Homework 4

Xixiang Chen

1 Constructing kernels

(a)

This is not necessarily a kernel. Take $a = -1 < 0$. Then $K(x, x) = -K_1(x, x) \leq 0$ which violates the strict positive definiteness property.

(b)

This is not a kernel. $\langle \mathbf{0}, \mathbf{0} \rangle = 0^3 + (0-1)^2 = 1$ which violates the strict positive definiteness property.

(c)

This is not a kernel, it violates the linearity property:

$$\begin{aligned}
\langle \alpha u + \beta v, w \rangle &= ((\alpha u + \beta v)^T w)^2 + exp(-\left\|\alpha u + \beta v\right\|^2) exp(-\left\|w\right\|^2) \\
&= (\alpha u^T w + \beta v^T w)^2 + exp(-\left\|\alpha u + \beta v\right\|^2 - \left\|w\right\|^2) \\
\alpha\langle u, w \rangle + \beta\langle v, w \rangle &= \alpha((u^T w)^2 + exp(-\left\|u\right\|^2 - \left\|w\right\|^2)) + \beta((v^T w)^2 + exp(-\left\|v\right\|^2 - \left\|w\right\|^2)) \\
&\neq \langle \alpha u + \beta v, w \rangle
\end{aligned}$$

2 Reproducing kernel Hilbert spaces

Define f in the space $\mathscr{F}$ as:

$$f(\cdot) = \sum_{i=1}^{m} \alpha_i k(\cdot, x_i) = \sum_{i=1}^{m} \alpha_i (\cdot) x_i = (\sum_{i=1}^{m} \alpha_i x_i)(\cdot)$$

And since $\alpha_i,\ x_i \in R$ can be anything, for any arbitrary a, let $\sum_{i=1}^{m} \alpha_i x_i = a$, we have $\langle f(\cdot), k(\cdot, x) \rangle_{\mathscr{F}} = ax = f(x)$ so kernel $k(x, y) = xy$ has the reproducing property and k spans $\mathscr{F}$. And since $k(x, y) = xy$ is just a dot product, it is obviously a valid kernel.

3 Convexity and KKT conditions

(a)

The constrains are:

$$\begin{aligned}
y_i - \langle w, x_i \rangle - \epsilon - \eta_i &\leq 0 \\
\langle w, x_i \rangle - y_i - \epsilon - \eta_i^* &\leq 0 \\
-\eta_i &\leq 0, i = 1, ..., n \\
-\eta_i^* &\leq 0, i = 1, ..., n
\end{aligned}$$

The Lagrangian is:

$$\mathcal{L} = \frac{1}{2}\left\|w\right\|^2 + C\sum_{i=1}^{n}(\eta_i + \eta_i^*) + \sum_{i=1}^{n}\alpha_i(y_i - \langle w, x_i \rangle - \epsilon - \eta_i) + \sum_{i=1}^{n}\alpha_i^*(\langle w, x_i \rangle - y_i - \epsilon - \eta_i^*) + \sum_{i=1}^{n}\beta_i(-\eta_i) + \sum_{i=1}^{n}\beta_i^*(-\eta_i^*) \quad (1)$$

where $\alpha_i, \alpha_i^*, \beta_i, \beta_i^* \geq 0, i = 1, ..., n$.
Apply Lagrangian stationary:

$$\nabla_w \mathcal{L} = w - \sum_{i=1}^{n}\alpha_i x_i + \sum_{i=1}^{n}\alpha_i^* x_i = 0 \implies w = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*) x_i \quad (2)$$

$$\frac{\partial}{\partial_{\eta i}}\mathcal{L} = C - \alpha_i - \beta_i = 0 \tag{3}$$

$$\frac{\partial}{\partial_{\eta i}^*}\mathcal{L} = C - \alpha_i^* - \beta_i^* = 0 \tag{4}$$

