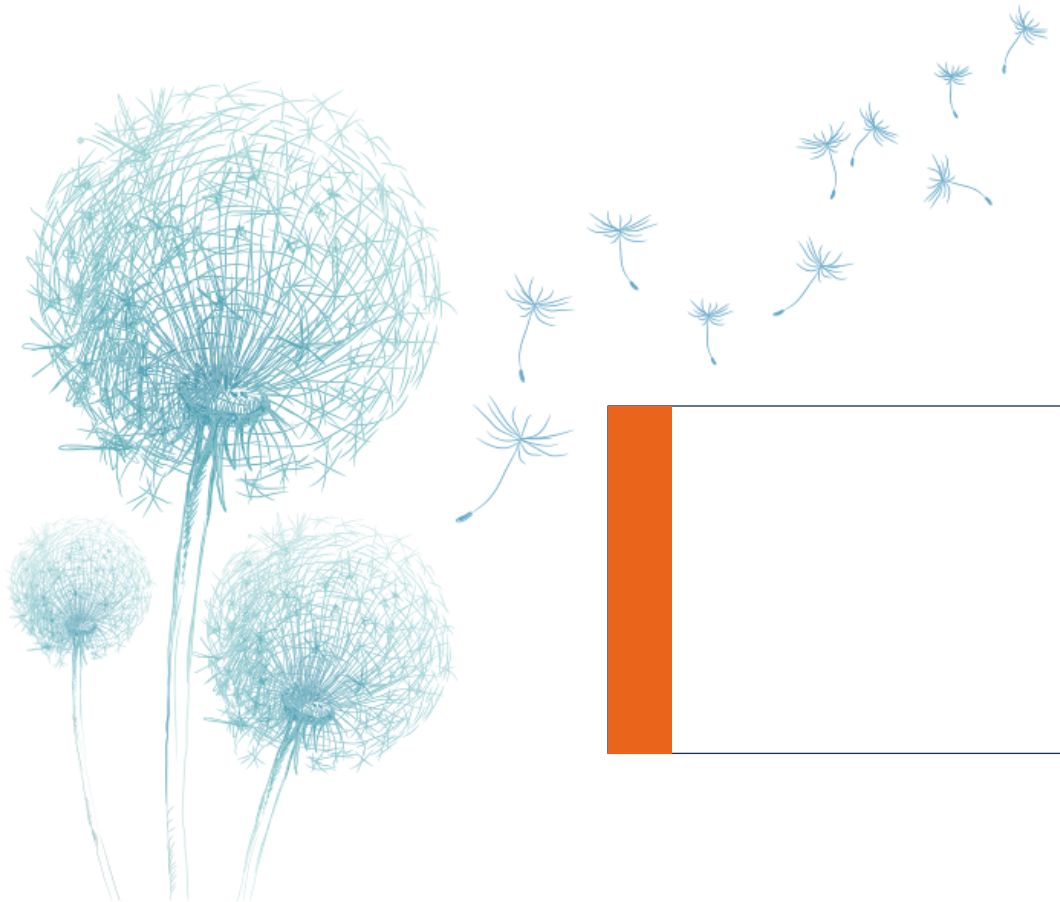
Neural networks and deep learning



KNN

Kun Suo

Computer Science, Kennesaw State University

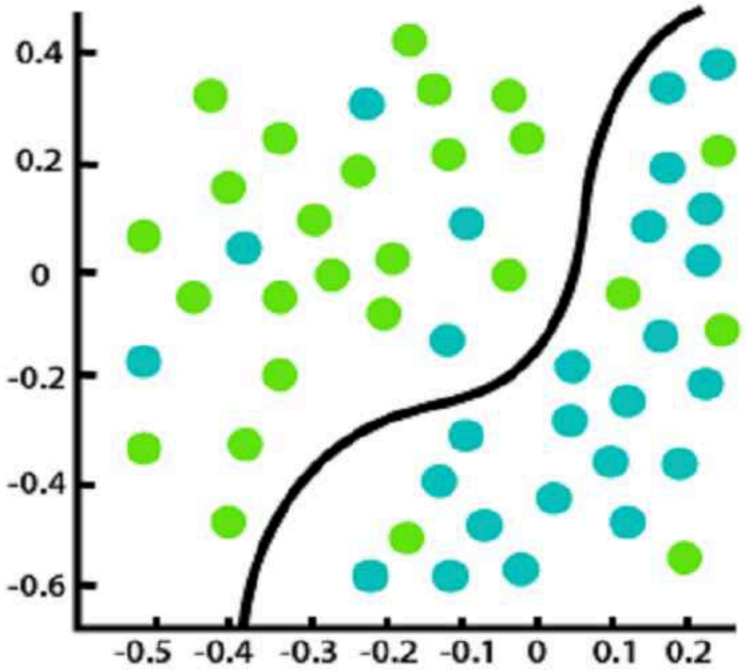
https://kevinsuo.github.io/

Classification

Classification

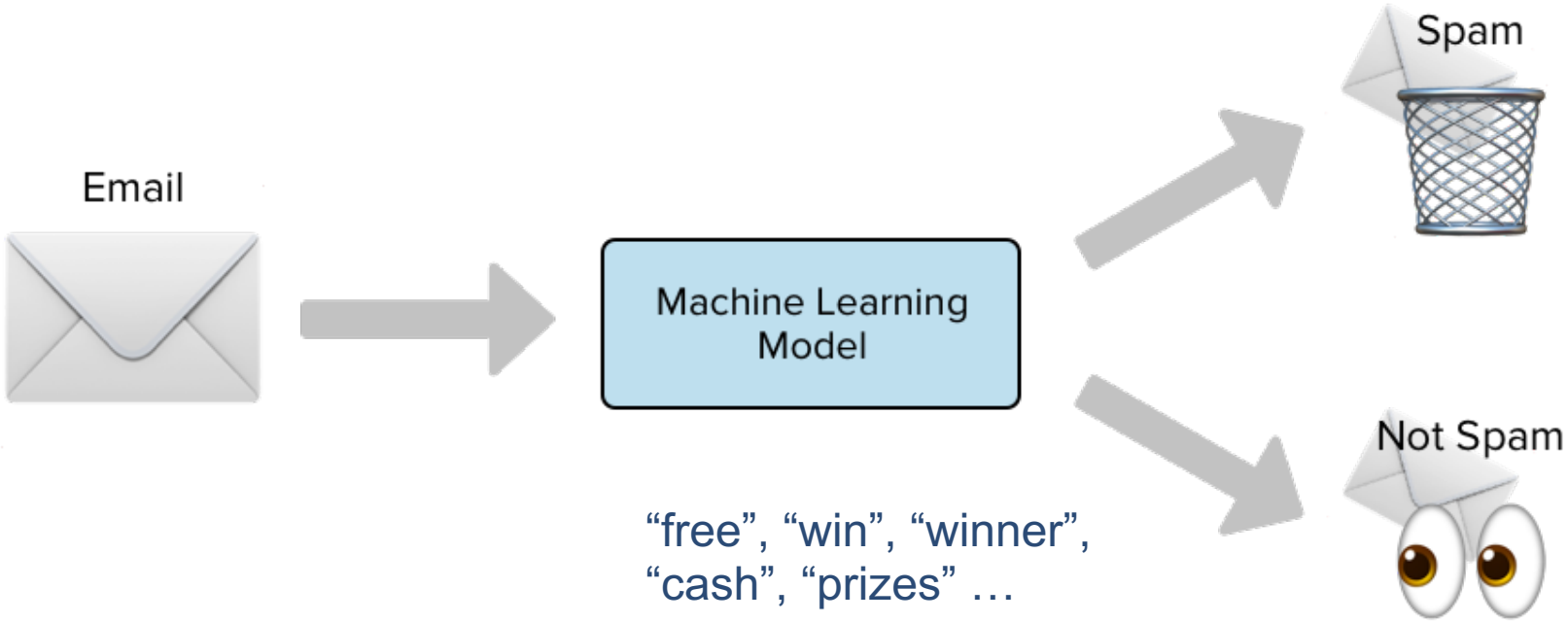
- Classification is a technique for determining which class the dependent belongs to based on one or more independent variables.

- Classification problems are when we are training a model to predict qualitative targets. For example: a type of fruit.



