Lecture 9:
Modeling Data with KNN

COSC 526:
Introduction to Data Mining
Spring 2021

THE UNIVERSITY OF
TENNESSEE
KNOXVILLE
KNOXVILLE
BIG ORANGE. BIG IDEAS.®

Lecture Outline

- Assignment 8
  - Discuss results
- Assignment 9
  - Apply the DBSCAN clustering method
- Project discussion
  - Title and 3 key paragraphs
- Modeling data with KNN
  - Use Case: Modeling soil moisture

Assignment 8

Assignment 8 – Problem 1

- Data file: ./data/data-1.csv
  - Contain no missing values
- Problem: Cluster food items in the file based on carbohydrate and fat:
  - *Use the k-Means from the Spark MLlib to cluster data points*
  - *Determine the optimal value for k using the* **elbow method**
- Metric of quality: Use Within Set Sum of Squared Errors
  - This method is built into the Spark kMeans model and can be accessed with model.computeCost()
- Note: the clusters provide a ground truth for comparison to clusters we find later using data with missing values.

Assignment 8 – Problem 1

Steps:

- Define the optimal value for K
- Cluster food items (using K)
- Plot clusters by *fat* and *carbohydrate* content

Assignment 8 – Problem 2

- Data file: ./data/data-2.csv
  - Missing values for the carbohydrate content of some food items
  - The data were removed from a specific set of food items (i.e., food items with carbohydrate value near 0.5)
- Define a method to remove food items with missing any macronutrient values and apply this method to the data.
- Cluster the modified data and plot the results
  - Use the same K value as in Problem 1
- Note: we provide you with the code that loads the data and reports the percentage of values missing for each macronutrient (i.e., carbohydrates and fat)

Assignment 8 – Problem 3

- Data file: ./data/data-3.csv
  - Missing values for the fat or carbohydrate content of some food items (but not both for a single food item).
  - The data were removed from a food items randomly

PART 1:

- Define and apply a method to fill missing values with the mean of other values
  - *E.g., for missing values in fat, fill with the mean of fat values that are present*
- *Cluster the modified data and plot the results*
  - Use the same K value as in Problem 1

Assignment 8 – Problem 3

PART 2:

- *Use the code for Problem 2 to remove data with missing values rather than filling the gaps (as you did in PART 1)*
- *Cluster the modified data and plot the results*
  - Use the same K value as in Problem 1

Assignment 8 – Problem 4

- **Observe and describe:** Can you summarize your findings in each problem? Can you compare and contrast the findings across problems? How did each method for dealing with missing data (i.e., remove of filling) change the clustering outcome?
- **Impact of K:** What value did you choose for K in Problems 1-3? You based the selection of your K on the first dataset (i.e., no missing data). Do you expect a different value of K if you had used the elbow method with the second or third dataset? If yes, propose changes to your current solutions.

Assignment 8 – Problem 4

- **Building assumptions on data distributions:** Now look at the plot of clusters in Problem 1. Logically, there cannot be more than 1 gram of (carbohydrate + fat) in 1 gram of food. In your plot this can be seen in the form of a diagonal line from the top-left to bottom-right (where the sum of fat and carbohydrate content is equal to 1). How can you use this information to improve the way you fill missing values? Can you think of other methods to fill missing values? (HINT: logistic regression)

