SSID: WWCode
Password: password

WOMEN WHO CODE MANILA

Machine Learning & AI Study Group

Twitter: @wwcodemanila
FB: fb.com/wwcodemanila

#WWCodeManila
#YourProgrammingLanguage
#StudyGroup

Issa Tingzon

Research Fellow

Philippine-California Advanced Research Institutes

OUR MISSION

Inspiring women to excel in technology careers.

WOMEN WHO CODE MANILA

OUR VISION

A world where women are representative as technical executives, founders, VCs, board members and software engineers.

WOMEN WHO CODE MANILA

STUDY GROUP

Study groups are events where women can come together and help each other learn and understand a specific programming language, technology, or anything related to coding or engineering.

GUIDELINES

- If you have a question, just **ask**
- If you have an idea, **share it**
- **Make friends** and learn from your study groupmates
- **Do not** promote your recruit or promote your business

New Member's Introduction

I am <name>

<your current profession>

<why did you join this study group?>

SHOW & TELL

STUDY GROUPS

Study Group 1: Machine Learning Basics

Study Group 2: Data Preprocessing

AGENDA

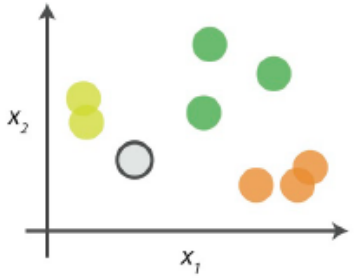
1. **Quick Review**: KNN Algorithm
2. **New Topic**: Data Preprocessing
3. Exercise
4. Presentations

REVIEW

K-Nearest Neighbor (KNN)

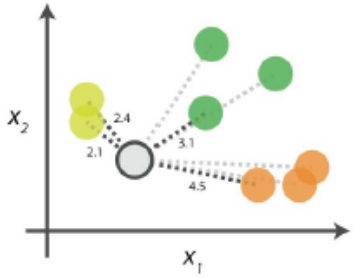
- One of the simplest ML Algorithms
- Steps:
  1. Compute the Euclidean distance between the “new observation” and all training data points
  2. Select the K nearest observations and perform a majority vote
  3. Assign the corresponding label to the observation

0. Look at the data



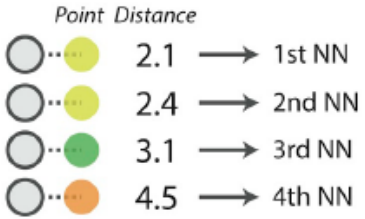
Say you want to classify the grey point into a class. Here, there are three potential classes - lime green, green and orange.

1. Calculate distances



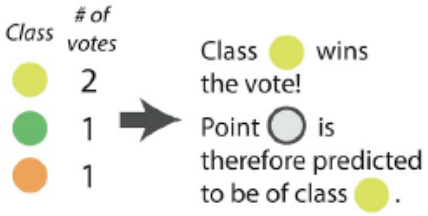
Start by calculating the distances between the grey point and all other points.

2. Find neighbours



Next, find the nearest neighbours by ranking points by increasing distance. The nearest neighbours (NNs) of the grey point are the ones closest in dataspace.

3. Vote on labels



Vote on the predicted class labels based on the classes of the k nearest neighbours. Here, the labels were predicted based on the k=3 nearest neighbours.

KNN Cheat Sheet

Importing the library:

```
from sklearn.neighbors import KNeighborsClassifier
```

Instantiating a model:

```
knn = KNeighborsClassifier(n_neighbors=3)
```

Fitting model to training set:

