Individual Coursework (ICW)

Semester 3 | 2020/2021

Module: CSC7073 | Principles & Practices of Machine Learning
Lecturer: Dr Reza Rafiee
Moderator: Professor Philip Hanna
Date released: 08/09/2021
Submission Due: 14/09/2021 (16:59) + an extra time because of the COVID-19 situation
Programme: MSc Software Development (Distance Learning)

Content Page

1. Individual coursework (ICW) project details

1.1 The dataset

In this project, there are three questions including unsupervised and supervised learning tasks. To answer these questions, you are required to utilise a **dataset** including 9126 samples with 440 features. The dataset is on Canvas and you can download it for this project.

The downloaded dataset, which is a tabular data, has the following format (see the **Figure 1**):
In this dataset, there are **9126 samples** in which every sample has **440 features**. A row in this dataset represents the values of a **feature** across all samples. ***ACTL6A_S5*** is **an example of a feature name.** A column in this dataset represents all features' values for a one specific sample. **Sample name** comprising a "sample id" (e.g., **TCGA.02.0047**), a "second name" (e.g., **GBM**) and a "subgroup" (e.g., **C4**) which all are attached together using a "**.**". As it has been illustrated in the **Figure 1**, these data in this dataset are represented by numeric values (e.g., 745.567). There is no missing data in this dataset.

1		TCGA.02.0047.GBM.C4	TCGA.02.0055.GBM.C4	TCGA.02.2483.GBM.C4	TCGA.02.2485.GBM.C4	TCGA.02.2486.GBM.C4	TCGA.04.1348.OV.C2	TCGA.04.1357.OV.C2	TCGA.04.1362.OV.C4
2	ACTL6A_S5	745.567	1154.31	1498.68	1320	1404.27	3504.629533	1293.399416	2882.595819
3	ADAM9_S2	4287.78	9475.54	2307.12	2685.71	2843.9	1107.601536	1064.576057	2617.298339
4	ADAMTS1_S5	241.556	6098.95	433.984	911.905	321.951	1956.185336	1916.188921	1367.380382
5	ADCY7_S3	1067.64	556.132	497.309	316.667	637.805	561.6090406	886.6116343	473.0080846
6	AIMP2_S5	406.736	537.088	752.148	785.552	792.963	1838.741423	805.56362	363.8475073
7	ALKBH7_S5	518.148	942.957	656.042	953.809	815.244	917.6066505	1461.944321	1250.306764
8	ALOX5AP_S3	1326.41	4211.35	566.543	307.143	5671.95	444.9242287	1130.604935	335.5019684
9	AMPD3_S3	326.992	361.598	196.728	80	542.683	241.0610312	318.0188471	214.5426725
10	APITD1_S5	184.308	319.535	311.443	260.462	494.488	466.9301231	336.125279	974.747003
11	APOC1_S3	1370.66	3093.48	3504.38	2482.86	12512.8	2287.531642	2819.476026	345.7059584
12	APOE_S3	32631	22377.6	20453.4	25919.5	67605.5	17272.72033	21633.67598	2686.479776
13	APOO_S5	374.935	558.935	390.501	411.429	540.244	465.869358	544.2361806	287.8394527
14	ARHGAP1_S2	2296.94	2491.94	2451.93	2808.57	2457.93	3029.908825	3414.708982	3271.568941
15	ARHGAP15_S3	153.047	214.156	95.8311	91.4286	234.146	88.54044478	242.1366028	34.69469372
16	ARHGDIA_S2	9756.29	7079.47	7478.63	5357.14	7302.44	8962.814476	6655.172151	4784.697387
17	ARRB2_S3	1828.57	2284.51	2240.42	1145.71	2794.51	1079.476551	1232.109697	633.6748711
18	B2M_S3	38492.3	119431	43296.3	45077.6	142230	78445.014	119750.8031	18271.47063
19	BCCIP_S5	1115.47	1175.04	1024.85	628.324	1010.76	1852.42725	1227.805073	1483.960418
20	BRCA2_S5	73.143	78.4863	76.8338	210	43.2927	224.0774349	60.74369163	180.5632399
21	BRIP1_S5	94.6556	43.7281	139.314	126.667	5.4878	95.93716627	31.94371897	74.51595117
22	BSG_S2	8056.79	19168.6	17766.8	18674.8	11197.6	19390.97735	22383.89089	11285.23148
23	BTK_S3	264.913	339.173	168.443	124.286	812.805	182.6892163	291.5055161	71.58660454
24	C11orf24_S5	749.869	1437.98	694.881	541.429	734.146	986.4355798	778.2170521	1642.232051
25	C12orf24_S5	184.394	352.067	537.414	277.143	357.927	271.5303189	70.81535273	116.9350421
26	C13orf1_S5	389.072	380.098	253.72	320.476	445.122	345.3082668	187.8937086	281.8963339
27	C13orf18_S3	235.41	311.142	97.0976	393.333	304.268	39.78503308	84.93013578	39.50361335
28	C13orf27_S5	115.554	220.322	570.343	200.476	239.024	260.175385	89.24907164	147.465956
29	C16orf61_S5	310.618	670.57	676.682	392.4	382.683	413.9935039	569.9874933	171.8548711

Figure 1 | Illustration of the dataset which will be used in ICW (partially illustrated)

Each sample in this datatset belongs to one of the 6 subgroups/subtypes (i.e., C1, C2, C3, C4, C5 and C6). You can find the subgroup of a sample in the last row of this dataset (i.e., reference subgroup).

