NYCU Introduction to Machine Learning, Homework 4

110550088 李杰穎

1 Part. 1, Coding

1.1 Support Vector Machine

1. Show the accuracy score of the testing data using `linear_kernel`. Your accuracy score should be higher than 0.8.

   As in Figure 1. I set $C=4$ and achieve the accuracy of 0.83.

2. Tune the hyperparameters of the `polynomial_kernel`. Show the accuracy score of the testing data using `polynomial_kernel` and the hyperparameters you used.

   As in Figure 1. I set $C=1$ and degree $=3$. In addition, the positive constant c is set to 1. Under this setting, the accuracy is 0.98.

3. Tune the hyperparameters of the `rbf_kernel`. Show the accuracy score of the testing data using `rbf_kernel` and the hyperparameters you used.

   As in Figure 1. I set $C=1$ and $\gamma=0.5$. The accuracy is 0.98.

```
Accuracy of using linear kernel (C = 4):  0.83
Accuracy of using polynomial kernel (C = 1, degree = 3):  0.98
Accuracy of using rbf kernel (C = 1, gamma = 0.5):  0.98
```

Figure 1: The accuracy of using different kernel functions. The hyperparameters are shown in the screenshot.

2 Part. 2, Questions

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

   a. $k(\mathbf{x},\mathbf{x}')=k_1(\mathbf{x},\mathbf{x}')+\exp\left(\mathbf{x}^\top\mathbf{x}'\right)$

   We first prove that $k_2(\mathbf{x},\mathbf{x}')=\mathbf{x}^\top\mathbf{x}'$ is a valid kernel. Suppose the basis function $\phi(\mathbf{x})=\mathbf{x}$, it's clear that the corresponding kernel function is $k_2(\mathbf{x},\mathbf{x}')$. Therefore, k_2 is a valid kernel.

   Second, by utilizing (6.16), $\exp\left(\mathbf{x}^\top\mathbf{x}'\right)$ is also a valid kernel.

   Lastly, by utilizing (6.17), we can finally prove that $k(\mathbf{x},\mathbf{x}')=k_1(\mathbf{x},\mathbf{x}')+\exp\left(\mathbf{x}^\top\mathbf{x}'\right)$ is a valid kernel.

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

   Suppose $\mathbf{X}=\{x_1,x_2\}$, where $x_1=(1,1)^\top, x_2=(3,4)^\top$ and k_1 is the linear kernel. The gram matrix $\mathbf{K}=\begin{bmatrix}1 & 6\\ 6 & 24\end{bmatrix}$. To calculate the eigenvalue of $\mathbf{K}$, we can solve $\det(\mathbf{K}-\lambda\mathbf{I})=0$, the solutions are $\lambda=\frac{25\pm\sqrt{673}}{2}$, one of the eigenvalues is less than zero. Therefore, $\mathbf{K}$ is not positive semi-definite, leading that $k(\mathbf{x},\mathbf{x}')=k_1(\mathbf{x},\mathbf{x}')+\exp\left(\mathbf{x}^\top\mathbf{x}'\right)$ is not a valid kernel.

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

   Suppose, $\mathbf{X}=\{x_1,x_2\}$, where $x_1=(1,1)^\top, x_2=(3,4)^\top$ and k_1 is the linear kernel. The gram matrix $\mathbf{K}=\begin{bmatrix}1 & \exp(13)\\ \exp(13) & 1\end{bmatrix}$. To calculate the eigenvalue of $\mathbf{K}$, we can solve $\det(\mathbf{K}-\lambda\mathbf{I})=0$, the solutions are $\lambda=1\pm\exp(13)$, one of the eigenvalues is less than zero. Therefore, $\mathbf{K}$ is not positive semi-definite, leading that $k(\mathbf{x},\mathbf{x}')=\exp\left(\|\mathbf{x}-\mathbf{x}'\|^2\right)$ is not a valid kernel.

