COMP30810
Intro to Text Analytics

Dr. Binh Thanh Le

thanhbinh.le@ucd.ie

Insight Centre for Data Analytics

School of Computer Science

University College Dublin

Today goals

- Understand the KNN classifier
- Apply KNN to text

What is KNN?

- KNN is the short name of K-Nearest-Neighbour classifier.

- It is very simple, easy to understand, versatile and one of the topmost Supervised Learning algorithms.

- **KNN** is a **non-parametric** learning algorithm.

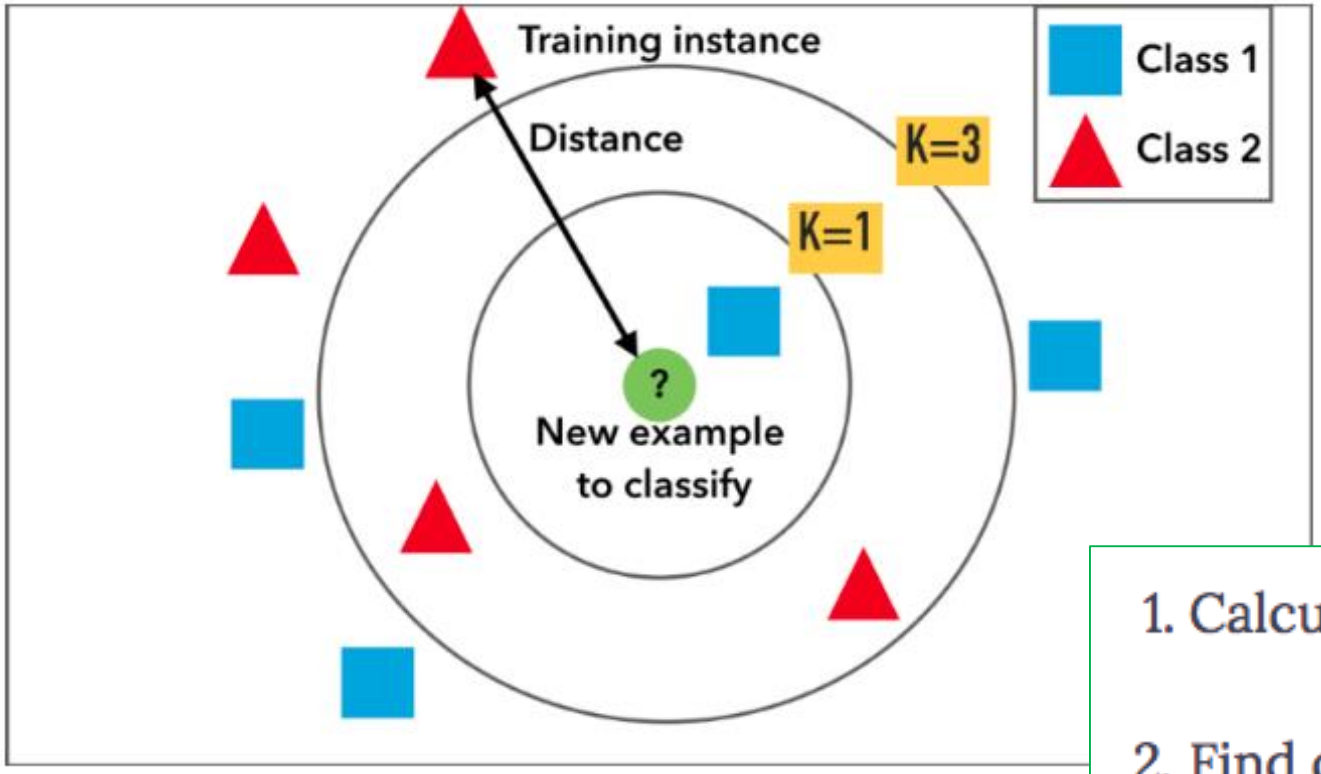
  ➔ It does not make any *assumptions* on the underlying data distribution.

- **KNN** classifier is also an **instance-based** learning algorithm

➔ does not use the training data points to do any *generalization.*

Minimal training but expensive testing.

