Statistical Machine Learning
Homework1

方言
2021210929

2022 年 3 月 22 日

1 Kernel Methods

1.1 Problem 1

1.1.1 Prove that $k(x, y) = (1 + xy)^n$ is a kernel on $\mathcal{X} = \mathbb{R}$

Proof:

As we have:

$$(1 + xy)^n = \sum_{i=0}^{n} \mathsf{C}_n^i x^i y^i$$

where C_n^i denotes the combinatorial number.

We can take $\mathcal{F} = \mathbb{R}^{n+1}$ and $\phi(x)$ as:

$$\phi(x) = (1, \sqrt{\mathsf{C}_n^1}x, ..., \sqrt{\mathsf{C}_n^i}x^i, ..., \sqrt{\mathsf{C}_n^n}x^n)$$

then we have:

$$\langle \phi(x), \phi(y) \rangle = \sum_{i=0}^{n} \mathsf{C}_n^i x^i y^i = (1 + xy)^n$$

which means that $k(x, y) = (1 + xy)^n$ is a kernel on $\mathcal{X} = \mathbb{R}$.

1.1.2 Prove that $k(x, y) = xy - 1$ is not a kernel on $\mathcal{X} = \mathbb{R}$

Proof:

We assume that $k(x, y) = xy - 1$ is a kernel on $\mathcal{X} = \mathbb{R}$, then there exists a $\phi(x)$ such that $k(x, y) = \langle \phi(x), \phi(y) \rangle$.

We take $x = y$ and we have:

$$k(x, x) = x^2 - 1 = \langle \phi(x), \phi(x) \rangle = ||\phi(x)||^2$$

For $x \in (-1, 1)$ we have $k(x, x) < 0$, which contradicts with $||\phi(x)||^2 \geq 0$. Therefore, the original assumption is false and $k(x, y) = xy - 1$ is not a kernel on $\mathcal{X} = \mathbb{R}$.

1.1.3 Prove that $k(x, y) = \min(x, y)$ is a kernel on $\mathcal{X} = [0, 1]$

Proof:

We can define $\phi(x) = 1_{[0,x]}$ and $\mathcal{F} = L^2(\mathbb{R})$, where $1_{[0,x]}$ denotes a function f:

$$f = 1_{[0,x]} = \begin{cases} 1 & t \in [0, x] \\ 0 & other \end{cases}$$

then we have:

$$k(x,y) = \langle \phi(x), \phi(y) \rangle = \langle f, g \rangle = \int_0^1 f(t)g(t)\,dx = \min(x,y)$$

Therefore, $k(x,y) = \min(x,y)$ is a kernel on $\mathcal{X} = [0,1]$.

1.2 Problem 2

Given a training dataset $\{(x_i, y_i)\}_{i=1}^N$, where $x_i \in \mathbb{R}^d$ and $y_i \in R$. Let $\phi : \mathbb{R}^d \to \mathbb{R}^m$ be a feature map. Consider the following regression problem:

$$\min_{w \in \mathbb{R}^m} \quad \frac{\lambda}{2}||w||^2 + \frac{1}{2}\sum_{i=1}^N (w^T\phi(x_i) - y_i)^2 \tag{1}$$

1.2.1 Derive the solution $\hat{w}$ of (1)

We first differentiate the above function to w and set it to 0:

$$\lambda w + \sum_{i=1}^N (w^T\phi(x_i) - y_i)\phi(x_i) = 0$$

We can solve the equation as:

$$\hat{w} = (\lambda \mathsf{I} + \sum_{i=1}^N \phi(x_i)\phi(x_i)^T)^{-1} \cdot ([\phi(x_1), \phi(x_2), ..., \phi(x_N)] \cdot [y_1, y_2, .., y_N]^T)$$

where I is the indentity matrix. We denote $[y_1, y_2, ..., y_N]$ as Y and $[\phi(x_1), \phi(x_2), ..., \phi(x_N)]$ as $\phi(X)$ for simplicity. Here we introduce the Matrix Inverse Lemma:

$$(P^{-1} + B^TR^{-1}B)^{-1}B^TR^{-1} = PB^T(BPB^T + R)^{-1} \tag{2}$$

then we have:

$$\begin{aligned}
\hat{w} &= (\lambda \mathsf{I} + \sum_{i=1}^N \phi(x_i)\phi(x_i)^T)^{-1} \cdot ([\phi(x_1), \phi(x_2), ..., \phi(x_N)] \cdot [y_1, y_2, .., y_N]^T) \\
&= (\phi(X)\phi(X)^T + \lambda \mathsf{I})^{-1}\phi(X)Y^T \\
&= \frac{1}{\lambda}\mathsf{I}\phi(X)(\mathsf{I} + \frac{1}{\lambda}\phi(X)^T\phi(X))^{-1}Y^T \\
&= \phi(X)(\lambda \mathsf{I} + \phi(X)^T\phi(X))^{-1}Y^T
\end{aligned}$$

1.2.2 Express the prediction function $f(x) = \hat{w}^T\phi(x)$ using the kernel $k(x,y) = \phi(x)^T\phi(y)$. The feature map ϕ is not allowed to appear in the result.