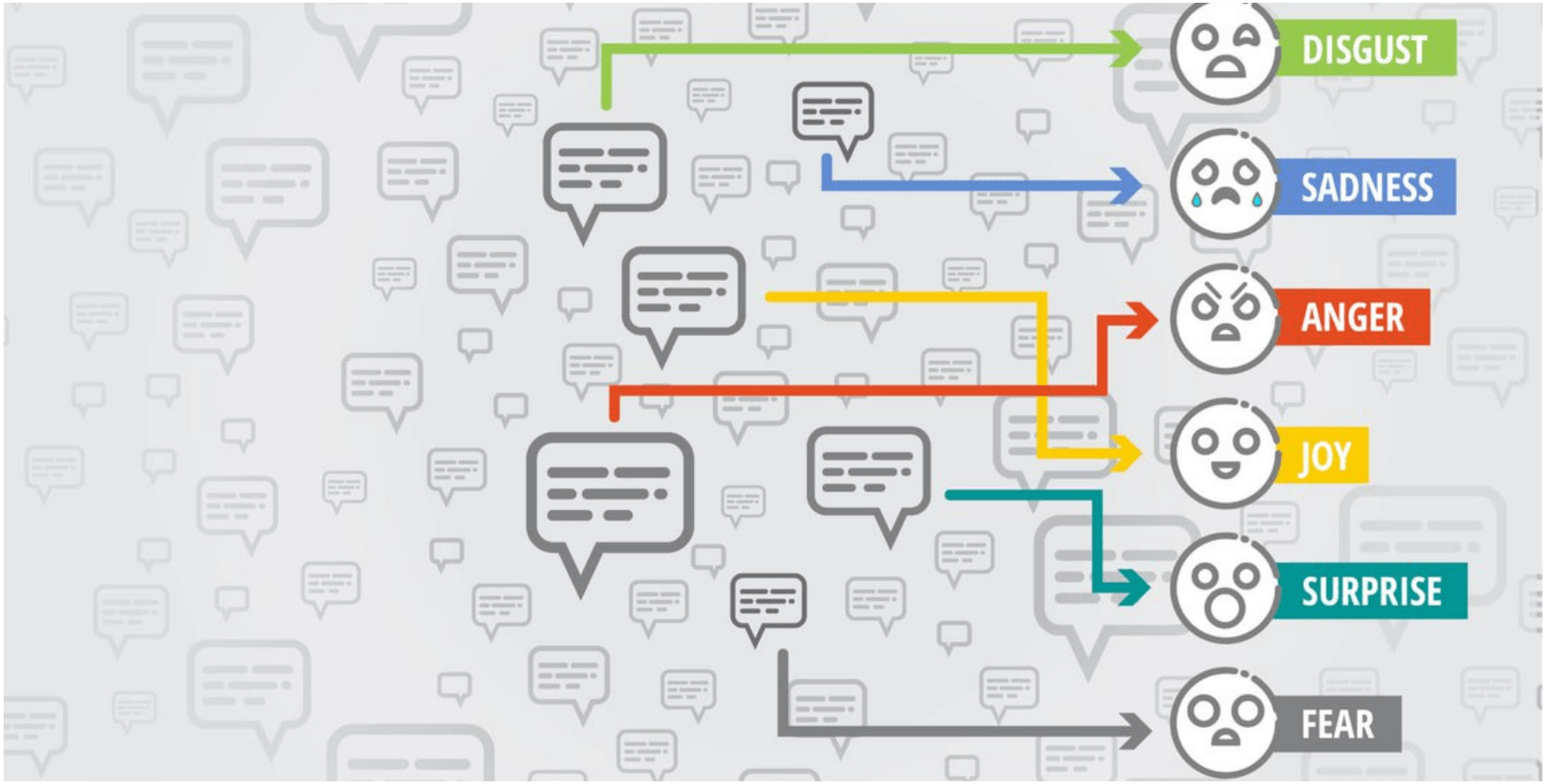
Classification examples

- Spam email detection



Classification examples

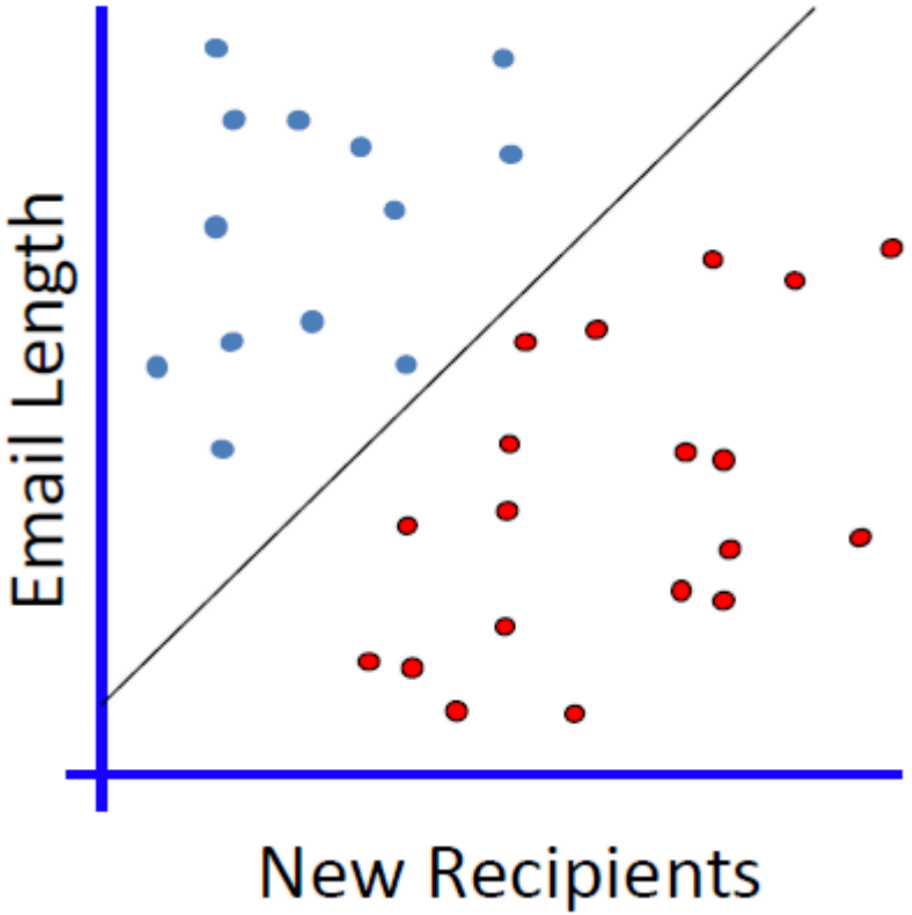
- Emotion analysis



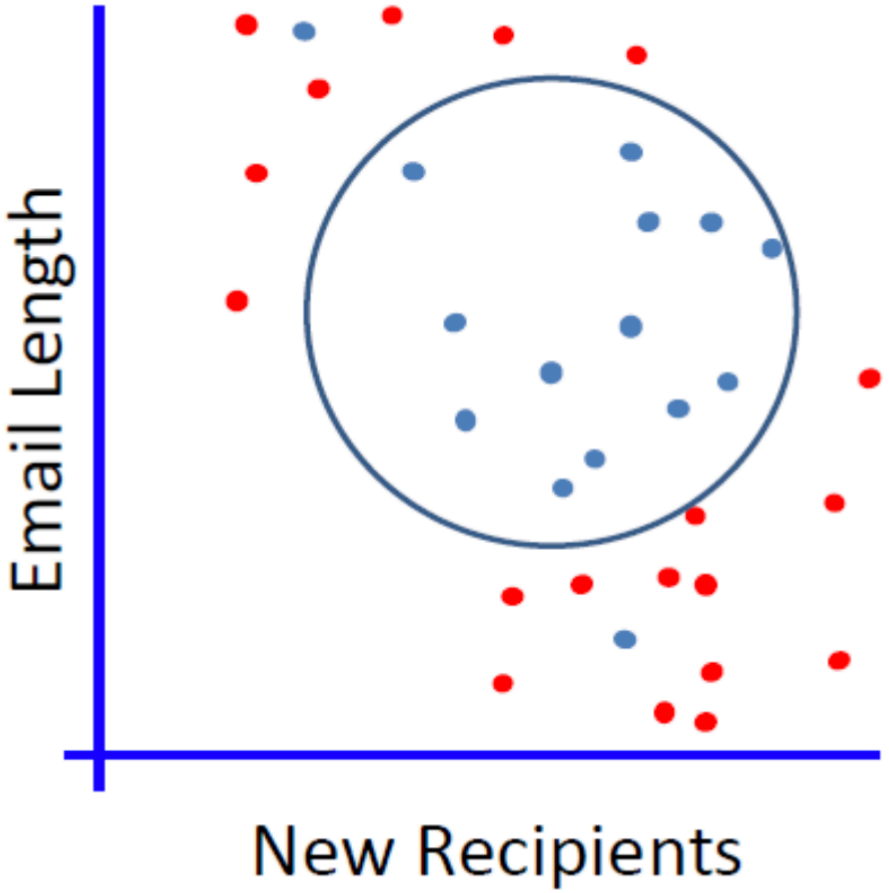
Classification examples

- Dog breed detection

Linear Classification



Non-Linear Classification



Common classification methods

- Naive Bayes (NB)
- Logistic Regression (LR)
- Decision Tree (DT)
- Support Vector Machine (SVM)
- K Nearest Neighbors (KNN)
- …