The KNN algorithm



1. Calculate distance

2. Find closest neighbors

3. Vote for labels

Training data

sepal-length	sepal-width	petal-length	petal-width	Class
6.1	2.8	4.7	1.2	Iris-versicolor
5.7	3.8	1.7	0.3	Iris-setosa
7.7	2.6	6.9	2.3	Iris-virginica
6.0	2.9	4.5	1.5	Iris-versicolor
6.8	2.8	4.8	1.4	Iris-versicolor
5.4	3.4	1.5	0.4	Iris-setosa
5.6	2.9	3.6	1.3	Iris-versicolor
6.9	3.1	5.1	2.3	Iris-virginica
6.2	2.2	4.5	1.5	Iris-versicolor
5.8	2.7	3.9	1.2	Iris-versicolor
6.5	3.2	5.1	2.0	Iris-virginica
4.8	3.0	1.4	0.1	Iris-setosa

Test data

sepal-length	sepal-width	petal-length	petal-width
5.1	3.5	1.4	0.3

?

Iris Setosa

Iris versicolor

Iris virginica

Training data

sepal-length	sepal-width	petal-length	petal-width	Class
6.1	2.8	4.7	1.2	Iris-versicolor
5.7	3.8	1.7	0.3	Iris-setosa
7.7	2.6	6.9	2.3	Iris-virginica
6.0	2.9	4.5	1.5	Iris-versicolor
6.8	2.8	4.8	1.4	Iris-versicolor
5.4	3.4	1.5	0.4	Iris-setosa
5.6	2.9	3.6	1.3	Iris-versicolor
6.9	3.1	5.1	2.3	Iris-virginica
6.2	2.2	4.5	1.5	Iris-versicolor
5.8	2.7	3.9	1.2	Iris-versicolor
6.5	3.2	5.1	2.0	Iris-virginica
4.8	3.0	1.4	0.1	Iris-setosa

Test data

sepal-length	sepal-width	petal-length	petal-width
5.1	3.5	1.4	0.3

Euclidean distances

KNN with K = **1**

```
[3.63180396]
[0.73484692]
[6.46683849]
[3.49571166]
[4.01870626]
[0.34641016]
[2.53968502]
[4.592385  ]
[3.73496988]
[2.8618176 ]
[4.31624837]
[0.6164414 ]
```

Iris-setosa

Training data

sepal-length	sepal-width	petal-length	petal-width	Class
6.1	2.8	4.7	1.2	Iris-versicolor
5.7	3.8	1.7	0.3	Iris-setosa
7.7	2.6	6.9	2.3	Iris-virginica
6.0	2.9	4.5	1.5	Iris-versicolor
6.8	2.8	4.8	1.4	Iris-versicolor
5.4	3.4	1.5	0.4	Iris-setosa
5.6	2.9	3.6	1.3	Iris-versicolor
6.9	3.1	5.1	2.3	Iris-virginica
6.2	2.2	4.5	1.5	Iris-versicolor
5.8	2.7	3.9	1.2	Iris-versicolor
6.5	3.2	5.1	2.0	Iris-virginica
4.8	3.0	1.4	0.1	Iris-setosa

Test data

sepal-length	sepal-width	petal-length	petal-width
5.1	3.5	1.4	0.3

Euclidean distances

KNN with K = **3**

[3.63180396]
[0.73484692]
[6.46683849]
[3.49571166]
[4.01870626]
[0.34641016]
[2.53968502]
[4.592385]
[3.73496988]
[2.8618176]
[4.31624837]
[0.6164414]