Following the solution $\hat{w}$, we have:

$$\begin{aligned}
f(x) &= Y(\lambda \mathsf{I} + \phi(X)^T\phi(X))^{-1}\phi(X)^T\phi(x) \\
&= Y \cdot (\lambda \mathsf{I} + \begin{bmatrix} k(x_1,x_1) & k(x_1,x_2) & \cdots & k(x_1,x_N) \\ k(x_2,x_1) & k(x_2,x_2) & \cdots & k(x_2,x_N) \\ \vdots & \vdots & \ddots & \vdots \\ k(x_N,x_1) & k(x_N,x_2) & \cdots & k(x_N,x_N) \end{bmatrix})^{-1} \cdot \begin{bmatrix} k(x_1,x) \\ k(x_2,x) \\ \vdots \\ k(x_N,x) \end{bmatrix}
\end{aligned}$$

where $Y = [y_1, y_2, ..., y_N]$ and I is the indentity matrix.

1.3 Problem 3

1.3.1 Find the hypothesis space corresponds to the kernel $k(x, y) = (1 + xy)^n$, where $x, y \in \mathbb{R}$

In the previous problem, we have

$$\phi(x) = (1, \sqrt{\mathsf{C}_n^1}x, ..., \sqrt{\mathsf{C}_n^i}x^i, ..., \sqrt{\mathsf{C}_n^n}x^n)$$

For $\forall c \in \mathbb{R}^n$:

$$c^T\phi(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$$

Therefore, the hypothesis space is:

$$\mathcal{H} = \{f : f(x) = p(x), p(x) \in \mathcal{P}_n(x)\}$$

where $\mathcal{P}_n(x)$ is the set of polynomials of degree less than or equal to n.

1.3.2 Show that the function $k_x(y) := k(x, y)$ belongs to $\mathcal{H}$ for every x; and compute $\langle k_x, k_y\rangle_{\mathcal{H}}$

For $\forall x$, $\phi(x) \in \mathbb{R}^m$, so we have:

$$k_x(y) = k(x, y) = \phi(x)^T\phi(y) \in \mathcal{H}$$

In the same way, we can get:

$$k_y(x) = \phi(y)^T\phi(x)$$

then we have:

$$\langle k_x, k_y\rangle_{\mathcal{H}} = \phi(x)^T\phi(y) = k(x, y)$$

1.3.3 For $f \in \mathcal{H}$ and $x \in \mathcal{X}$, show that $\langle f, k_x\rangle_{\mathcal{H}} = f(x)$

Since $\langle f, g\rangle_{\mathcal{H}} = c_f^T c_g$, where $f = c_f^T\phi$, $\quad g = c_g^T\phi$, then we have:

$$\langle f, k_x\rangle_{\mathcal{H}} = c_f^T\phi(x) = f(x)$$

1.3.4 Let $\hat{f}$ be the KRR prediction function obtained in the previous problem, show that

$$\hat{f} = \arg\min_{f\in\mathcal{H}} \frac{\lambda}{2}||f||_{\mathcal{H}}^2 + \frac{1}{2}\sum_{i=1}^{N}(f(x_i) - y_i)^2$$

From the previous problem, we have:

$$\hat{f} = \hat{w}^T\phi(x)$$

where $\hat{w}$ is the solution to KRR:

$$\hat{w} = \arg\min_{w\in\mathbb{R}^m} \frac{\lambda}{2}||w||^2 + \frac{1}{2}\sum_{i=1}^{N}(w^T\phi(x_i) - y_i)^2$$

then we have:

$$||f||_{\mathcal{H}}^2 = \langle f, f\rangle_{\mathcal{H}} = w^T w = ||w||^2$$

For $\forall w \in \mathbb{R}^m$, $f = w^T\phi(x) \in \mathcal{H}$, and for $\forall f \in \mathcal{H}$, the corresponding $w \in \mathbb{R}^m$. Therefore, we notice that:

$$\begin{aligned}
\hat{f} &= \hat{w}^T\phi(x) \\
&= (\arg\min_{w\in\mathbb{R}^m} \frac{\lambda}{2}||w||^2 + \frac{1}{2}\sum_{i=1}^{N}(w^T\phi(x_i) - y_i)^2)^T\phi(x) \\
&= \arg\min_{f\in\mathcal{H}} \frac{\lambda}{2}||f||_{\mathcal{H}}^2 + \frac{1}{2}\sum_{i=1}^{N}(f(x_i) - y_i)^2
\end{aligned}$$