1.2 The Project

ICW		
Topic		**Mark**
Dimension Reduction & Cluster Analysis	**A.** We aim to cluster 9126 samples of the dataset introduced in Section 1.1 into the six distinct subgroups (C1, C2, C3, C4, C5 & C6) using an unsupervised learning approach. Write a Python script in Jupyter Notebook for the following analyses: 1. **PCA projection**: Apply the principal component analysis (PCA) method on this dataset and identify the most significant principal components as the new features for the clustering analysis that you will be doing in section 3. Most significant components for this question are the ones by which the cumulative explained variances are more than 70% (5%). 2. **Visualisation of the projected dataset**: Illustrate a 3D PCA visualisation of samples (see the Note in below). Illustrate the plot of Cumulative Explained Variances vs. Number of Components (5%). 3. **Cluster Analysis**: By using a machine learning package in Python, apply k-means clustering algorithm on your projected dataset and cluster all samples into six subgroups. Evaluate the correctness of "the number of clusters" by illustrating a WCSS (**W**ithin-**C**luster-**S**um-of-**S**quares) graph. Is there any specific "the number of clusters" that you might find useful from this graph? Discuss your result (10%). 4. **Visualisation of samples** (i.e., data points): Visualise 3 principal components of samples according to the cluster labels obtained from your clustering result (5%). 5. **Quantitative performance evaluation:** Calculate the average silhouette of samples for your cluster analysis. Moreover, compare your clustering output with the reference subgroup. Provide this evaluation in the shape of a **confusion matrix** (10%). 6. **Interpret the quantitative performance result and suggest for improving the results**: Explain which subgroup comparatively represents the best and worst clustering performance result and any other conclusions on your results. You also must suggest how to improve the cluster analysis for obtaining a better result (5%). Note: Use the following colours for samples belonging to the various subgroups; C1: red, C2: yellow, C3: green, C4: cyan, C5: blue, C6: purple.	40%

ICW		
Topic		**Mark**
Supervised Learning - Classification (Optimal Model Complexity)	**B.** We aim to design a classification model using K-nearest neighbours (KNN) model for the dataset provided in Section 1.1. By Using this dataset, write a Python script in Jupyter Notebook for the following analyses: 1. Divide your dataset into 70% and 30% for a training and a test set, respectively. Train a KNN model by using the training set (2.5%). 2. Plot the model accuracy (i.e., model score) vs. different values of K (2.5%). 3. What is the best value of K (or range of K) by which you could obtain the optimal KNN model (2.5%)? 4. Explain your results in terms of bias and variance trade-off (2.5%).	10%
Supervised Learning – Classification (SVM)	**C.** We aim to design a classification model using support vector machines (SVM) for the dataset provided in Section 1.1. This dataset has 9126 samples and 440 features. Each sample belongs to one of the 6 subgroups (C1, C2, C3, C4, C5 & C6). By Using this dataset, write a Python script in Jupyter Notebook for the following analyses: 1. **SVM model development**: Design a SVM classifier model with a radial basis function (RBF) kernel in which their hyperparameters have been optimised: a. Read your dataset and then divide it into 80% and 20% for training/validation and test datasets, respectively (5%). b. What are the optimal values of cost and gamma parameters in your model (5%)? c. Train your model with the selected training set (5%). 2. **Quantitative performance evaluation of the model:** Evaluate the model by obtaining an accuracy and a confusion matrix: a. Obtain the training and test accuracies (5%). b. Show the performance of your model (i.e., training/validation and test) by illustrating confusion matrices (5%). 3. **Improve the model:** By discussing about the issues of the model, explain how you can address these issues and improve the classification model for this dataset (5%).	30%

2. Submission instructions

The submission due date of this ICW is on **September 14th, 2021 (by 16:59)**. Late submission penalties and rules will be applied in accordance with the QUB policy on late submission. For more information on this or any other QUB policy with regards to assessment please see:

https://www.qub.ac.uk/directorates/AcademicStudentAffairs/AcademicAffairs/ExaminationsandAssessment/MarkSchemesandClassifications/

If you have any questions or issues about this ICW please contact the module lecturer in the first instance.

You MUST submit **a single ipynb file** (including the following items) on Canvas (CSC7073 module, Assignments section, Individual Coursework Project):

1) Your script answers for question A, B & C must include the outputs of your running and also any interpretations, explanation and conclusions in the form of comments or markdown.

You must use the following naming format for your single file name:
StudentNumber_Firstname_Surname_ CSC7073_ICW_SubmissionDate.ipynb

Example:

1234567_Reza_Rafiee_CSC7073_ICW_041220.ipynb

Please remember to write **student number, first name, surname, CSC7073 and your submission date** as well as the following highlighted texts in the beginning of your script:

By submitting the work, I declare that:

1. I have read and understood the University regulations relating to academic offences, including collusion and plagiarism: http://www.qub.ac.uk/directorates/AcademicStudentAffairs/AcademicAffairs/GeneralRegulations/Procedures/ProceduresforDealingwithAcademicOffences/

2. The submission is my own original work and no part of it has been submitted for any other assignments, except as otherwise permitted.

3. All sources used, published or unpublished, have been acknowledged.

4. I give my consent for the work to be scanned using a plagiarism detection software.

3. Feedback mechanism

Feedback in the form of marks and some comments will be available as soon as attainable after submission with the expectation that marking will be completed and accordingly marks are provided by **three weeks (i.e., 21 working days)**.

Good Luck!