K-Nearest Neighbors

Ko, Youngjoong

Dept. of Computer Science & Engineering,
SKKU

Index

- **Introduction**
- **K-NN**
- **Assignment**
- **Cautions**

Introduction

- **Goal**
  - The HW5 is to implement K-NN model for text classification.

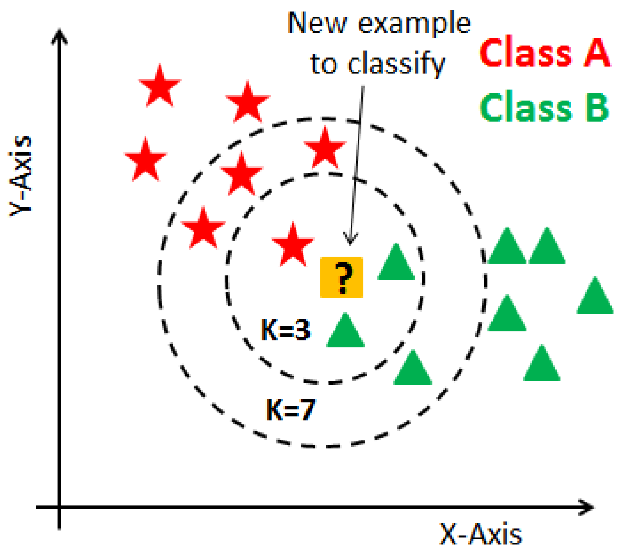
  - First, we will briefly explain how the K-NN model works.

K-NN

How The K-NN Model Works?

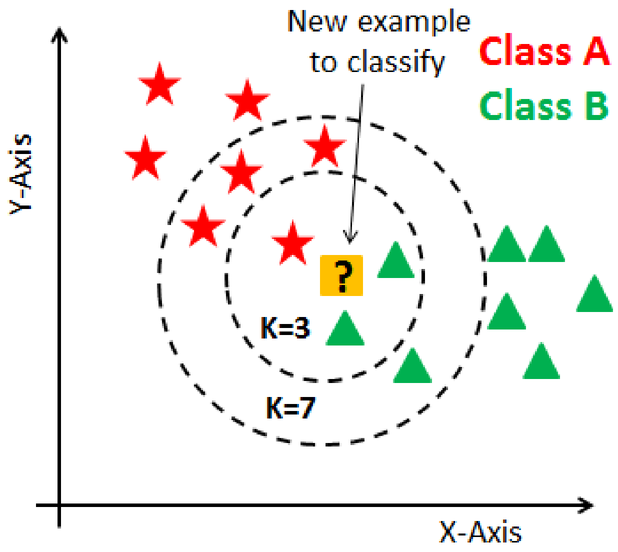
- Assume that there are two classes for the data and that there are already labeled data.

- Let’s consider the below figure.



K-NN

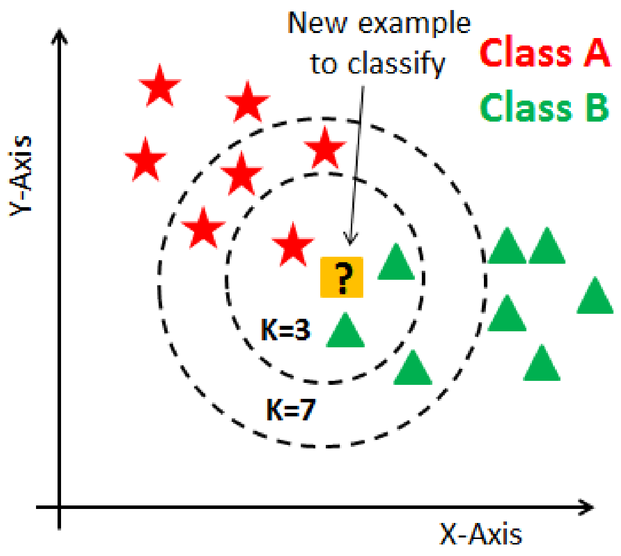
- **How The K-NN Model Works?**
  - To assign a label to new data, you have to find K number of nearest data (neighbors).
  - The label to be assigned can be different depending on the **K** that you can select.



K-NN

How The K-NN Model Works?

- If you choose 3 as the K, the new data will be assigned as triangle.
- If you choose 7 as the K, the new data will be assigned as star.



K-NN

- **How The K-NN Model Works?**
  - Depending on what **distance metric** you want to use, the nearest neighbors can be different and this means that the label to be assigned also can be different.
  - There are several metrics for calculating distance between two data points such as Manhattan distance, Euclidean distance, Cosine similarity and so on.

Assignment

- **Various Possible Inputs**
  - TF.
  - TF-IDF.
  - Bag-of-words.
    - 1 if a word in document else 0.
    - For example,

      doc_a = I love dog doc_b = I like cat

      Vocab = { cat, dog, i, like, love }

      vec_a = [0, 1, 1, 0, 1] vec_b = [1, 0, 1, 1, 0]

Assignment

Distance Metrics

- Formula for the distance function
  - Cosine similarity between two vectors:

$$similarity = cos(\Theta) = \frac{A \cdot B}{||A||\ ||B||} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}}$$

  - Euclidean distance between two vectors:

$$\mathrm{D}_{\mathrm{euclidean}} = \sqrt{\sum_{i=0}^{n}(p_i - q_i)^2}$$

where $p_i \in P, q_i \in Q$ and $P, Q \in \mathbb{R}^n$.

