Artificial Intelligence I 2023/2024

Week 6 Tutorial and Additional Exercises

K-means & WEKA GUI

School of Computer Science

19th February 2024

In this tutorial...

In this tutorial we will be covering

- K-means Clustering.
- WEKA Exercises

K-means clustering

Recall the formal algorithm of *K-means clustering*:

Algorithm 1: K-means clustering.

Input: $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}$: data set, K: number of clusters, $\mathbf{c}_1, \ldots, \mathbf{c}_K$: initial centroids.

Output: Clusters.

1 **repeat**

2 **Assignment step**: For all $i = 1, \ldots, N$, form clusters by assigning

$$Cluster(\mathbf{x}^{(i)}) \leftarrow \arg \min_{k=1,\ldots,K} Dist(\mathbf{x}^{(i)}, \mathbf{c}_k)$$

where $Dist(\cdot)$ is some distance function, e.g. squared Euclidean distance;

3 **Refitting step**: For all $k = 1, \ldots, K$ compute the centroid of the obtained clusters as

$$\mathbf{c}_k \leftarrow \frac{1}{n_k} \sum_{\{i:Cluster(\mathbf{x}^{(i)})=k\}} \mathbf{x}^{(i)}$$

where n_k is the number of examples in the k-th cluster.

4 **until** $\mathbf{c}_1, \ldots, \mathbf{c}_K$ *"stop changing"*;

5 **return** *Clusters.*

Exercise 1

- Consider the following data set of 8 examples, each consisting of 2 features

$$\mathbf{x}^{(1)} = \begin{bmatrix} 3 \\ 1 \end{bmatrix}, \mathbf{x}^{(2)} = \begin{bmatrix} 5 \\ 1 \end{bmatrix}, \mathbf{x}^{(3)} = \begin{bmatrix} 4 \\ 2 \end{bmatrix}, \mathbf{x}^{(4)} = \begin{bmatrix} 5 \\ 2 \end{bmatrix},$$

$$\mathbf{x}^{(5)} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}, \mathbf{x}^{(6)} = \begin{bmatrix} 7 \\ 4 \end{bmatrix}, \mathbf{x}^{(7)} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \mathbf{x}^{(8)} = \begin{bmatrix} 8 \\ 0 \end{bmatrix}.$$

- Use K-means in algorithm 1 to cluster these examples.
- Use $K = 2$, the squared Euclidean distance as the distance function, and the following initial centroids:

$$\mathbf{c}_1 = \begin{bmatrix} 3 \\ 0 \end{bmatrix} \text{ and } \mathbf{c}_2 = \begin{bmatrix} 5 \\ 7 \end{bmatrix}.$$

- Find the clusters and the new centroids at the end of one iteration of algorithm 1 (lines 2 and 3).

Exercise 1: Solution

- We first find the squared Euclidean distance of each example from the two centroids:

	$\mathbf{x}^{(1)}$	$\mathbf{x}^{(2)}$	$\mathbf{x}^{(3)}$	$\mathbf{x}^{(4)}$	$\mathbf{x}^{(5)}$	$\mathbf{x}^{(6)}$	$\mathbf{x}^{(7)}$	$\mathbf{x}^{(8)}$
$L^2(\cdot, \mathbf{c}_1)^2$	1	5	5	8	10	32	4	25
$L^2(\cdot, \mathbf{c}_2)^2$	40	36	26	25	25	13	65	58

- The assignment therefore is the following:

	$\mathbf{x}^{(1)}$	$\mathbf{x}^{(2)}$	$\mathbf{x}^{(3)}$	$\mathbf{x}^{(4)}$	$\mathbf{x}^{(5)}$	$\mathbf{x}^{(6)}$	$\mathbf{x}^{(7)}$	$\mathbf{x}^{(8)}$
$Cluster(\cdot)$	1	1	1	1	1	2	1	1

- The new centroids are the following:

$$\mathbf{c}_1 = \begin{bmatrix} 4 \\ 9/7 \end{bmatrix}, \text{ and } \mathbf{c}_2 = \begin{bmatrix} 7 \\ 4 \end{bmatrix}.$$

- Run a second iteration of algorithm 1. Find the new clusters and the new centroids.

Exercise 2

- Reconsider the following set with 8 examples and 2 variables

$$\mathbf{x}^{(1)} = \begin{bmatrix} 3 \\ 1 \end{bmatrix}, \mathbf{x}^{(2)} = \begin{bmatrix} 5 \\ 1 \end{bmatrix}, \mathbf{x}^{(3)} = \begin{bmatrix} 4 \\ 2 \end{bmatrix}, \mathbf{x}^{(4)} = \begin{bmatrix} 5 \\ 2 \end{bmatrix},$$

$$\mathbf{x}^{(5)} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}, \mathbf{x}^{(6)} = \begin{bmatrix} 7 \\ 4 \end{bmatrix}, \mathbf{x}^{(7)} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \mathbf{x}^{(8)} = \begin{bmatrix} 8 \\ 0 \end{bmatrix}.$$

- Use K-means in algorithm 1 to cluster these examples.
- Use $K = 2$, the squared Euclidean distance as the distance function, but the initial centroids are now:

$$\mathbf{c}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \text{ and } \mathbf{c}_2 = \begin{bmatrix} 4 \\ 0 \end{bmatrix}.$$

- Find the clusters and the new centroids at the end of one iteration of algorithm 1 (lines 2 and 3).

