BENG 183

**Homework assignment 5**

**You need to clearly label the answers, and append the code you used to solve them.**

**All the files you need to download have been saved on google drive.** [**https://drive.google.com/drive/folders/1kyTPyr-1W7AjHjKNWC-fbmfGPqvdlci5?usp=sharing**](https://drive.google.com/drive/folders/1kyTPyr-1W7AjHjKNWC-fbmfGPqvdlci5?usp=sharing)

**A link to the** [**tutorial on Github**](https://github.com/jtburrows/BENG183Tutorials2023.git) **for this assignment**

**If you cannot access google from your place, email me (**[**jtburrow@ucsd.edu**](mailto:jtburrow@ucsd.edu)**) and I will send you the files directly.**

* 1. Assign the following models/algorithms A-J into the three following groups based on what they are usually used for. Some choices may appear in more than one category, but all choices will appear in at least one. (**3 pt**; 1 pt each category)

1. Affinity propagation
2. DBSCAN
3. K means
4. K nearest neighbor
5. Support vector machine
6. Linear regression
7. Logistic regression
8. Louvain method
9. Random forest

Classification: **Support Vector Machine, K nearest neighbor, Random Forest**

Regression: **Linear Regression, Logistic Regression**

Clustering: **Affinity Propagation, K means, DBSCAN, Louvain Method**

* 1. Briefly describe classification, regression, and clustering and how these methods differ from each other. (1 pt)

**Classification tries to predict the category or class of the input based on its features. Regression attempts to predict a numerical model that can continually produce accurate output values given input values. Clustering attempts to group the vast input based on similarities between various individual datapoints in the input.**

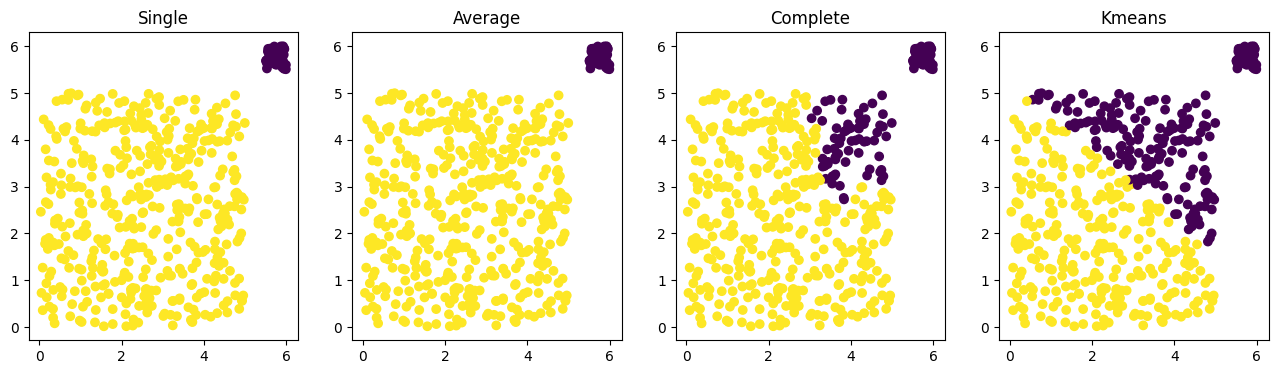
1. Download two datasets in Q2/. The two columns represent the (x,y) coordinates of data points. Use K-Means, hierarchical clustering with Euclidean distance and single, average, and complete linkage to partition the data into 2 clusters. For each clustering result, plot the data as a scatter plot and color the data by the cluster label you assigned. Based on the results, what advantages or limitations do you see in any of these methods? (**2.5 pt**; 2 datasets x 4 clustering methods, 0.25 pt each; conclusion 0.5 pt). Linked [here](https://github.com/jtburrows/BENG183Tutorials2023.git) is a tutorial for questions 2 and 3 of this homework.

Dataset 1:

A yellow and purple circles

Description automatically generated

Dataset 2:



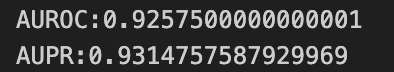
1. Area under the receiver operating characteristic curve (AUROC) and area under the precision recall curve (AUPR) are two metrics commonly used to measure the performance of binary classification. ROC curve is the true positive rate v.s. false positive rate. PR curve is the precision v.s. recall.

Download the single-cell gene expression data of three cell types in Q3/.

[Here](https://github.com/jtburrows/BENG183Tutorials2023.git) is another link to the tutorial for this section of the homework.

1. Train a logistic regression model to predict whether a cell belongs to cell type A or cell type B based on gene expression. Use five-fold cross-validation to measure the performance of your models, and **report the average AUROC and AUPR of the five testing sets**. When splitting the data into training and testing sets, make sure the proportion of positive and negative samples of each fold equals the proportion in the full dataset. You can achieve this by splitting positive and negative samples separately, and *scikit-learn* also has functions to do that. Use cell type A as positive label and cell type B as negative label first, and then exchange (A as negative and B as positive). Use the probability returned from the model (a real value between 0 to 1) to compute AUROC and AUPR, rather than the binary prediction result. (**0.5 pt**)

**For cell A positive and cell B negative:**



**For cell A negative and cell B positive:**

A black background with white numbers

Description automatically generated

1. Train a logistic regression model to predict whether a cell belongs to cell type A or cell type C based on gene expression. Use five-fold cross-validation to measure the performance of your models, and **report the average AUROC and AUPR of the five testing sets**. When splitting the data into training and testing sets, make sure the proportion of positive and negative samples of each fold equals the proportion in the full dataset. You can achieve this by splitting positive and negative samples separately, and *scikit-learn* also has functions to do that. Use cell type A as positive label and cell type C as negative label first, and then exchange (A as negative and C as positive). Use the probability returned from the model (a real value between 0 to 1) to compute AUROC and AUPR, rather than the binary prediction result. (**0.5 pt**)

**For cell A positive and cell C negative:**

A black background with white numbers

Description automatically generated

**For cell A negative and cell C positive:**

A number on a black background

Description automatically generated

1. What did you observe by exchanging the positive and negative labels? (**1 pt**)

**There weren’t any major differences when the positive and negative lables were swapped. The AUROC average increases slightly and the AUPR average decreases slightly.**

1. Use the data in b, and save the predicted probability of all the five folds into the same vector so that you will get the prediction probability of all data points when it serves in the testing set. Plot the ROC curve and PR curve using the cell type label (C as positive and A as negative) and the prediction probability. (**0.5 pt**) Explain why by changing different decision thresholds, the curves show the current shape (i.e. **what does the starting point and the ending point represent, why y goes up/down when x increases and whether the trend corresponds to increasing or decreasing thresholds**). (**2 pt**; 2 curves x 4 features to explain, 0.25 pt each)

**ROC Curve:**

A graph with blue lines

Description automatically generated

**PC Curve:**

A graph of a number of blue dots

Description automatically generated with medium confidence