Iris-setosa

Example for Text Analytics

Training data	Training class
1. Cricket is a bat and ball game played between two team 1. Each phase of play is called an innings during which o 1. The teams have one or two innings apiece and, when the 1. Before a match begins, the two team captains meet on t 1. Two batsmen and eleven fielders then enter the field a 1. The most common dismissal in cricket match are bowled, 1. Runs are scored by two main methods: either by hitting 1. The main objective of each team is to score more runs 1. If the team batting last is all out having scored fewe 1. The role of striker batsman is to prevent the ball fro	Cricket
2. Artificial intelligence is intelligence exhibited by m 2. the field of AI research defines itself as the study o 2. The overall research goal of artificial intelligence i 2. Natural language processing[77] gives machines the abi 2. AI researchers developed sophisticated mathematical to 2. An intelligent agent is a system that perceives its en 2. AI techniques have become an essential part of the tec 2. Recent advancements in AI, and specifically in machine 2. AI research was revived by the commercial success of e 2. Advanced statistical techniques (loosely known as deep	Artificial Intelligence
3. A compound is a pure chemical substance composed of mo 3. Since the properties of an element are mostly determin 3. The property of inertness of noble gases makes them ve 3. The atom is also the smallest entity that can be envis 3. The nucleus is made up of positively charged protons a 3. The atom is the basic unit of chemistry. It consists o 3. A chemical reaction is a transformation of some substa 3. Chemistry is sometimes called the central science beca 3. Chemistry includes topics such as the properties of in 3. Chemistry is a branch of physical science that studies	Chemistry

Test data	
	Chemical compunds are used for preparing bombs based on some reactions ?
	Cricket is a boring game where the batsman only enjoys the game ?
	Machine learning is a area of Artificial intelligence ?

Example for Text Analytics

Texts

==> Stop words removal

==> Punctuation removal

==> Word Lemmatization

==> Digit removal

==> Feature Extraction (Tf-Idf)

X

ability	access	action	advance	advanced	advancement	agent	ai	allows	analytical	...	topic	toss	transformation
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.000000	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.000000	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.000000	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.320343	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.000000	0.0

```
1  y_train
```

executed in 5ms, finished 14:19:39 2018-10-30

```
array([0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 1., 1., 1., 1., 1., 1., 1.,
       1., 1., 1., 2., 2., 2., 2., 2., 2., 2., 2., 2., 2.])
```

==> Model training

```
1  from sklearn.neighbors import KNeighborsClassifier
2  modelknn = KNeighborsClassifier(n_neighbors=5)
3  modelknn.fit(X,y_train)
```

```
# Predicting it on test data : Testing Phase
test_sentences = ["Chemical compunds are used for preparing bombs based on some reactions",\
"Cricket is a boring game where the batsman only enjoys the game",\
"Machine learning is a area of Artificial intelligence"]

test_clean_sentence = []
for test in test_sentences:
    cleaned_test = clean(test)
    cleaned = ' '.join(cleaned_test)
    cleaned = re.sub(r"\d+","",cleaned)
    test_clean_sentence.append(cleaned)

Test = vectorizer.transform(test_clean_sentence)

true_test_labels = ['Cricket','AI','Chemistry']
predicted_labels_knn = modelknn.predict(Test)

print("\nBelow 3 sentences of test data:\n1. ",\
test_sentences[0],"\n2. ",test_sentences[1],"\n3. ",test_sentences[2])
print("\n------------------------------PREDICTIONS BY KNN----------------------------------------")
print("\n",test_sentences[0],":",true_test_labels[np.int(predicted_labels_knn[0])],\
"\n",test_sentences[1],":",true_test_labels[np.int(predicted_labels_knn[1])],\
"\n",test_sentences[2],":",true_test_labels[np.int(predicted_labels_knn[2])])
```

```
Below 3 sentences of test data:
1.  Chemical compunds are used for preparing bombs based on some reactions
2.  Cricket is a boring game where the batsman only enjoys the game
3.  Machine learning is a area of Artificial intelligence

------------------------------PREDICTIONS BY KNN----------------------------------------

 Chemical compunds are used for preparing bombs based on some reactions : Chemistry
 Cricket is a boring game where the batsman only enjoys the game : Cricket
 Machine learning is a area of Artificial intelligence : AI
```

Pros and Cons

- **Pros**
  - No assumptions about data — useful, for example, for nonlinear data
  - Simple algorithm — to explain and understand/interpret
  - High accuracy (relatively) — it is pretty high but not competitive in comparison to better supervised learning models
  - Versatile — useful for classification or regression
- **Cons**
  - Computationally expensive
  - High memory requirement
  - Stores all (or almost all) of the training data
  - Prediction stage might be slow
  - Sensitive to irrelevant features and the scale of the data