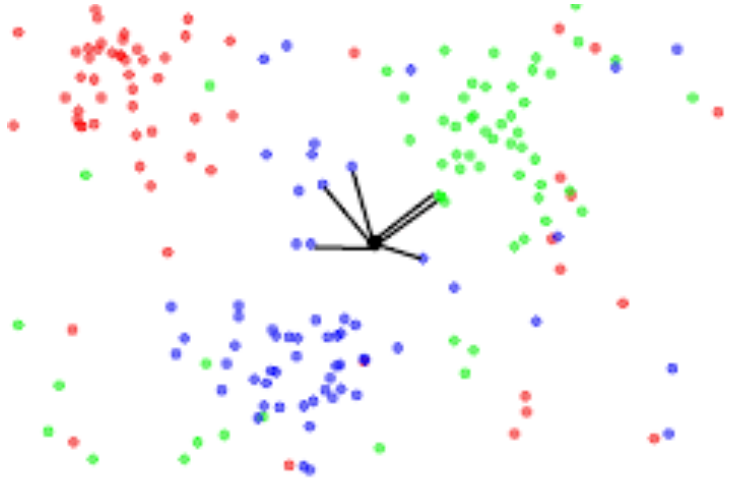
K-Nearest Neighbor Classification

K-Nearest Neighbor Classification

- Definition: An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors

- Input: Its k nearest neighbors

- Output: is a class



K-Nearest Neighbor Classification

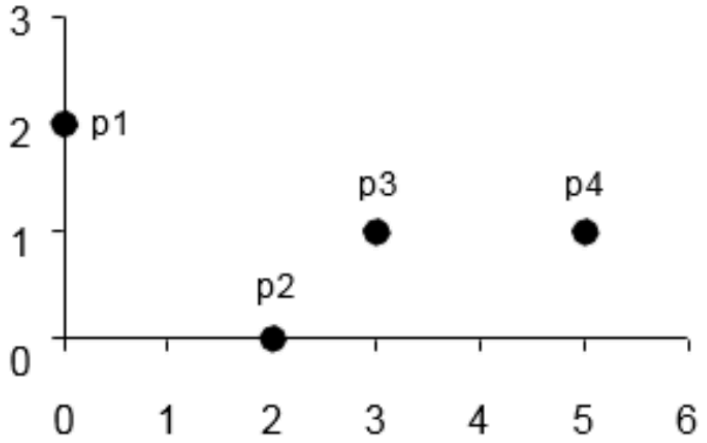
- Euclidean Distance

$$dist = \sqrt{\sum_{k=1}^{p}(a_k - b_k)^2}$$

Where p is the number of dimensions (attributes), a_k and b_k are, respectively, the k-th attributes (components) or data of object a and b

K-Nearest Neighbor Classification

- Euclidean Distance

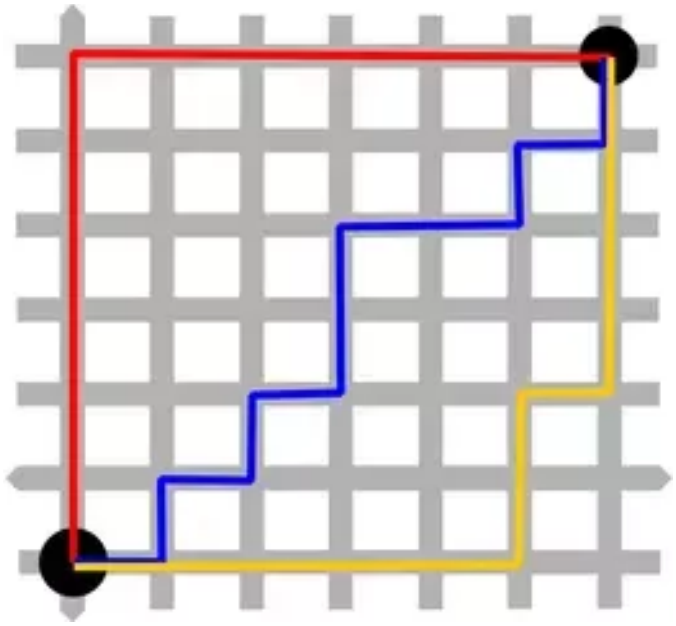


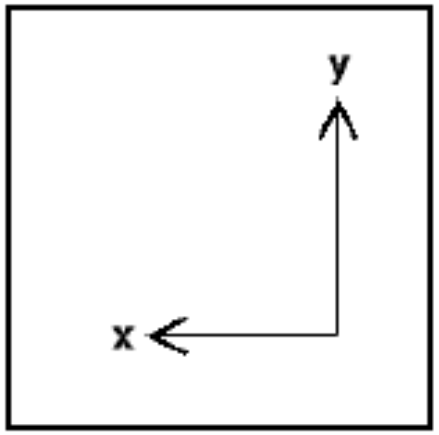
point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