Substitute back into equation (1):

$$\begin{aligned}
\mathcal{L} &= \frac{1}{2}\|w\|^2 + \sum_{i=1}^{n}((C-\alpha_i-\beta_i)\eta_i + (C-\alpha_i^*-\beta_i^*)\eta_i^*) - \sum_{i=1}^{n}\epsilon(\alpha_i^*+\alpha_i) + \sum_{i=1}^{n}y_i(\alpha_i-\alpha_i^*) + \sum_{i=1}^{n}(\alpha_i^*-\alpha_i)\langle w, x_i\rangle \\
&= \frac{1}{2}\left\|\sum_{i=1}^{n}(\alpha_i-\alpha_i^*)x_i\right\|^2 - \sum_{i=1}^{n}\epsilon(\alpha_i^*+\alpha_i) + \sum_{i=1}^{n}y_i(\alpha_i-\alpha_i^*) + \sum_{i=1}^{n}(\alpha_i^*-\alpha_i)\langle\sum_{j=1}^{n}(\alpha_j-\alpha_j^*)x_j, x_i\rangle \\
&= -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)\langle x_j, x_i\rangle - \sum_{i=1}^{n}\epsilon(\alpha_i+\alpha_i^*) + \sum_{i=1}^{n}y_i(\alpha_i-\alpha_i^*)
\end{aligned}$$

Therefore the dual form is:

$$\max_{\alpha_i,\alpha_i^*} -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)\langle x_j, x_i\rangle - \sum_{i=1}^{n}\epsilon(\alpha_i+\alpha_i^*) + \sum_{i=1}^{n}y_i(\alpha_i-\alpha_i^*)$$

subject to:

$$0 \le \alpha_i, \alpha_i^* \le C,\ i = 1, \dots, n$$

(b)

From complementary slackness, we have:

$$\begin{aligned}
\alpha_i(y_i - \langle w, x_i\rangle - \epsilon - \eta_i) &= 0 \\
\alpha_i^*(\langle w, x_i\rangle - y_i - \epsilon - \eta_i^*) &= 0 \\
\beta_i\eta_i &= 0 \\
\beta_i^*\eta_i^* &= 0
\end{aligned}$$

When $\alpha_i > 0$, $y_i - \langle w, x_i\rangle - \epsilon - \eta_i = 0 \implies y_i - \langle w, x_i\rangle - \epsilon = \eta_i$. When $\eta_i = 0$, the loss is 0, so x_i is a support vector. Similarly when $\alpha_i^* > 0$ and $\eta_i^* = 0$, x_i is a support vector.

(c)

Increasing ϵ makes the model less likely to overfit in general in a sense that it is "insensitive" to larger errors.

(d)

Increasing C makes the model more likely to overfit since it gives a larger penalty for errors.

(e)

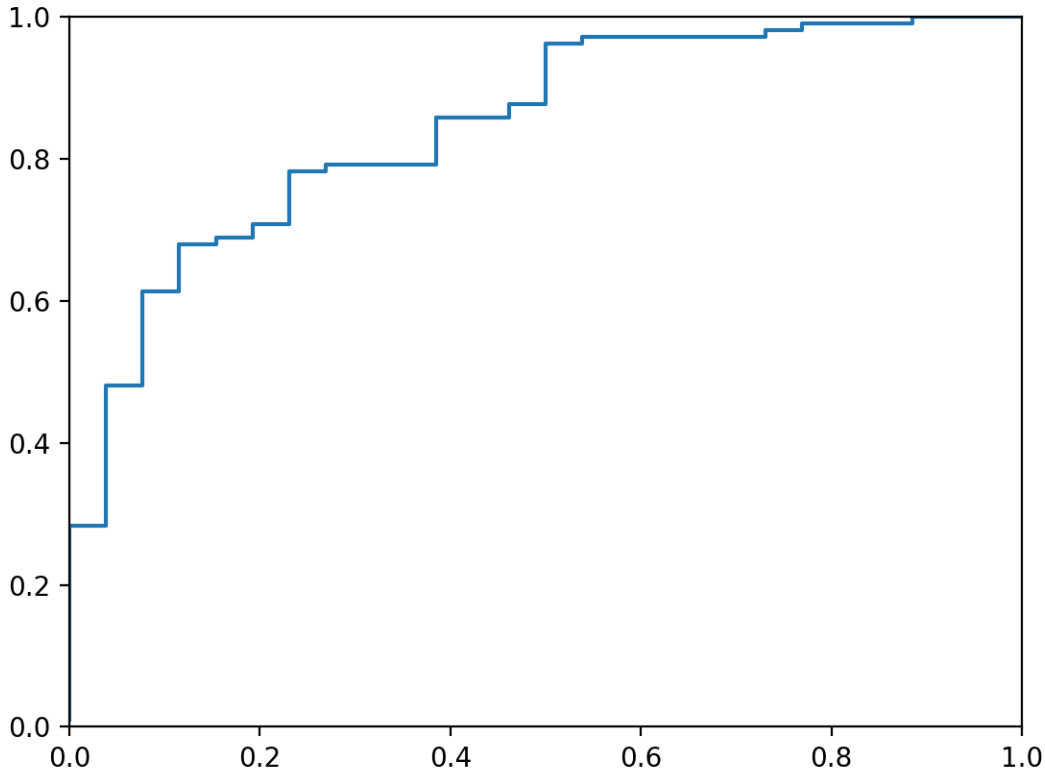
$$f(x) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)\langle x, x_i\rangle$$

