Introduction to Data Science

Course 094201

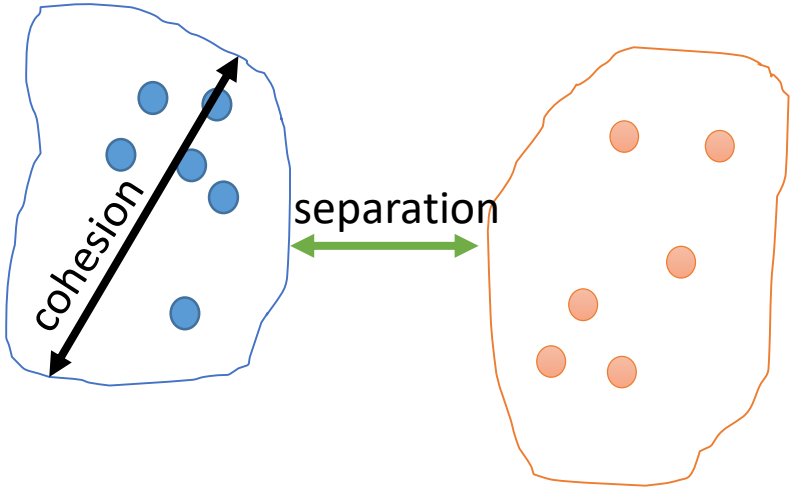
Lab 5:

Clustering Evaluation

Spring 2017

Cluster cohesion and separation

- Good clustering is one with tightly packed and well separated clusters



Cluster cohesion and separation

- **W**ithin cluster **s**um of **s**quares (WSS) is a measure of cluster **cohesion**:

$$WSS = \sum_{c_i \in C} \sum_{x_{j \in} c_i} (x_j - m_i)^2$$

Also called …

where x_j are points, C is a set of clusters, m_i is the centroid of cluster c_i

- **B**etween clusters **s**um of **s**quares (BSS) is a measure of cluster **separation**:

$$BSS = \sum_{c_i \in C} |c_i|(m - m_i)^2$$

Reminiscent of …

where m is the centroid of all points, $|c_i|$ is the number of points in cluster c_i

- **T**otal **s**um of **s**quares (TSS) is a property of the data: $TSS = \sum_{1 \le j \le n} (x_j - m)^2$
- There is a proof that: TSS=BSS + WSS
- Since TSS is a constant, regardless of the number of clusters or the partition, when we reduce WSS, BSS is grows and vice verse

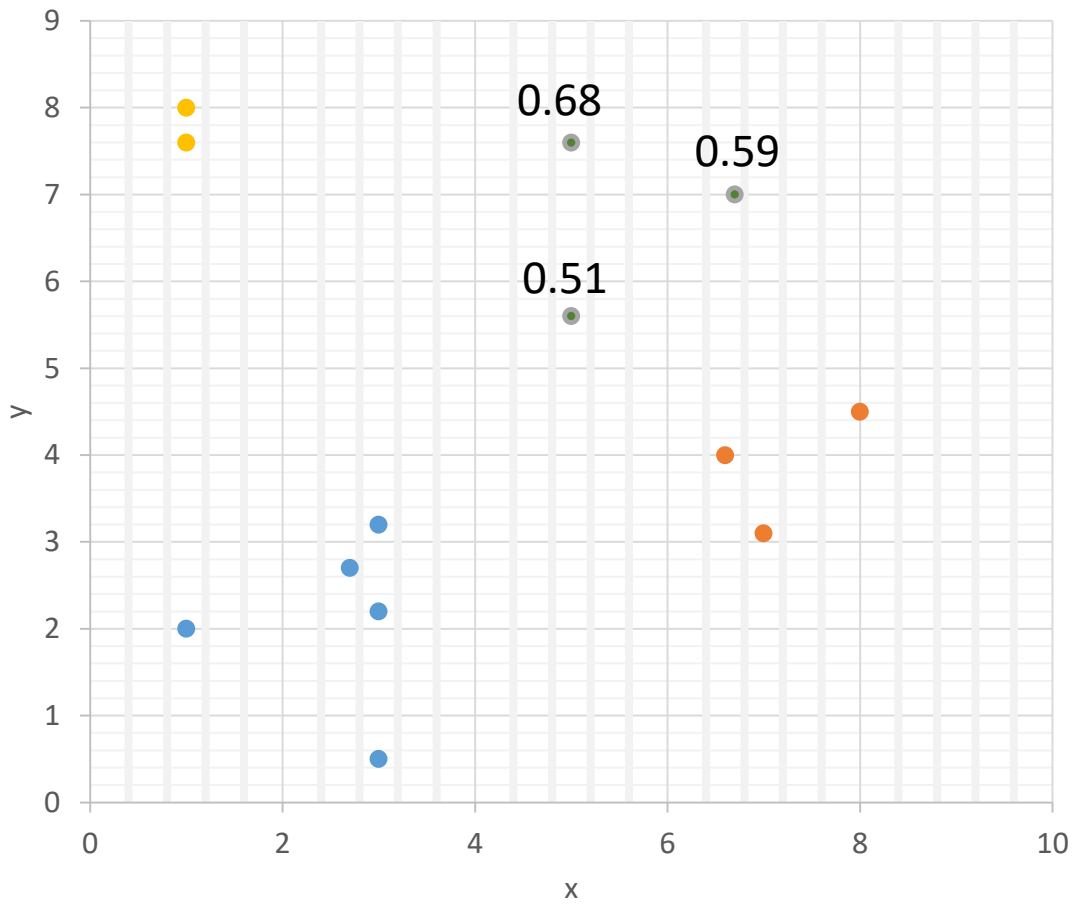
Silhouette coefficient

- Combines the cohesion and separation measures
- **Silhouette coefficient for point $\boldsymbol{x_j}$** in cluster c_i:
  - $in(x_j)$- the average distance of x_j to points $x_k \in c_i$;
  - $out(x_j)$- the minimal distance of x_j to other clusters. Distance between x_j and cluster c_m, $m \neq i$, is measured as the average distance between x_j and $x_v \in c_m$

$$s(x_j) = \frac{out(x_j) - in(x_j)}{\max(in(x_j), out(x_j))}$$

  - Ranges between -1 and 1
- Averaging over cluster's points – we get **Silhouette cluster coefficient**
- Averaging over all points – we get **Silhouette clustering coefficient**

