FMAN45 - Assignment 2

Jingmo Bai

May 1, 2023

1 Task T1

Given the non-linear feature map

$$\phi(x) = \begin{pmatrix} x \\ x^2 \end{pmatrix}$$

and the kernel

$$k(x, y) = \phi(x)^T \phi(y)$$

The kernel matrix can be computed as

$$K = k(x_i, x_j) = \phi(x_i)^T \phi(x_j) = x_i x_j + (x_i x_j)^2 = \begin{pmatrix} 20 & 6 & 2 & 12 \\ 6 & 2 & 0 & 2 \\ 2 & 0 & 2 & 6 \\ 12 & 2 & 6 & 20 \end{pmatrix}$$

2 Task T2

Given that the solution satisfies $\alpha = \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4$ we can rewrite the maximization problem as

$$\max_{\alpha} \left(4\alpha - \frac{\alpha^2}{2} \sum_{i,j=1}^{4} y_i y_j k(x_i, x_j) \right) \tag{1}$$

$$\textit{subject to } \alpha \geq 0 \textit{ and } \alpha \sum_{i=1}^{4} y_i = 0$$

By inserting the values from the kernel we get

$$\sum_{i,j=1}^{4} y_i y_j k(x_i, x_j) = (2 \cdot 20 + 2 \cdot 12 + 2 \cdot 2 + 2 \cdot 0) - (4 \cdot 6 + 4 \cdot 2) = 36$$

Inserted to (1) we can get

$$\max_{\alpha}(4\alpha - 18\alpha^2)$$

To maximize this function, we just need to investigate the first and second derivative

$$\frac{d}{d\alpha}(4\alpha - 18\alpha^2) = 4 - 36\alpha$$

$$\frac{d^2}{d\alpha^2}(4\alpha - 18\alpha^2) = -36 < 0$$

Therefore the function is concave and we can find the maximum point when the first derivative equals 0

$$\alpha = \frac{1}{9}$$

3 Task T3

To reduce the classifier function to the simplest form

$$g(x) = \sum_{j=1}^{4} \alpha y_i k(x_j, x) + b = \frac{1}{9} \sum_{j=1}^{4} y_i k(x_j, x) + b$$

$$g(x) = \frac{1}{9}\big((-2x + 4x^2) - (-x + x^2) - (x + x^2) + (2x + 4x^2)\big) + b$$

$$= \frac{2}{3}x^2 + b$$

Given that

$$y_s\bigg(\sum_{j=1}^{4} \alpha_j y_j k(x_j, x_s) + b\bigg) = 1$$

$$= y_s\bigg(\frac{2}{3}x^2 + b\bigg) = 1\bigg(\frac{2}{3}(-2)^2 + b\bigg) = \frac{8}{3} + b \iff$$

$$b = -\frac{5}{3}$$

Therefore we obtain the simplest form.

$$g(x) = \frac{2}{3}x^2 - \frac{5}{3}$$

4 Task T4

We notice that x_2, x_3, x_5, x_6 are the same as the previous dataset. Thus we try to investigate if the three new data also suit the previous classifier.

$$g(-3) = \frac{13}{3} \geq 1$$

$$g(0) = -\frac{5}{3} \leq -1$$

$$g(4) = \frac{27}{3} \geq 1$$

Thus all the data points are located outside the margin we obtained before, so the solution $g(x)$ of the nonlinear kernel SVM is the same as the previous one.

$$g(x) = \frac{2}{3}x^2 - \frac{5}{3}$$

5 Task T5

Given the primal formulation of the linear soft margin classifier

$$\min_{w,b,\xi} \frac{1}{2}||w||^2 + C\sum_{i=1}^{n}\xi_i$$

$$subject\ to\ \ y_i(w^T x_i + b) \geq 1 - \xi_i\ \ \forall i$$

$$\xi_i \geq 0\ \ \forall i$$

The minimization problem can be derived from the corresponding Lagrangian function.

$$L(w,b,\xi) = \frac{1}{2}||w||^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}a_i(\xi - 1 + u_i(w^T x_i + b)) - \sum_{i=1}^{n}\lambda_i\xi_i \quad (2)$$

$$subject\ to\ \ a_i, \lambda_i \geq 0$$

Now we need to minimize L with respect to w, b, ξ, by setting the differentiation to zero.