1.3.5 Find the hypothesis space of $k(x, y) = \min(x, y)$ considered in Problem 1

From Problem 1 we have:

$$\phi(x) = 1_{[0,x]}, \quad k(x, y) = \int_0^1 1_{[0,x]} 1_{[0,y]} \, dt = \min(x, y)$$

So, for every $f(t)$, $k_x(y) = k(x, y) \in \mathcal{H}$:

$$k(x, y) = \int_0^1 f(t) 1_{[0,x]} \, dt = \int_0^x f(t) \, dt$$

For every $g(t) \in \mathcal{H}$, we have:

$$g(t) = \int_0^1 \phi(x) f(t) \, dt = \int_0^1 f(t) 1_{[0,x]} \, dt = \int_0^x f(t) \, dt$$

then we can get $f(t) = g'(t)$

Therefore, we have

$$\mathcal{H} = \{g : g = \int_0^x f(t) \, dt, \; f \in \mathcal{L}^1(\mathbb{R})\}$$

1.4 Problem 4

1.4.1 Dataset Generation

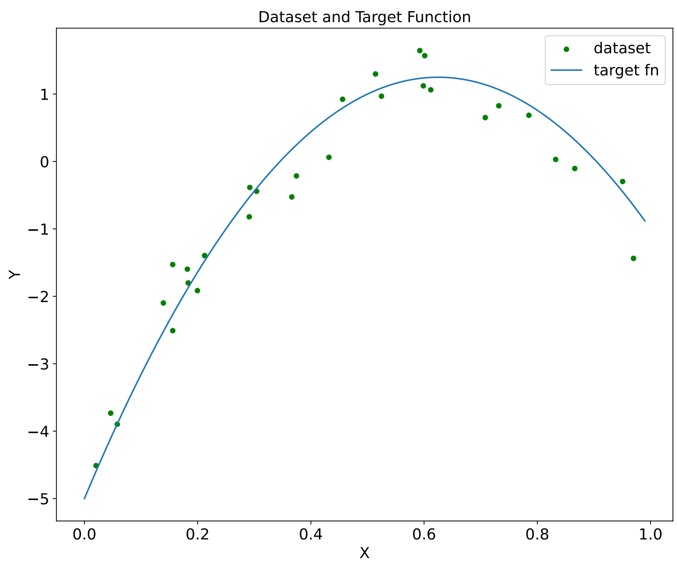

图 1: Data points and target function

The **generate_dataset** method is used for dataset generation. Fig.(1) plots the data points and the target function.

1.4.2 Try to fit the dataset using different λ

We select λ as 0, 0.001, 0.01, 0.1, 1.0, and use kernel function $(1 + xy)^9$ and $\min(x, y)$. Fig.(2) plots the prediction function.

When λ is small, the prediction function tries to fit every data point, which may cause the overfit problem. On the contrary, the prediction can not fit the dataset well with a relatively large λ.