Example



Silhouette coefficient overall: 0.75

- Cluster 1 : 0.72
- Cluster 2 : 0.73
- Cluster 3 : 0.6
- Cluster 4 : 0.96

The dataset and the code

- The code and the data can be found at:

/mnt/share/students/LAB5

- Copy everything to your **local folder** and unzip the code:

unzip lab5-students.zip

- Input argument for the code: the clustering results file
- The data folder contains several clustering outcomes for the Iris dataset and a single example given in class

The code – what do we have?

1. We have the classes “**Point**” and “**Cluster**” we already know. Some **additional methods** were added to both classes – these are used by the different evaluation measures.
2. The classes for “Silhouette coefficient” (SilhouetteCoeff) and “Purity” are given evaluation measures. **Both classes are inherit from the class “EvaluationMeasure” and use its members** (look into it).
3. Each of the classes for cluster evaluation includes a method called “calculate”. SilhouetteCoeff has three “calculate” methods, make sure you understand the goal of each.
4. The class ClusterReader reads the clustering output file (no need to invest your time in the parsing of the file) however notice that **this is the class that holds the clusters**, while the **evaluation measures classes** hold **only a reference** to the clusters vector (how and why is it preferable?)

Run the code

- Create a new project with the source files
- Change the C++ standard to be 98 in the file CMakeLists.txt
- Give as the argument a path to the example file shown in the lecture (shown in slide 5): /data/sample_out
- Understand what the output for Silhouette coefficient demonstrates
- Why calculating Purity on the artificial lecture example is not appropriate?

Assignment – In class

- Complete the class WSS:
  - Look at main.cpp to understand which methods you need to implement
  - Look at Silhouette coefficient class to see how it makes use of its base class members (specifically, look at the constructor)
  - See what member functions of Cluster class you can use
  - After you've finished you can remove the comments from printWSS function in main and run the code

Input files for home assignment

אחרי שתסיימו את המימוש של המחלקות לפי התיאור בעמוד הבא,
השלימו את הטבלה הבאה, שמתארת את קבצי הקלט השונים
ואת המדדים שתחשבו באמצעות הקוד.

קובץ	תיאור	Silhouette	BSS	WSS	TSS	Purity
out_average_iris	Average-link, 3 clusters					
out_complete_iris	Complete-link, 3 clusters					
out_single_iris	Single-link, 3 clusters					
iris_kmeans_3	KMeans, 3 clusters					
iris_kmeans_4	KMeans, 4 clusters					
iris_kmeans_5	KMeans, 5 clusters					
iris_kmeans_6	KMeans, 6 clusters					

Assignment - Home

1. השלימו גם את המחלקות BSS ו TSS. שימו לב, בעוד WSS וגם BSS הם מדדים לאיכות לקלאסטרינג, ה TSS אינו כזה. ה TSS הוא מדד לפיזור של הנתונים. עם זאת ממשו אותו לפי הנוסחה המקורית שלו (ולא כסכום של BSS ו WSS) על מנת שתוכלו לוודא את המימוש של מרכיביו. לצורך התרגיל על ה TSS לרשת ממחלקת EvaluationMeasure.

2. שימו לב בכל המימושים עליכם להשתמש בפונקציות מתאימות במחלקות האחרות, מה שיצמצם (מאוד) את כמות הקוד שתכתבו בפועל.

3. הריצו את כל המדדים הנ"ל על הקבצים בטבלה בשקף 10 ומלאו את הטבלה.

4. תחילה נדרג (נסדר לפי סדר המדד מהטוב לפחות טוב) את התוצאות השונות (הקבצים) לפי מדדים פנימיים (כאלה שלא דורשים ידע חיצוני על תיוג האירוסים). צרו טבלה חדשה בה בכל עמודה מופיעים הקבצים בסדר האיכות לפי אותו מדד. האם המדדים מסכימים ביניהם? מה הסיבה לאי-הסכמה לדעתכם? באופן ספציפי התייחסו להבדלים ב Silhouette בין ה single-link וה complete-link: הסבירו ממה נובע ה Silhouette הגבוה ב single-link

5. דרגו את תוצאות הקלאסטרינג לפי Purity. מה היא השיטה הטובה ביותר לקלאסטרינג? עם אילו מדדים פנימיים ישנה הסכמה?

6. יש המשתמשים ב Silhouette coefficient לבחירת מספר הקלאסטרים עבור Kmeans. מבין קבצי הקלט הנתונים לכם, כמה קלאסטרים תבחרו? האם הבחירה מעידה על כמות סוגי האירוסים הקיימים בפועל? הסבר ונמק.

עליכם להגיש את כל קבצי הקוד ללא קבצי הפרויקט וקובץ PDF בודד עם תשובות לשאלות הנ"ל.