$$\frac{dL}{dw} = 0 \Longleftrightarrow w - \sum_{i=1}^{n}a_i y_i x_i = 0 \longleftrightarrow w = \sum_{i=1}^{n}a_i y_i x_i$$

$$\frac{dL}{db} = 0 \Longleftrightarrow \sum_{i=1}^{n}a_i y_i = 0$$

$$\frac{dL}{d\xi} = 0 \Longleftrightarrow \lambda_i = C - a_i$$

Inserted to (2) we get

$$\max_{a_1,\ldots,a_n} L = \frac{1}{2}||\sum_{i=1}^{n}a_i y_i x_i||^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}a_i(\xi - 1 + u_i((\sum_{i=1}^{n}a_i y_i x_i)^T x_i + b)) - \sum_{i=1}^{n}\lambda_i\xi_i$$

$$= -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}a_i a_j y_i y_j x_i^T x_j + C\sum_{i=1}^{n}\xi_i(C - a_i - \lambda_i) + \sum_{i=1}^{n}a_i - \sum_{i=1}^{n}a_i y_i b$$

$$= \sum_{i=1}^{n}a_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}a_i a_j y_j y_j x_i^T x_j \quad (3)$$

Given that $a_i, \lambda_i \geq 0$ and $\lambda_i = C - a_i$ we can get the constraint

$$0 \leq a_i \leq C$$

$$\sum_{i=1}^{n}a_i y_i = 0$$

We have shown that the Lagrangian dual problem is given by (3)

6 Task T6

We want to show that support vectors with $y_i(w^T x_i + b) < 1$ have coefficient $a_i = C$ using complementary slackness of the KKT conditions.

$$a_i \Big(y_i(w^t x_i + b) - 1 + \xi_i \Big) = 0$$

The complementary slackness states that

$$\beta_i \xi_i = 0$$

We can get

$$\xi_i = \max(0, 1 - y_i(w^T x_i + b))$$
$$y_i(w^T x_i + b) < 1$$
$$\xi_i > 0$$

Thus we can get $\beta_i = 0$. Given that $\lambda_i = C - a_i$, we can show $a_i = C$

7 Task E1

In this task, to ensure the necessary condition to apply PCA is met, we need to normalize the data to zero-mean data and then apply the singular value decomposition(SVD). The plot is displayed below.

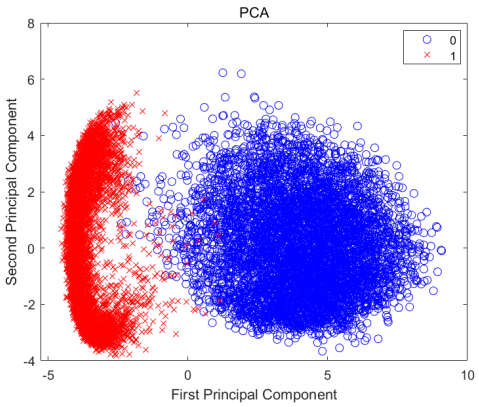

Figure 1: Visualization of the training data through PCA

8 Task E2

The figures below show the visualization of K-means clustering when K=2 and K=5 clusters.
The reason clusters seem to overlap for K=5 is that we only have two digits "0" and "1" in the data, so when we try to classify them into 5 classes, the result is that some classes are the same digit but just look a bit different.

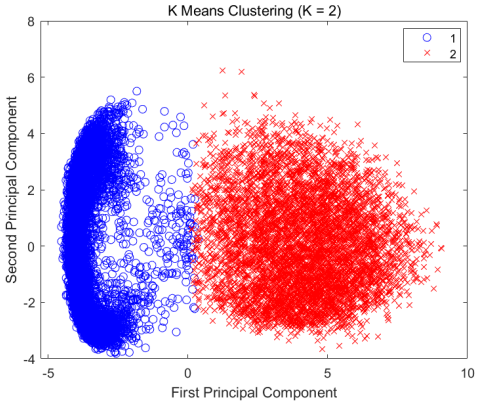

Figure 2: Visualization of K-means clustering(K=2)