Given Data

- **train.json & test.json**
  - Data Format : a list of dictionaries
  - Data Example :

```
[
    {
        “paragraph” : “Since Game of Throne first aired, …”
        “label”: “tv”,
        “id”: “31”     *this is an article id
    },
    …
]
```

  - The number of Labels : 5 (finance, lifestyle, tv, sports, entertainment)

Assignment

- **Assignment**
  - Implement the ‘K-NN’ model for text classification in main.py.
    - Use TF, TF-IDF and bag-of-words vectors as input for the model respectively.
      - ✓ You can use the TF-IDF function that you have already implemented for hw3.
    - Use Euclidean distance and cosine similarity as distance metrics respectively.
    - Use k = 11.
  - See page 13 for specific output format.

Assignment

- **Submissions**
  1) StudentName _StudentID_main.py (python version 3.x)
     - e.g., 홍길동_2020123123_hw5_main.py.
       e.g., MichaelJackson_2020123123_hw5_main.py

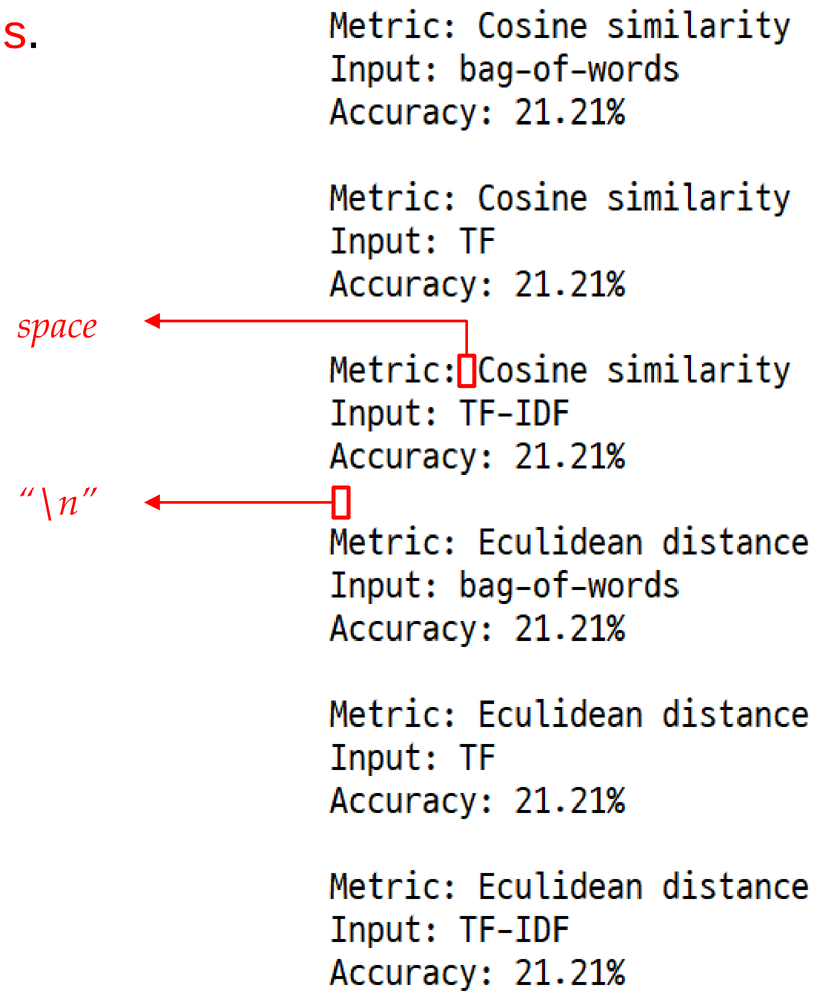
  2) StudentName _StudentID.txt
     - e.g., 홍길동_2020123123_hw5.txt
       e.g., MichaelJackson_2020123123_hw5.txt

Assignment

Outlook of the Text File

- Be careful of capitals and spellings.

- Round to second decimal place.



```
Metric: Cosine similarity
Input: bag-of-words
Accuracy: 21.21%

Metric: Cosine similarity
Input: TF
Accuracy: 21.21%

Metric: Cosine similarity
Input: TF-IDF
Accuracy: 21.21%

Metric: Eculidean distance
Input: bag-of-words
Accuracy: 21.21%

Metric: Eculidean distance
Input: TF
Accuracy: 21.21%

Metric: Eculidean distance
Input: TF-IDF
Accuracy: 21.21%
```

Cautions

- **Cautions**
  - Use ‘Python3’ and ‘Google Colab’.
  - Do not import any library except already imported libraries.
  - Copy will be scored 0.

Thank you for your attention!

고 영 중 (Ko, Youngjoong)

http://nlp.skku.edu/