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

   First, we can expand $\exp(k_1(\mathbf{x},\mathbf{x}'))$ to $1+k_1(\mathbf{x},\mathbf{x}')+\frac{k_1(\mathbf{x},\mathbf{x}')^2}{2!}+\frac{k_1(\mathbf{x},\mathbf{x}')^n}{n!}+\cdots$ using Taylor's expansions. Therefore, $k(\mathbf{x},\mathbf{x}')=\exp(k_1(\mathbf{x},\mathbf{x}'))-k_1(\mathbf{x},\mathbf{x}')=1+\frac{k_1(\mathbf{x},\mathbf{x}')^2}{2!}+\frac{k_1(\mathbf{x},\mathbf{x}')^n}{n!}+\cdots$. It's trivial that each term in the RHS is a valid kernel by (6.18) and (6.16). For the constant 1, the eigenvalues of its corresponding gram matrix are 0 and 1. This make it also a valid

kernel. Finally, by applying (6.17), we prove that $k(\mathbf{x},\mathbf{x}') = \exp(k_1(\mathbf{x},\mathbf{x}')) - k_1(\mathbf{x},\mathbf{x}')$ is a valid kernel.

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

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

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

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

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

$$K(\mathbf{x},\mathbf{x}') = \left(1 + \left(\frac{\mathbf{x}}{\|\mathbf{x}\|}\right)^\top \left(\frac{\mathbf{x}'}{\|\mathbf{x}'\|}\right)\right)^3$$

   You can assume that you already have a constant kernel $K_0(\mathbf{x},\mathbf{x}') = 1$ and a linear kernel $K_1(\mathbf{x},\mathbf{x}') = \mathbf{x}^\top \mathbf{x}'$. Identify which rules you are employing at each step.

   I solve the question by following steps:

   (a) Scaling the linear kernel, with $f(\mathbf{x}) = \frac{1}{\|\mathbf{x}\|}$

$$\mathbf{x}^\top \mathbf{x}' \Rightarrow \frac{1}{\|\mathbf{x}\|}\mathbf{x}^\top \mathbf{x}' \frac{1}{\|\mathbf{x}'\|} = \left(\frac{\mathbf{x}}{\|\mathbf{x}\|}\right)^\top \left(\frac{\mathbf{x}'}{\|\mathbf{x}'\|}\right)$$

   (b) Apply the sum construction rule, with K_1 is a constant kernel and K_2 the result from last step.

$$\left(\frac{\mathbf{x}}{\|\mathbf{x}\|}\right)^\top \left(\frac{\mathbf{x}'}{\|\mathbf{x}'\|}\right) \Rightarrow 1 + \left(\frac{\mathbf{x}}{\|\mathbf{x}\|}\right)^\top \left(\frac{\mathbf{x}'}{\|\mathbf{x}'\|}\right)$$

   (c) Apply the product construction rule for two times, we will get the final kernel.

$$1 + \left(\frac{\mathbf{x}}{\|\mathbf{x}\|}\right)^\top \left(\frac{\mathbf{x}'}{\|\mathbf{x}'\|}\right) \Rightarrow \left(1 + \left(\frac{\mathbf{x}}{\|\mathbf{x}\|}\right)^\top \left(\frac{\mathbf{x}'}{\|\mathbf{x}'\|}\right)\right)^3$$

3. 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]

      For 'One-versus-one', an SVM is trained for every pair of classes. Therefore, to classify N classes, we need $\frac{N(N-1)}{2}$ SVMs. To classify an input, we often use majority voting to decide the output category.

      As for 'One-versus-the-rest', one SVM is trained per class, distinguishing that class from all other classes. Therefore, to classify N classes, we only need $N-1$ classifiers to classify an input.

   b. Key trade offs involved (such as complexity and robustness).

      The complexity of 'One-versus-one' is much higher when N is large, compared with 'One-versus-the-rest'. However, the robustness for 'One-versus-one' is higher than 'One-versus-the-rest' because each classifier is trained on only two classes at a time, which might make them more specialized and potentially more accurate in distinguishing between classes.

   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.

      As I mentioned earlier, we need lots of SVMs using 'One-versus-one' formulation when N is large. Therefore, I suggest that 'One-versus-the-rest' is a better method under the circumstance.