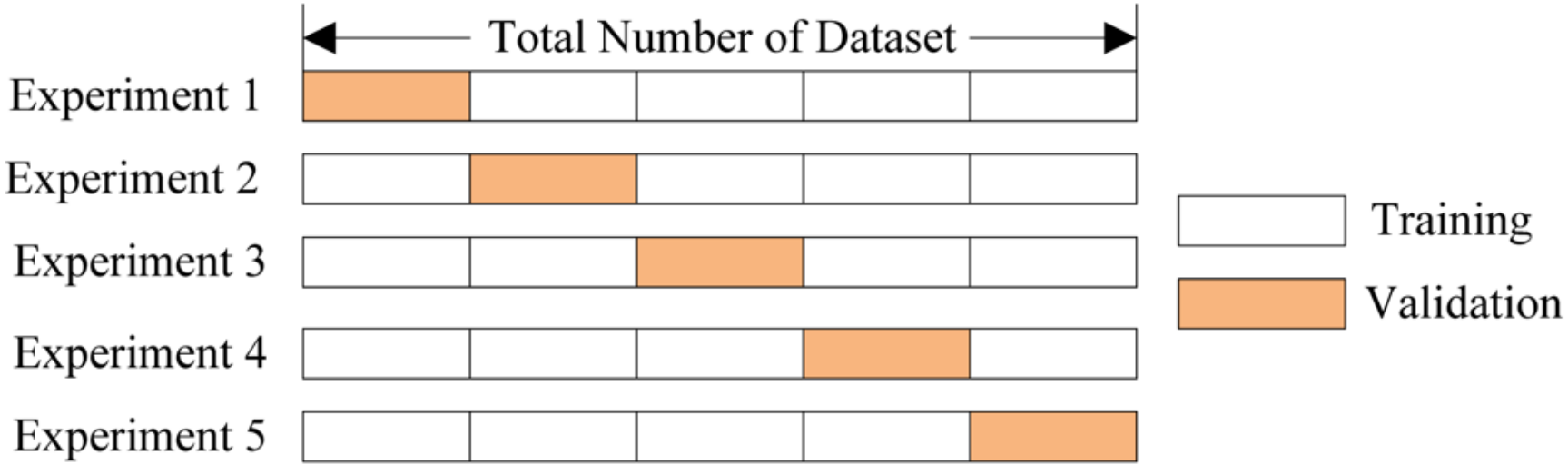
```
knn.fit(X_train, y_train)
```

Predicting test set:

```
y_pred = knn.predict(X_test)
```

k-fold Cross Validation

For hyperparameter tuning (i.e. choosing the “right” K)



TODAY'S TOPIC

DATA PRE-PROCESSING: FEATURE SCALING

FEATURE SCALING

Different features → measured on different scales.

- height – centimetres
- weight – kilograms
- blood pressure in mmHg
- etc.

Some classifiers combine and compare feature values (e.g. Euclidean distance).

FEATURE SCALING

Features with a broad range of values → dominate features with a smaller range of values:

- percentage of unemployment in a city - ranges from 0.0 to 1.0
- population of the city - can range up to 500,000

Scaling transforms the data so that the features have, more or less, uniform range.

Min-max Scaling

Scales values to a range of [0, 1].

Min-max Scaling

Computing the norm of feature vector X:

$$z_i = \frac{x_i - x_{min}}{x_{max} - x_{min}}$$

Example: For feature *age:*

$$z_1 = \frac{22 - 22}{42 - 22} = 0$$

ID	Age	Age_{scaled}
1	22	0.00
2	25	
3	30	
4	42	

Min-max Scaling

Computing the norm of feature vector X:

$$z_i = \frac{x_i - x_{min}}{x_{max} - x_{min}}$$

Example: For feature *age*:

$$z_2 = \frac{25 - 22}{42 - 22} = 0.15$$

ID	Age	Age_{scaled}
1	22	0.00
2	25	0.15
3	30	
4	42	

Min-max Scaling

Computing the norm of feature vector X:

$$z_i = \frac{x_i - x_{min}}{x_{max} - x_{min}}$$

Example: For feature *age*:

$$z_3 = \frac{30 - 22}{42 - 22} = 0.4$$

ID	Age	Age_{scaled}
1	22	0.00
2	25	0.15
3	30	0.40
4	42	

Min-max Scaling

Computing the norm of feature vector X:

$$z_i = \frac{x_i - x_{min}}{x_{max} - x_{min}}$$

Example: For feature *age*:

$$z_4 = \frac{42 - 22}{42 - 22} = 1$$

ID	Age	Age_{scaled}
1	22	0.00
2	25	0.15
3	30	0.40
4	42	1.00

FEATURE SCALING

- Standardization
- Min-max Scaling
- Normalization
- Binarization

Task: Read the Feature Scaling Tutorial

Partner/Group/Individual Exercise:

WINE DATA CLASSIFICATION

Note: Python beginners can partner up with more advanced users for better guidance

Partner/Group/Individual Presentation

Assignment

Binarize features in the Handwritten Digit Recognition Exercise

References:

WWCodeLondon Slides

https://kevinzakka.github.io/2016/07/13/k-nearest-neighbor/

http://scikit-learn.org/stable/modules/preprocessing.html

http://sebastianraschka.com/Articles/2014_about_feature_scaling.html

T.I.L.

SHARE IT!
In front!

On Twitter: @wwcodemanila
Or FB: fb.com/wwcodemanila

Don’t forget to tag WWCodeManila so we can retweet or share it.

THANK YOU :)