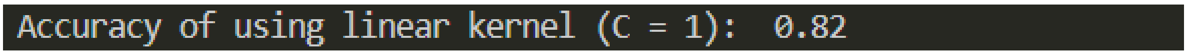
NYCU Introduction to Machine Learning, Homework 4

110612117 張仲瑜

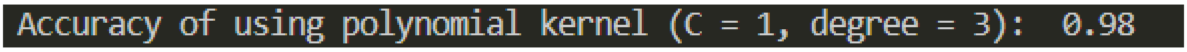
Part. 1, Coding (50%):

1. (10%) Show the accuracy score of the testing data using linear_kernel. Your accuracy score should be higher than 0.8.



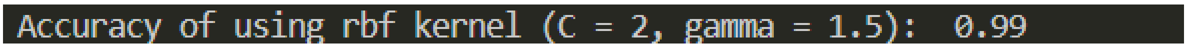
```
Accuracy of using linear kernel (C = 1):  0.82
```

2. ((20%) Tune the hyperparameters of the polynomial_kernel. Show the accuracy score of thetesting data using.



```
Accuracy of using polynomial kernel (C = 1, degree = 3):  0.98
```

3. (20%) Tune the hyperparameters of the rbf_kernel. Show the accuracy score of the testing data using rbf_kernel and the hyperparameters you used.



```
Accuracy of using rbf kernel (C = 2, gamma = 1.5):  0.99
```

Part. 2, Questions (50%):

1. (20%) Given a valid kernel $\boldsymbol{k_1(x,x')}$, prove that the following proposed functions are or are not valid kernels. If one is not a valid kernel, give an example of $\boldsymbol{k(x,x')}$ that the corresponding $\boldsymbol{K}$ is not positive semidefinite and shows its eigenvalues.

   a. $\boldsymbol{k(x,x') = k_1(x,x') + exp(x^T x')}$

   b. $\boldsymbol{k(x,x') = k_1(x,x') - 1}$

   c. $\boldsymbol{k(x,x') = exp(\|x - x'\|^2)}$

   d. $\boldsymbol{k(x,x') = exp\left(k_1(x,x')\right) - k_1(x,x')}$

a. $k(x, x') = k_1(x, x') + \exp(x^T x')$

Prove I is a symmetric positive semidefinite matrix first. (SPSM)

$u^T I u = u^T u = \|u\|^2 \geq 0 \quad \forall\, u \neq \vec{0}$

∴ I is a SPSM.

$x^T A x'$ is a valid kernel (6.20)

$\Rightarrow x^T x'$ is a valid kernel (when $A = I$)

$\Rightarrow \exp(x^T x')$ is a valid kernel (6.16)

$\Rightarrow k(x, x') = k_1(x, x') + \exp(x^T x')$ is a valid kernel (6.17) ※

↳ valid kernel given by the question

b. $k(x, x') = k_1(x, x') - 1$

let $x_1 = [1, 0]^T$, $x_2 = [0, 1]^T$

$\begin{bmatrix} x_1^T x_1 & x_1^T x_2 \\ x_2^T x_1 & x_2^T x_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = I$ is SPSM $\Rightarrow k_1(x_1, x_2)$ is a valid kernel

$K = \begin{bmatrix} 1-1 & 0-1 \\ 0-1 & 1-1 \end{bmatrix} = \begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix}$

SPSM ⇔ all of the eigenvalue should ≥ 0

$\begin{vmatrix} 0-\lambda & -1 \\ -1 & 0-\lambda \end{vmatrix} = \lambda^2 - 1 = 0 \Rightarrow \lambda = \pm 1$