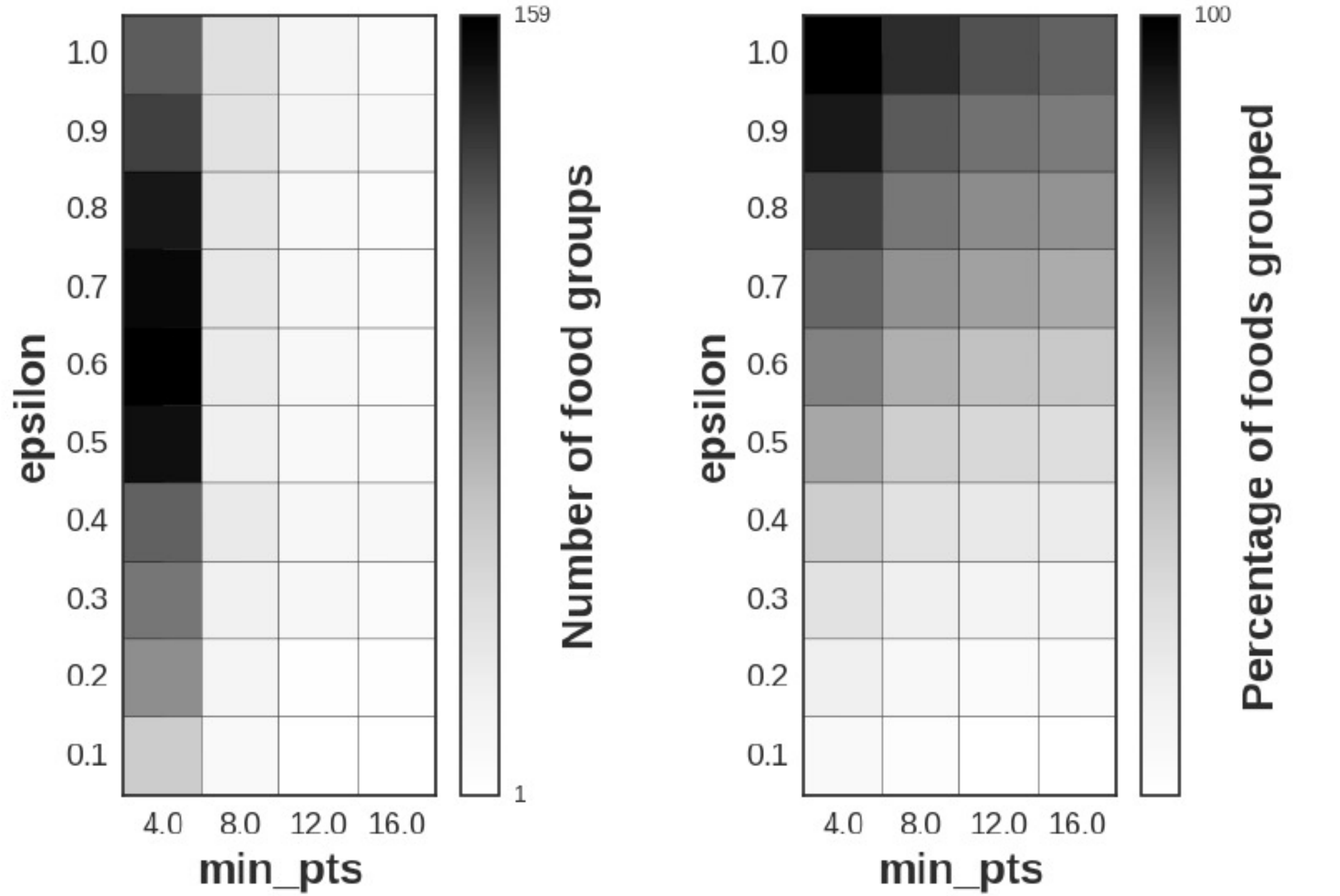
Assignment 9

Assignment 9

- We use Density-based spatial clustering of applications with noise (DBSCAN) to cluster food items based on macronutrient and micronutrient content.
  - We learned about DBSCAN in the last lecture.
- We will reference the paper *Development of a Scalable Method for Creating Food Groups Using the NHANES Dataset and MapReduce
  - The paper can be downloaded from the ACM Digital Library -- available through the University
  - The paper is also in the course repo
- We use the data preprocessing code and the DBSCAN code provided in this paper's git repo.
  - https://github.com/TauferLab/NHANES-Analytics

