### **Hong Kong Baptist University**

### Department of Computer Science

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

Name: Chi Zhang\_\_\_\_\_\_\_\_\_ SID: \_\_22407987

**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** | 145 | 144 | 140 |
| **Number of misclassified instances** | 5 | 6 | 10 |
| **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**) = \_\_\_91.2281\_\_\_\_\_\_\_\_%, which is higher (higher/ lower) when comparing with the one using kNN value 5.
2. Correctly Classified Instances (for **KNN value 15**) = \_\_\_\_80.7018\_\_\_\_\_\_\_\_\_\_\_\_\_%, which is lower (higher/ lower) when comparing with the one using kNN value 10
3. Accuracy is improved as k \_\_\_\_\_\_increases\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ (increases/decreases) from 1 to 10, then \_\_\_\_\_\_\_\_\_\_\_\_decreases\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ (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 | 50 | 60 | 70 | 80 |
| % of testing data | 50 | 40 | 30 | 20 |
| Number of testing data  (Number of Instances) | 28 | 23 | 17 | 11 |
| Accuracy (%)  /Correctly Classified Instances (%) | 78.5714 % | 73.913 % | 82.3529 % | 81.8182 % |

**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 = 3 | Numbers of iteration = 4 |
| Within cluster sum of squared errors = 13.619813754306595 | Within cluster sum of squared errors = 5.933983166627734 |
| Initial starting points: (screenshot)  IMG_256 | Initial starting points: (screenshot)  IMG_256 |
| Final cluster centroids: (screenshot)  IMG_256 | Final cluster centroids: (screenshot)  IMG_256 |
| Clustered Instances: (screenshot)  IMG_256 | Clustered Instances: (screenshot)  IMG_256 |

# **Submission**

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