$\Rightarrow$ negative λ exist

$\Rightarrow$ x SPSM

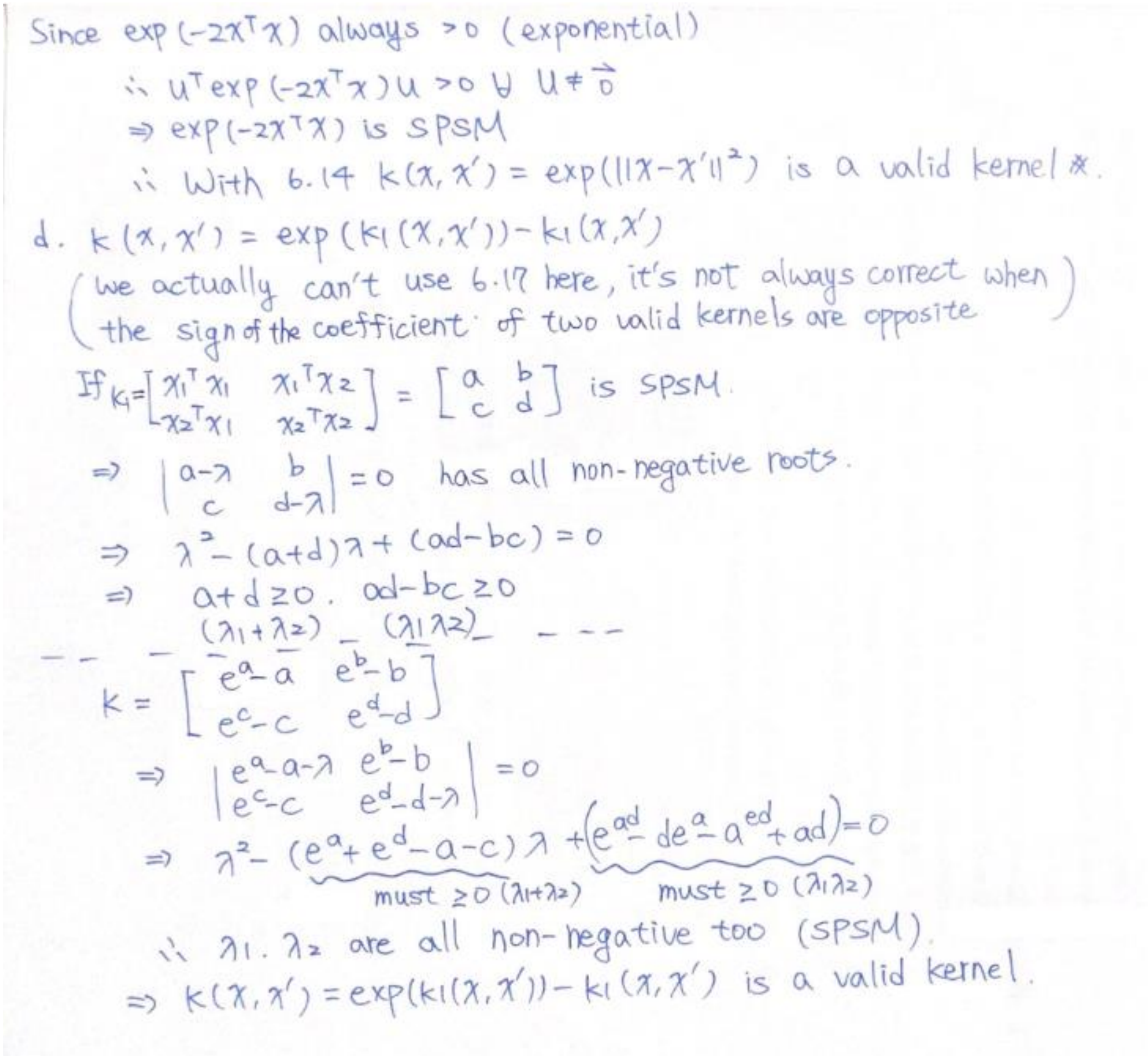
∴ $k(x, x') = k_1(x, x') - 1$ isn't a valid kernel ※

c. $k(x, x') = \exp(\|x - x'\|^2)$

$\|x - x'\|^2 = x^T x - 2x^T x + (x')^T x'$

$\Rightarrow \exp(\|x - x'\|^2) = \underbrace{\exp(x^T x)}_{f(x)} \cdot \exp(-2x^T x) \cdot \underbrace{\exp(x'^T x')}_{f(x')}$

We can use 6.14 as long as $\exp(-2x^T x)$ is SPSM($\Rightarrow$ valid kernel)



Since $\exp(-2x^T x)$ always >0 (exponential)

$\therefore u^T \exp(-2x^T x) u > 0 \ \forall \ u \neq \vec{0}$

$\Rightarrow \exp(-2x^T x)$ is SPSM

$\therefore$ With 6.14 $k(x, x') = \exp(\|x-x'\|^2)$ is a valid kernel ※.

d. $k(x, x') = \exp(k_1(x, x')) - k_1(x, x')$

(we actually can't use 6.17 here, it's not always correct when the sign of the coefficient of two valid kernels are opposite)

If $K_1 = \begin{bmatrix} x_1^T x_1 & x_1^T x_2 \\ x_2^T x_1 & x_2^T x_2 \end{bmatrix} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ is SPSM.