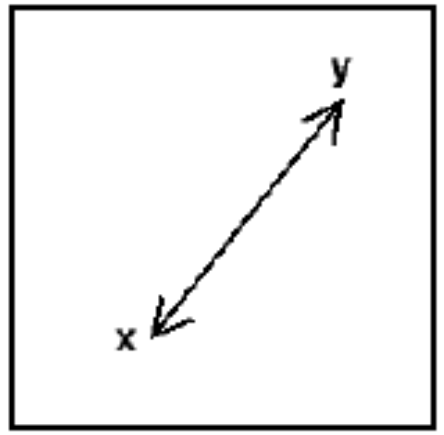
K-Nearest Neighbor Classification

- Manhattan Distance
  - In a plane with p1 at (x1, y1) and p2 at (x2, y2), it is |x1 - x2| + |y1 - y2|





Manhattan

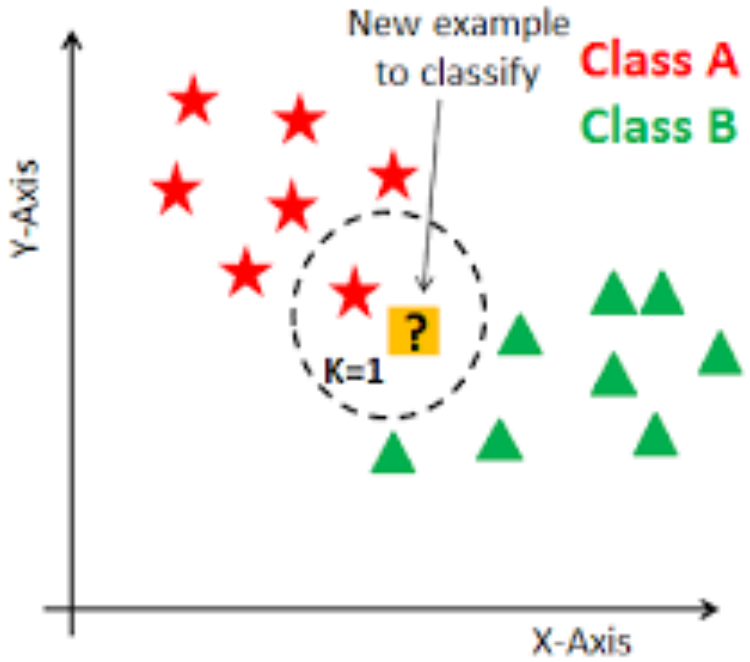


Euclidean

K-Nearest Neighbor Classification

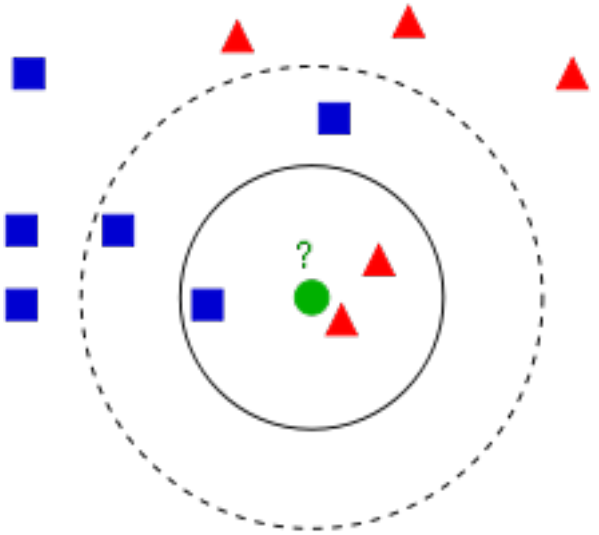
- If k = 1, then the object is simply assigned to the class of that single nearest neighbor

- K should be an odd number to avoid equal votes



Example: KNN

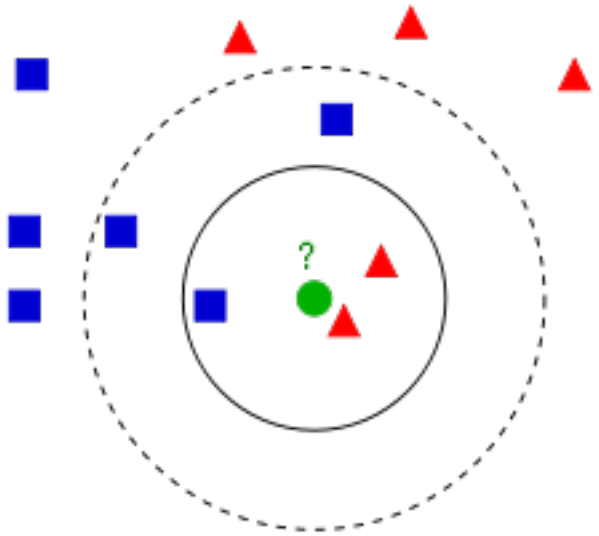
- There are two types of sample data in the figure:
  - blue square
  - red triangle
  - The green circle is the data we want to classify



Example: KNN

- If K=3, which class the green point is classified to?