Assignment 9

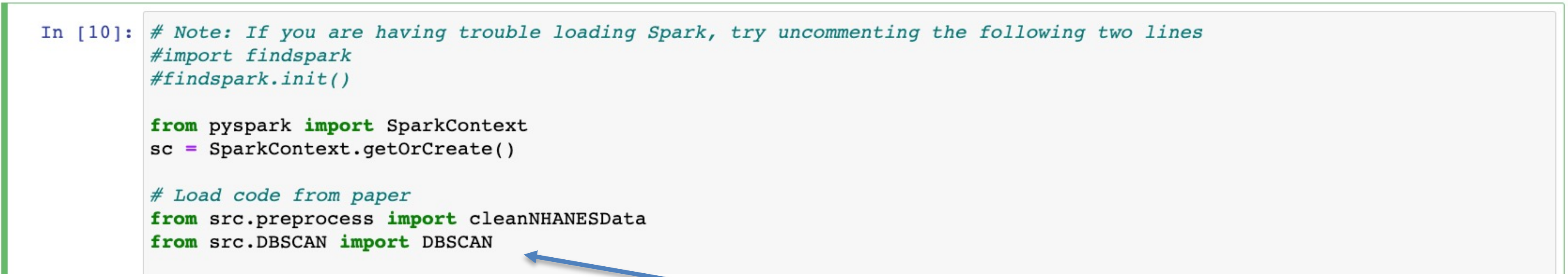
- We preprocess the raw NHANES dietary data
  - Data location: ./data/NHANES-20**-dietary.csv
- We cluster the data with DBSCAN
- We reproduce Figure 6 from the
  - Your figures may be slightly different because you are only using 2 years of NHANES data, where the paper used 7 years of data



(a) Number of clusters

(b) Clustered items

Figure 6: Heat maps showing the number of clusters (a) and the percentage of clustered food items (b) when the Euclidean distance is used.

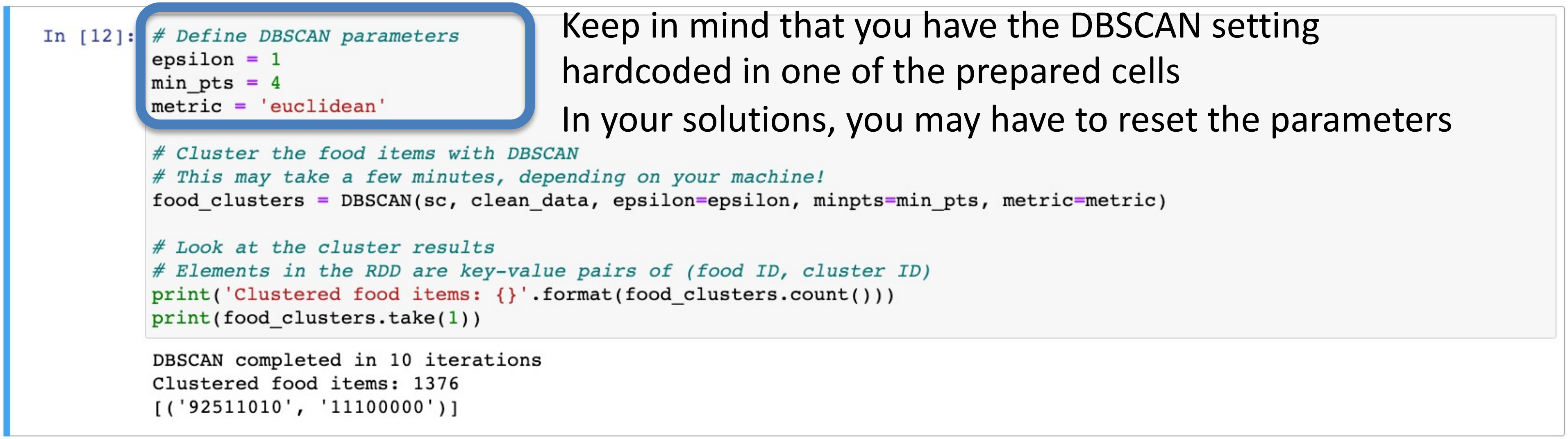


```
In [10]: # Note: If you are having trouble loading Spark, try uncommenting the following two lines
         #import findspark
         #findspark.init()

         from pyspark import SparkContext
         sc = SparkContext.getOrCreate()

         # Load code from paper
         from src.preprocess import cleanNHANESData
         from src.DBSCAN import DBSCAN
```

We are giving you the DBSCAN sw
Running on spark



```
In [12]: # Define DBSCAN parameters
         epsilon = 1
         min_pts = 4
         metric = 'euclidean'

         # Cluster the food items with DBSCAN
         # This may take a few minutes, depending on your machine!
         food_clusters = DBSCAN(sc, clean_data, epsilon=epsilon, minpts=min_pts, metric=metric)

         # Look at the cluster results
         # Elements in the RDD are key-value pairs of (food ID, cluster ID)
         print('Clustered food items: {}'.format(food_clusters.count()))
         print(food_clusters.take(1))
```

```
DBSCAN completed in 10 iterations
Clustered food items: 1376
[('92511010', '11100000')]
```