$\Rightarrow \begin{vmatrix} a-\lambda & b \\ c & d-\lambda \end{vmatrix} = 0$ has all non-negative roots.

$\Rightarrow \lambda^2 - (a+d)\lambda + (ad-bc) = 0$

$\Rightarrow \underset{(\lambda_1+\lambda_2)}{a+d \geq 0}.\ \underset{(\lambda_1\lambda_2)}{ad-bc \geq 0}$

$K = \begin{bmatrix} e^a - a & e^b - b \\ e^c - c & e^d - d \end{bmatrix}$

$\Rightarrow \begin{vmatrix} e^a - a - \lambda & e^b - b \\ e^c - c & e^d - d - \lambda \end{vmatrix} = 0$

$\Rightarrow \lambda^2 - (\underbrace{e^a + e^d - a - c}_{\text{must} \geq 0\ (\lambda_1+\lambda_2)})\lambda + (\underbrace{e^{ad} - de^a - a^{ed} + ad}_{\text{must} \geq 0\ (\lambda_1\lambda_2)}) = 0$

$\therefore \lambda_1 . \lambda_2$ are all non-negative too (SPSM).

$\Rightarrow k(x, x') = \exp(k_1(x, x')) - k_1(x, x')$ is a valid kernel.

2. ((15%) One way to construct kernels is to build them from simpler ones. Given three possible "construction rules" : assuming $\boldsymbol{K_1}(\boldsymbol{x}, \boldsymbol{x}')$ and $\boldsymbol{K_2}(\boldsymbol{x}, \boldsymbol{x}')$are kernels then so are

   a. (scaling) $\boldsymbol{f}(\boldsymbol{x})\boldsymbol{K_1}(\boldsymbol{x}, \boldsymbol{x}')\boldsymbol{f}(\boldsymbol{x}'), \boldsymbol{f}(\boldsymbol{x}) \in \boldsymbol{R}$

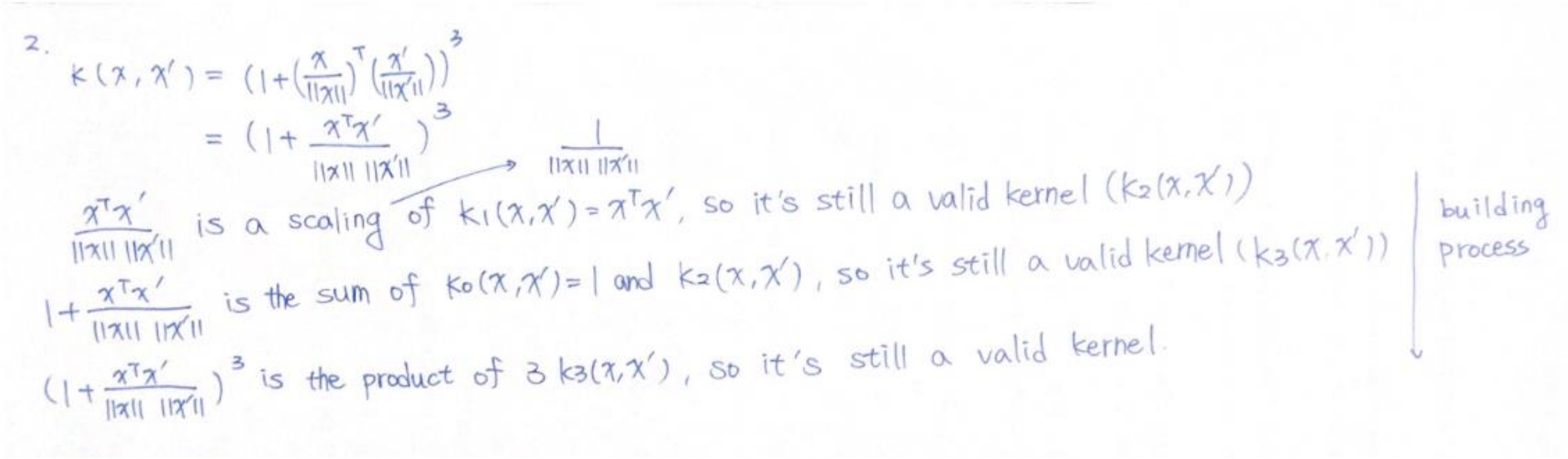
   b. (sum) $\boldsymbol{K_1}(\boldsymbol{x}, \boldsymbol{x}') + \boldsymbol{K_2}(\boldsymbol{x}, \boldsymbol{x}')$