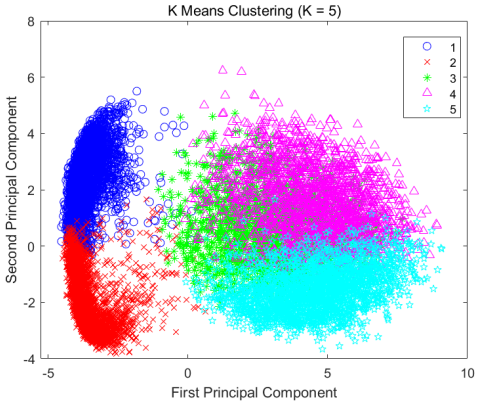

Figure 3: Visualization of K-means clustering(K=5)

9 Task E3

The stacked centroids are displayed below for K=2 and K= 5.

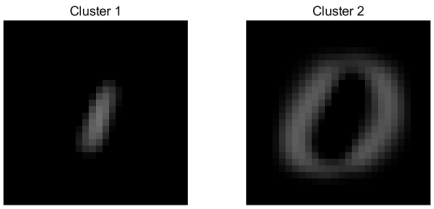

Figure 4: Centroids from each cluster(K=2)

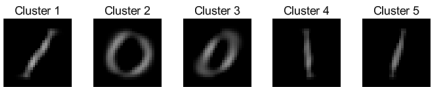

Figure 5: Centroids from each cluster(K=5)

10 Task E4

The table below shows how many misclassifications occur for the train and test set and the misclassification rate for K=2.

Table 1: K-means classification results

Training data	Cluster	# '0'	# '1'	Assigned to class	# misclassified
	1	112	6736	1	112
	⋮	⋮	⋮	⋮	⋮
	K	5811	6	0	6
$N_{\text{train}} = 12665$				Sum misclassified:	118
				Misclassification rate (%):	0.93
Testing data	Cluster	# '0'	# '1'	Assigned to class	# misclassified
	1	11	1135	1	11
	⋮	⋮	⋮	⋮	⋮
	K	969	0	0	0
$N_{\text{test}} = 2115$				Sum misclassified:	11
				Misclassification rate (%):	0.52

11 Task E5

In this task, we try to find out if we can lower the misclassification rate on test data by considering some different K values. After trying out values from 2-10, I find that the misclassification rate keeps going down when the number of clusters K increases. When k=10, the misclassification rate can reach 0.0014 for test data.

12 Task E6

The table below shows the classification results of the linear SVM.

Table 2: Linear SVM classification results

Training data	Predicted class	True class: # '0'	# '1'
	'0'	5923	0
	'1'	0	6742
$N_{\text{train}} = 12665$		Sum misclassified:	0
		Misclassification rate (%):	0
Testing data	Predicted class	True class: # '0'	# '1'
	'0'	979	1
	'1'	1	1134
$N_{\text{test}} = 2115$		Sum misclassified:	2
		Misclassification rate (%):	0.0946

13 Task E7

After trying out a few values of β, we find the optimal value is around 5-6. It gives 0 misclassifications for test data. Either larger or smaller values will give more than 0 misclassifications.
The table below shows the classification results for SVM with Gaussian kernel using the optimal $\beta = 6$.

Table 3: Gaussian kernel SVM classification results with optimal β

Training data	Predicted class	True class: # '0'	# '1'
	'0'	5923	0
	'1'	0	6742
$N_{\text{train}} = 12665$		Sum misclassified:	0
		Misclassification rate (%):	0
Testing data	Predicted class	True class: # '0'	# '1'
	'0'	980	0
	'1'	0	1135
$N_{\text{test}} = 2115$		Sum misclassified:	0
		Misclassification rate (%):	0

14 Task E8

No, we can't expect the same error on new images. Because during the process to tune the parameters to achieve the best performance on the test data, the trained models are already overfitted to the training and test data. So this model probably can not generalize well. We need to split the data into train validation and test set. When tuning the parameters on the validation data, we should not have access to the test data. Then when we have a good performance on the test data, we can say our model can adapt to new data well.