Keep in mind that you have the DBSCAN setting hardcoded in one of the prepared cells

In your solutions, you may have to reset the parameters

Assignment 9

- Problem 1: Measure metrics to characterize clusters
  - *Define the functions which intake food clusters (i.e., the output of DBSCAN) and returns **percentage of food items clustered** and **number of food clusters found**.*
- Problem 2: Plot a heatmap
  - *Define the function that intakes a 2D array and plots a heatmap*
- Problem 3: Recreate Figure 6 from the paper
  - *Cluster food items with the **Euclidean distance metric***
    and epsilon and min_pts values in ranges [2,4] and [4,7], respectively

Assignment 9

- Problem 4: Visualize n-dimension spaces
  - *Project the data down to 2 dimensions and visualize the clusters*
  - *Use t-Distributed Stochastic Neighbor Embedding (t-SNE) for dimensionality reduction*
- Problem 5: Euclidian, or not Euclidian, “that is the question”
  - *Use the given code to examine the clusters found using DBSCAN and Cosine Similarity and compare with clusters found using the Euclidian distance*

Project Steps

Project (I)

- Step 0: search for datasets
  - Discussion in class of datasets identified

Assignment 6 – Problem 5 DONE!

Find an interesting dataset.

- Is the dataset available? Do you have the dataset source (with url)?
- It your dataset large enough to allow for interesting analysis and non-trivial results?
- How were the data were collected?
- What is the significance of the data?
- Describe the number of rows, objects, or data points
- What information is contained in each rows, objects, or data points?
- What is the data types (int, str, char, float, etc.) and numerical ranges where appropriate
- What file(s) do you need to parse (if multiple files are available)?

Project (II)

- Step 1: Address (and answer) these questions
  - Do you work alone or in team? If in team, indicate who is your collaborator and what are the skills / expertise he/she is bringing to the project
  - What is the dataset you are considering?
  - What are the possible key question(s) you want to answer? Are the questions too general? Are the questions too narrow?

Assignment 7 – Project Questions DONE

Answer the following questions, in a couple sentences each, in the cells provided below. Note that you can indicate that you do not have a dataset or/and a set of questions, and you wish to be helped in completing this task.

- Do you wish to work alone or in team? If in team, indicate what are the skills / expertise he/she/they can bring to the project. Suggestion: you may consider to work in team with another student who works in a different scientific domain.
- Have you identified your dataset? If yes, what is the dataset you are considering?
- Have you identified possible key question(s) you want to answer? If yes, list the questions.

Project (III)

- What are the scientific questions you are answering with your tool/framework? Be as specific as possible.
- What is the tentative title of your project?
- What are the milestones you want to meet from now until the end of the semester when you will present your poster at the poster showcase? Be as specific as possible. Write the date of the deadline and the task(s) you want to achieve for that deadline.

Assignment 8 – Project Questions DONE

Answer the following questions, in a couple sentences each, in the cells provided below.

- What are the scientific questions you are answering with your tool/framework? Be as specific as possible.
- What is the tentative title of your project?
- What are the milestones you want to meet from now until the end of the semester when you will present your poster at the poster showcase? Be as specific as possible. Write the date of the deadline and the task(s) you want to achieve for that deadline.

Project (IV) Assignment 9 THIS WEEK

Motivation Describe the motivation of your work. To build the motivation, you can answer these questions:

- What is the problem you are tackling?
- How is the problem solved today?
- Write a paragraph of 200 - 300 words

Contributions List between 2 and 4 contributions of your work. Contributions are bullet points that define your solution. E.g.,

- We build a system that ....
- We validate the system accuracy by ....
- We measure the performance of the system by ...
- Write a section of 150 - 200 words

Project (IV) Assignment 9 THIS WEEK

Tests List the type of tests (measurements) you will perform. E.g.,

- What are your metrics of success?
- Where do you run your tests?
- What tests do you perform?
- How many times do you run each test?
- What do you measure?
- Write a section of 250 - 350 words.

Complete the answers in this form:

- https://forms.gle/gg73noS52HWW29bW7