   c. (product) $\boldsymbol{K_1}(\boldsymbol{x}, \boldsymbol{x}')\boldsymbol{K_2}(\boldsymbol{x}, \boldsymbol{x}')$

   Use the construction rules to build a normalized cubic polynomial kernel:

$$\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x}') = \left(\boldsymbol{1} + \left(\frac{\boldsymbol{x}}{||\boldsymbol{x}||}\right)^{\boldsymbol{T}} \left(\frac{\boldsymbol{x}'}{||\boldsymbol{x}'||}\right)\right)^{\boldsymbol{3}}$$

You can assume that you already have a constant kernel $K_0(x, x') = 1$ and a linear kernel $K_1(x, x') = x^T x'$. Identify which rules you are employing at each step.



2.

$$k(x, x') = \left(1 + \left(\frac{x}{\|x\|}\right)^T \left(\frac{x'}{\|x'\|}\right)\right)^3$$

$$= \left(1 + \frac{x^T x'}{\|x\|\,\|x'\|}\right)^3$$

$\frac{1}{\|x\|\,\|x'\|}$

$\frac{x^T x'}{\|x\|\,\|x'\|}$ is a scaling of $k_1(x, x') = x^T x'$, so it's still a valid kernel $(k_2(x, x'))$

$1 + \frac{x^T x'}{\|x\|\,\|x'\|}$ is the sum of $k_0(x, x') = 1$ and $k_2(x, x')$, so it's still a valid kernel $(k_3(x, x'))$

$\left(1 + \frac{x^T x'}{\|x\|\,\|x'\|}\right)^3$ is the product of 3 $k_3(x, x')$, so it's still a valid kernel.

building process ↓

3. (15%) A social media platform has posts with text and images spanning multiple topics like news, entertainment, tech, etc. They want to categorize posts into these topics using SVMs. Discuss two multi-class SVM formulations: `One-versus-one` and `One-versus-the-rest` for this task.
   a. The formulation of the method [how many classifiers are required]
   b. Key trade offs involved (such as complexity and robustness).
   c. If the platform has limited computing resources for the application in the inference phase and requires a faster method for the service, which method is better.

Ans

a. One-versus-one

(k-1) classifiers are needed for every topic (except for itself),

so k(k-1) / 2 classifiers are needed in total.

The topic with the most votes is assigned as the final prediction.

One-versus-the-rest

we need k classifiers in total. For each topic, we train a binary SVM classifier where samples from the target topic are labeled as positive, and samples from all other classes are labeled as negative.

The positive topic of the classifier that has the highest positive rate is the predicted topic.

b. One-versus-one

Advantage:

i. The influence of class imbalance problem will be mitigated compared to one-versus-the-rest, which means higher robustness.
ii. Though the number of classifiers is larger, fewer data need to be considered at a time during the training of a single classifier.
iii. Can handle non-linear decision problem better than the other strategy.

Disadvantage:

It costs more time and memory to train the model since more classifier compared to one-versus-the-rest.

One-versus-the-rest

Advantage:

i. It takes less time and memory to train since the number of the classifiers are less.
ii. The model can be use in various situations that need to output the probability of one specific topic.

Disadvantage:

i. It cost the unbalanced problem since that the number of one specific topic is usually less than the sum of the others, so the positive tags will be enormously less than the negative tags. This will cause the bias of the decision boundary.
ii. Can be problematic if classes are not well-separated.

c. If the time and the computing resources issue exists, the One-versus-the-rest strategy is suitable due to its lower time and memory cost compared to the other strategy.