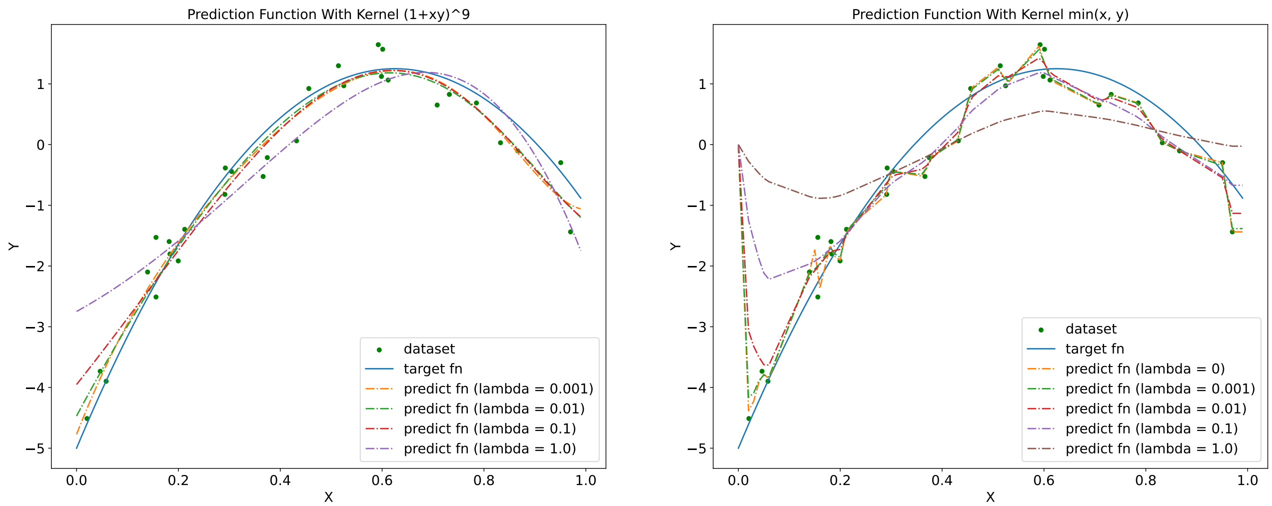

图 2: Prediction function with different λ and kernel functions.

1.4.3 Try to fit the dataset using different kernels.

We fit the dataset with $(1+xy)^n$, $\min(x,y)$ and $\exp(-(x-y)^2)$. Fig.(3) plots the prediction function. Other results are in the attachment (**figures**).

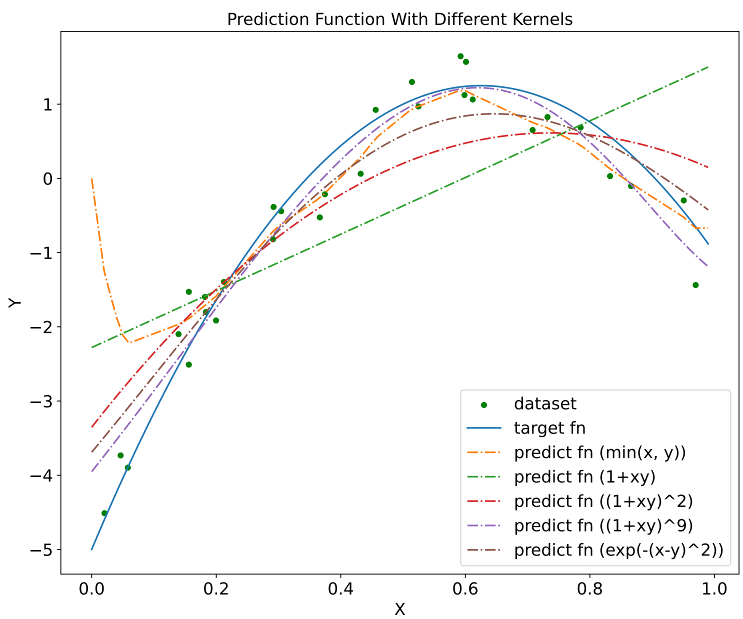

图 3: Prediction function with different kernel functions, λ is set to 0.01.

2 Exponential Families

2.1 Problem 5

2.1.1 Find the maximum likelihood eastimators (MLE) $\hat{\mu}_{\text{ML}}$ and $\hat{\Sigma}_{\text{ML}}$.

Since $x_1, x_2, ...x_N$ are i.i.d. samples, we have:

$$p(x_1, x_2, ...x_N|\mu, \Sigma) = \prod_{i=1}^{N} p(x_i|\mu, \Sigma)$$

We can get the log-likelihood form:

$$\hat{\mu}_{\text{ML}} = \arg\max_{\mu} \ \log p(x_1, x_2, ...x_N|\mu, \Sigma) = \sum_{i=1}^{N} \log p(x_i|\mu, \Sigma)$$

We differentiate it to μ and set it to 0:

$$\frac{\partial}{\partial \mu} \sum_{i=1}^{N} \log p(x_i|\mu, \Sigma) = \sum_{i=1}^{N} \Sigma^{-1}(x_i - \mu) = 0$$

then we get the solution:

$$\hat{\mu}_{\text{ML}} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

Accoring to the defination of covariance, we have:

$$\Sigma = \mathbb{E}[xx^T] - \mu\mu^T$$

Since $\hat{\mu}_{\text{ML}}$ is the MLE of μ, we can get the MLE of Σ:

$$\begin{aligned}\hat{\Sigma}_{\text{ML}} &= \mathbb{E}[xx^T] - \hat{\mu}_{\text{ML}}\hat{\mu}_{\text{ML}}^T \\ &= \frac{1}{N} \sum_{i=1}^{N} x_i x_i^T - \hat{\mu}_{\text{ML}}\hat{\mu}_{\text{ML}}^T\end{aligned}$$

2.1.2 Compute $\mathbb{E}[\hat{\mu}_{\text{ML}}]$ and $\mathbb{E}[\hat{\Sigma}_{\text{ML}}]$, where both expectations are taken with respect to $p(x_1, x_2, ..., x_N|\mu, \Sigma)$. Are these estimators unbiased ?