Exercise 2: Solution

- We first find the squared Euclidean distances of each example from the two centroids:

	$\mathbf{x}^{(1)}$	$\mathbf{x}^{(2)}$	$\mathbf{x}^{(3)}$	$\mathbf{x}^{(4)}$	$\mathbf{x}^{(5)}$	$\mathbf{x}^{(6)}$	$\mathbf{x}^{(7)}$	$\mathbf{x}^{(8)}$
$L^2(\cdot, \mathbf{c}_1)^2$	5	17	13	20	10	52	0	49
$L^2(\cdot, \mathbf{c}_2)^2$	2	2	4	5	13	25	9	16

- The assignment therefore is the following:

	$\mathbf{x}^{(1)}$	$\mathbf{x}^{(2)}$	$\mathbf{x}^{(3)}$	$\mathbf{x}^{(4)}$	$\mathbf{x}^{(5)}$	$\mathbf{x}^{(6)}$	$\mathbf{x}^{(7)}$	$\mathbf{x}^{(8)}$
$Cluster(\cdot)$	2	2	2	2	1	2	1	2

- The new centroids are the following:

$$\mathbf{c}_1 = \begin{bmatrix} 3/2 \\ 3/2 \end{bmatrix}, \text{ and } \mathbf{c}_2 = \begin{bmatrix} 16/3 \\ 5/3 \end{bmatrix}.$$

- Run a second iteration of algorithm 1. Find the new clusters and the new centroids.

WEKA

- *Waikato Environment for Knowledge Analysis* (WEKA) is a collection of machine learning algorithms and data pre-processing tools.
- It was developed by *University of Waikato, New Zealand*.
- Documentation can be found in the following links:
  1. `https://www.cs.waikato.ac.nz/ml/WEKA/Witten_et_al_2016_appendix.pdf`
  2. `https://user.eng.umd.edu/~austin/ence688p.d/handouts/WEKAManual2018.pdf`
- Access via the school server: `https://jupyterhub.oc1.aws.cs.bham.ac.uk`
- We next do exercises about loading datasets and implementing algorithms in WEKA.

Exercise 3: Loading Data in WEKA GUI

- Load soybean.arff dataset.
- Identify the number of instances and the number of non-class attributes.
- How many distinct class labels exist in the data set?
- How many numeric and how many nominal attributes are there in the dataset?
- From the dataset, delete the attribute 'temp'.
- Repeat the same for the dataset ionosphere.arff.

Exercise 4: Pre-processing via Filters

- Load the weather.arff dataset. Use the filter `WEKA.filters.unsupervised.attribute.Remove` to remove the humidity attribute. To do this, first set the filter as above. The text "Remove" will appear in the field next to the Choose button. Click on the field containing this text. The Generic Object Editor window opens and click on it to get a fuller description. Enter the correct attribute index and click the OK button. The window with the filter options closes. Now click the Apply button on the right, which runs the data through the filter.
- Undo the change to the dataset that you just performed, and verify that the data has reverted to its original state.

Exercise 4: Pre-processing via Filters (continued)

- Use the filter `WEKA.unsupervised.instance.RemoveWithValues` to remove all instances in which the humidity attribute has value higher than 70. To do this, first make the field next to the Choose button show the text RemoveWithValues. Then click on it to get the Generic Object Editor window, and figure out how to change the filter settings appropriately.

Exercise 5: Classification in WEKA

- Load housing.arff data file.
- Select Linear Regression from `WEKA.classifiers.function`
- Identify the default classifier settings and report it. To do this, after selecting the algorithm, click on the name of the algorithm to review the algorithm configuration.
- After training the model, what are the available test options in WEKA? Run the different test options and compare their outputs.
- Compare the results of running 5-fold and 10-fold cross-validation.

Exercise 6: K-Means in WEKA

- Load vote.arff.
- Choose SimpleKMeans clustering algorithm.
- Report the default number of clusters, the distance function used, the maximum number of iterations, the initialization method, and the seed number for the random number generator.
- Change the number of clusters to 4.
- Identify the various cluster modes.
- Choose 'use training set' cluster mode and run the algorithm. Report the within cluster sum of squared errors.
- Re-run it with 'use training set' cluster mode. Did the output change? Why do you think it has not changed? Try changing the random seed and run the algorithm again. Has the output changed?

Exercise 6: K-Means in WEKA (continued)

- Change the cluster mode to 'Classes to Clusters Evaluation'.
- How is classes to clusters evaluation performed?
  - During training, WEKA ignores the class attributes and generates the clustering.
  - During testing, it assigns classes to the clusters based on the majority value of the classes within each cluster. Then, it computes the classification error based on this assignment and shows the corresponding confusion matrix.
- Run the algorithm with this cluster mode. Report the confusion matrix. What is the majority class label in the 2nd cluster?