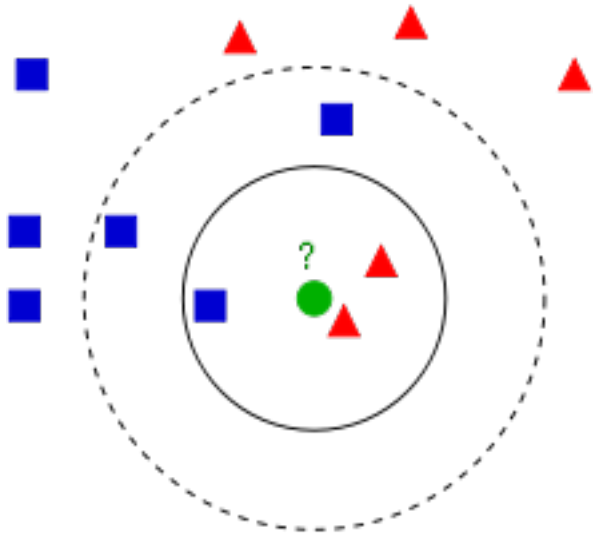
  - There are 2 red triangles and 1 blue square closest to the green point. Based on these 3 points vote, so the green point to be classified belongs to the red triangle.



Example: KNN

- If K=5, which class the green point is classified to?

  - There are 2 red triangles and 3 blue square closest to the green point. Based on these 5 points vote, so the green point to be classified belongs to the blue square.