We have:

$$\mathbb{E}[\hat{\mu}_{\text{ML}}] = \mathbb{E}[\frac{1}{N} \sum_{i=1}^{N} x_i] = \frac{1}{N} \sum_{i=1}^{N} \mu = \mu$$

Therefore, $\hat{\mu}_{\text{ML}}$ is unbiased.

We have:

$$\mathbb{E}[\hat{\Sigma}_{\text{ML}}] = \mathbb{E}[\frac{1}{N} \sum_{i=1}^{N} x_i x_i^T - \hat{\mu}_{\text{ML}}\hat{\mu}_{\text{ML}}^T]$$

For the first part:

$$\mathbb{E}[\frac{1}{N} \sum_{i=1}^{N} x_i x_i^T] = \mathbb{E}[xx^T] = \mu\mu^T + \Sigma$$

For the second part:

$$\begin{aligned}
\mathbb{E}[\hat{\mu}_{\mathrm{ML}}\hat{\mu}_{\mathrm{ML}}^T] &= \frac{1}{N^2}\mathbb{E}[(\sum_{i=1}^{N} x_i)(\sum_{i=1}^{N} x_i)^T] \\
&= \frac{1}{N^2}\mathbb{E}[\sum_{i=1}^{N} x_i x_i^T + \sum_{i\neq j} x_i x_j^T] \\
&= \frac{1}{N^2}[N\mathbb{E}[xx^T] + (N^2 - N)(\mathbb{E}[x])(\mathbb{E}[x])^T] \\
&= \frac{1}{N}(\mu\mu^T + \Sigma) + (1 - \frac{1}{N})\mu\mu^T \\
&= \mu\mu^T + \frac{1}{N}\Sigma
\end{aligned}$$

We then have:

$$\mathbb{E}[\hat{\Sigma}_{\mathrm{ML}}] = (\mu\mu^T + \Sigma) - (\mu\mu^T + \frac{1}{N}\Sigma) = \frac{N-1}{N}\Sigma \neq \Sigma$$

Therefore, $\hat{\Sigma}_{\mathrm{ML}}$ is not unbiased.

2.1.3 Show that

$$\mathbb{E}[||\hat{\mu}_{\mathrm{ML}} - \mu||^2] = \frac{\mathrm{Tr}\Sigma}{N}$$

where $\mathrm{Tr}\Sigma$ is the trace of the matrix Σ.

We have:

$$||\hat{\mu}_{\mathrm{ML}} - \mu||^2 = (\hat{\mu}_{\mathrm{ML}} - \mu)^T(\hat{\mu}_{\mathrm{ML}} - \mu) = \hat{\mu}_{\mathrm{ML}}^T\hat{\mu}_{\mathrm{ML}} + \mu^T\mu - \hat{\mu}_{\mathrm{ML}}^T\mu - \mu^T\hat{\mu}_{\mathrm{ML}}$$

So we get:

$$\begin{aligned}
\mathbb{E}[||\hat{\mu}_{\mathrm{ML}} - \mu||^2] &= \mathbb{E}[\hat{\mu}_{\mathrm{ML}}^T\hat{\mu}_{\mathrm{ML}}] + \mu^T\mu - \mathbb{E}[\hat{\mu}_{\mathrm{ML}}]^T\mu - \mu^T\mathbb{E}[\hat{\mu}_{\mathrm{ML}}] \\
&= \mathbb{E}[\hat{\mu}_{\mathrm{ML}}^T\hat{\mu}_{\mathrm{ML}}] - \mu^T\mu \\
&= \mathbb{E}[\mathrm{Tr}(\hat{\mu}_{\mathrm{ML}}\hat{\mu}_{\mathrm{ML}}^T)] - \mu^T\mu \\
&= \mathrm{Tr}(\mathbb{E}[\hat{\mu}_{\mathrm{ML}}\hat{\mu}_{\mathrm{ML}}^T]) - \mu^T\mu \\
&= \mathrm{Tr}(\mu\mu^T + \frac{1}{N}\Sigma) - \mu^T\mu \\
&= \frac{\mathrm{Tr}\Sigma}{N}
\end{aligned}$$