4 SVM implementation

(a)

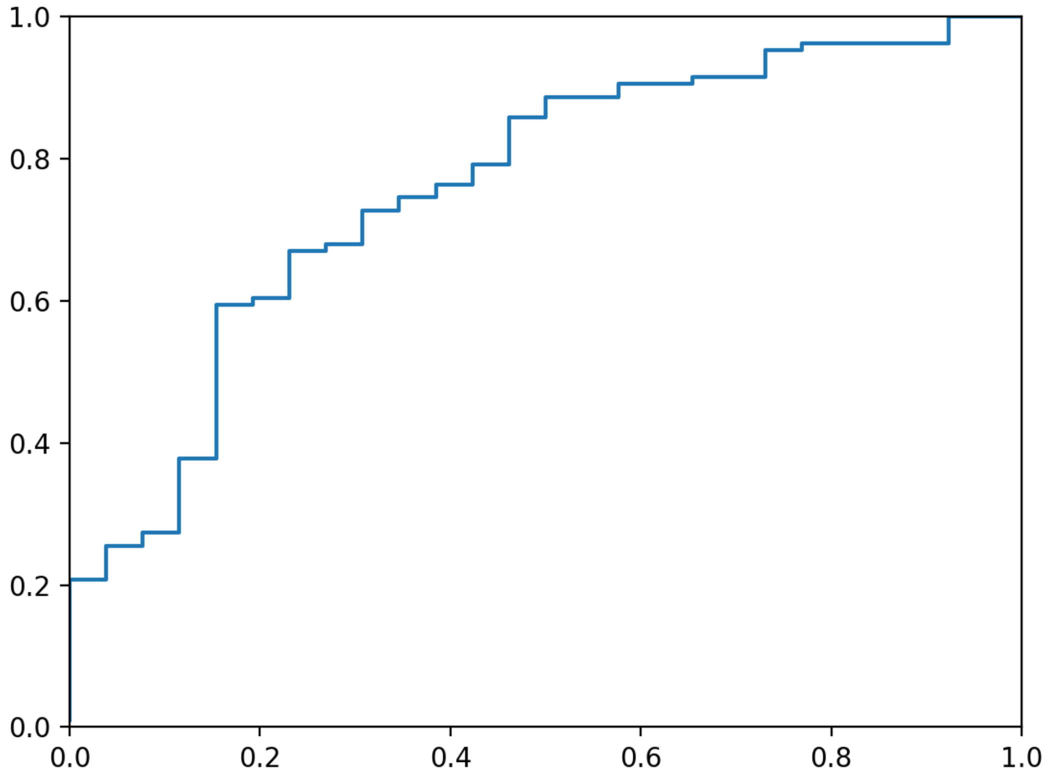
(b)

Accuracy = 0.8409090909
AUC = 0.8454281567



(c)

For $\sigma^2 = 5$:
Accuracy = 0.8030303030
AUC = 0.7601596517



For $\sigma^2 = 25$:
Accuracy = 0.8409090909
AUC = 0.8465166909