### **Hong Kong Baptist University**

### Department of Computer Science

*COMP 7990 Principles and Practices of data analytics (2022-23)*

Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ SID: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**In-class exercise A answer (Support Vector Machine):**

Open the file **iris.arff**. Use the default **Test-options Cross-validation folds 10**. Go to **Classify** tab and choose **SMO** algorithm. Click on **SMO** algorithm to review the algorithm configuration. Change the **filterType** to **No normalization/Standardization**

Fill in the following table by adjusting exponent and kernel:

|  |  |  |  |
| --- | --- | --- | --- |
| **Kernel and exponent used** | **PolyKernel**  **exponent = 1** | **PolyKernel**  **exponent = 2** | **RBFKernel** |
| **Correctly Classified Instances** |  |  |  |
| **Number of misclassified instances** |  |  |  |
| **Confusion matrix** **(screenshot)** |  |  |  |

**In-class exercise B answer (IBk / KNN):**

Continue to use the file **labor.arff** to answer the following questions:

1. Correctly Classified Instances (for **KNN value 10**) = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_%, which is \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ (higher/ lower) when comparing with the one using kNN value 5.
2. Correctly Classified Instances (for **KNN value 15**) = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_%, which is \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ (higher/ lower) when comparing with the one using kNN value 10
3. Accuracy is improved as k \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ (increases/decreases) from 1 to 10, then \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ (increases/decreases) when k=15.
4. **Change** the **KNN value = 10**, then change **Test option** to **Percentage split.** Try different percentages for training and testing set and fill in the table below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **KNN=10** | Percentage split=**50%** | Percentage split=**60%** | Percentage split=**70%** | Percentage split=**80%** |
| % of training data |  |  |  |  |
| % of testing data |  |  |  |  |
| Number of testing data  (Number of Instances) |  |  |  |  |
| Accuracy (%)  /Correctly Classified Instances (%) |  |  |  |  |

**In-class exercise C answer (Clustering using SimpleKMeans):**

Continue to use the file **weather.numeric.arff** to answer the following questions:

1. Change the classifier configuration. Try the following cases.

* Number of clusters = 2, seed = 11
* Number of clusters = 4, seed = 11

Remember to select **Use training set** in **Cluster mode**. **Ignore the attribute play**.

|  |  |
| --- | --- |
| **Number of clusters = 2, seed = 11** | **Number of clusters = 4, seed = 11** |
| Numbers of iteration = | Numbers of iteration = 4 |
| Within cluster sum of squared errors = | Within cluster sum of squared errors = |
| Initial starting points: (screenshot) | Initial starting points: (screenshot) |
| Final cluster centroids: (screenshot) | Final cluster centroids: (screenshot) |
| Clustered Instances: (screenshot) | Clustered Instances: (screenshot) |

# **Submission**

Submit the file **lab1-inclass-ans.docx** to bulearning website