(x)|x, \{(x_i, y_i)\}_{i=1}^n \sim \mathcal{N}\left(\boldsymbol{k}^\top (K + \sigma^2 I)^{-1} \mathbf{y}, k(x,x) - \boldsymbol{k}^\top (K + \sigma^2 I)^{-1} \boldsymbol{k}\right).$$

Compare this to kernel ridge regression. There will be homework exercise on this part.

Acknowledgement

Reference: Machine Learning: A Probabilistic Perspective by Kevin Murphy. The random Fourier feature is due to Rahimi and Recht. A thank you to wikipedia contributors.