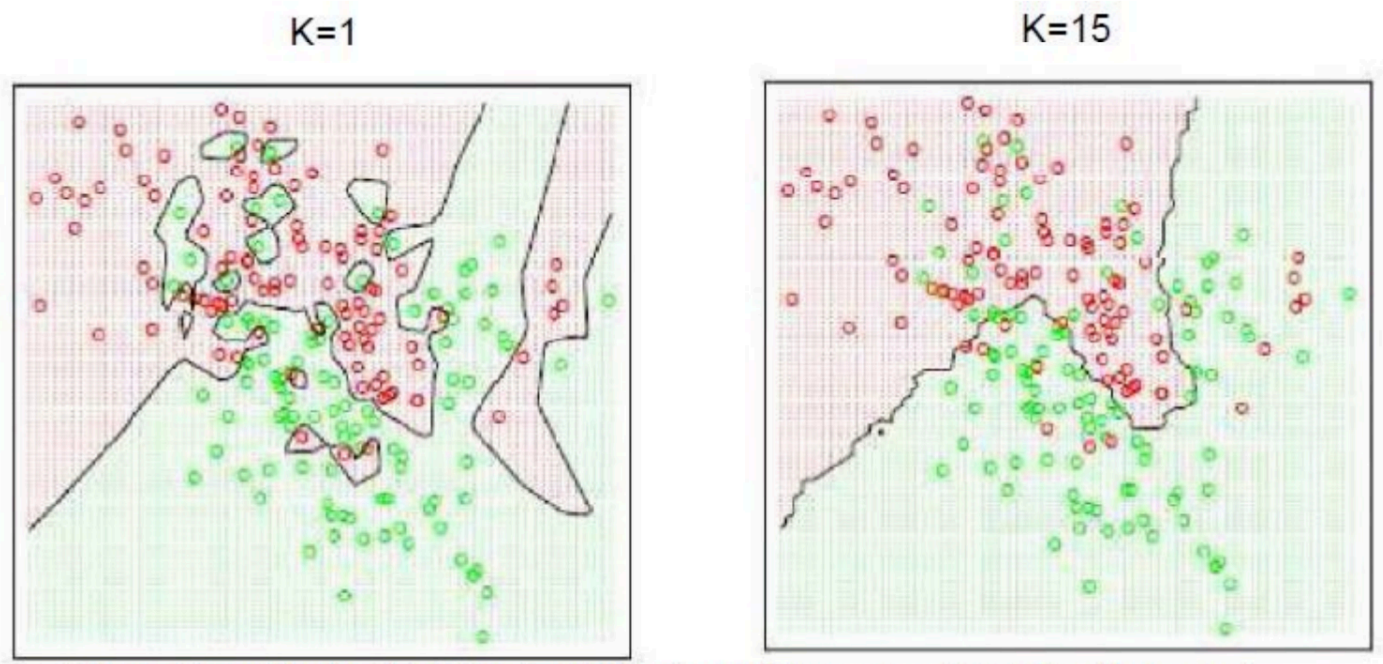


KNN: Pseudo code

- Step 1: Determine parameter K = number of nearest neighbors
- Step 2: Calculate the distance between the query-instance and all the training examples.
- Step 3: Sort the distance and determine nearest neighbors based on the k-th minimum distance.
- Step 4:Gather the category Y of the nearest neighbors.
- Step 5: Use simple majority of the category of nearest neighbors as the prediction value of the query instance.

Effect of K

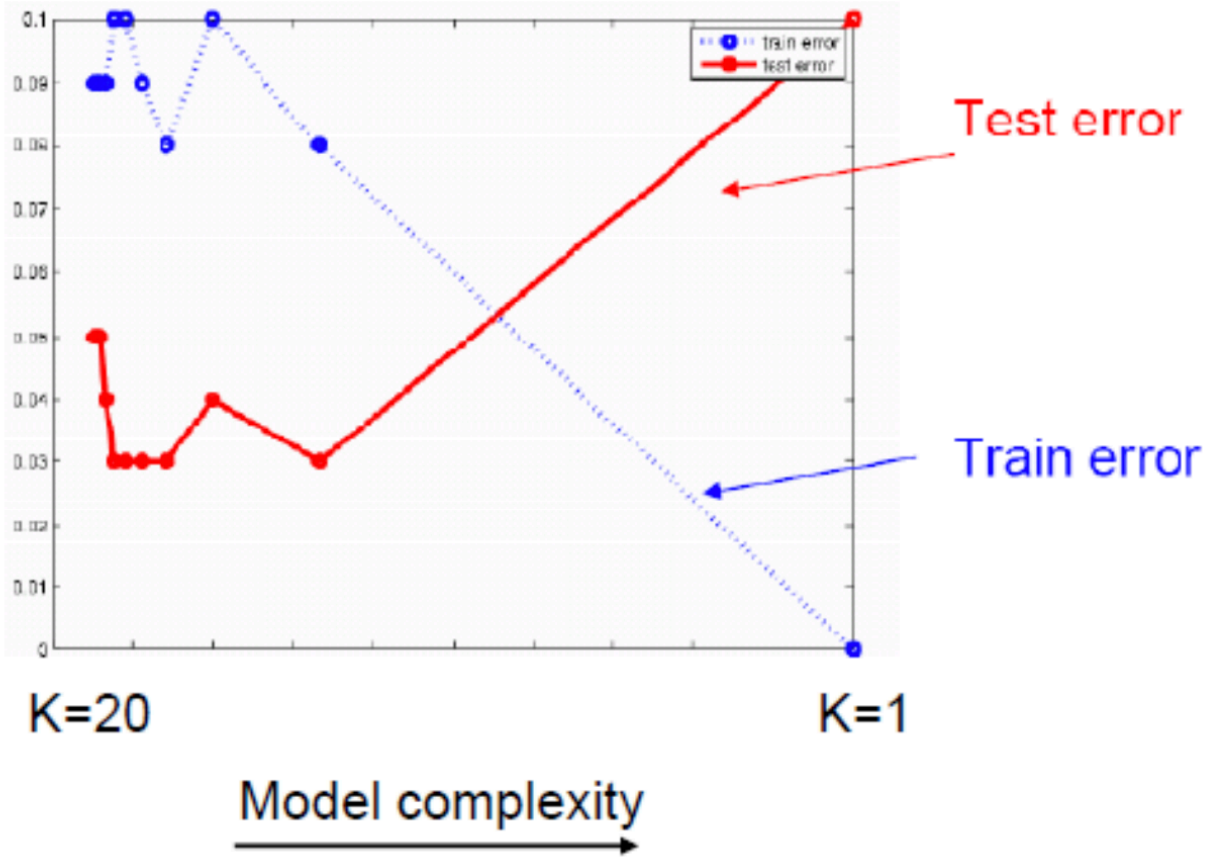
- Larger k produces smoother boundary effect
- When K==N, always predict the majority class



Figures from Hastie, Tibshirani and Friedman (Elements of Statistical Learning)

How to choose k?

- Empirically optimal k?



KNN Practice

Example: Product quality judgment

- Suppose we need to judge the quality of paper towels

The quality of paper towels can be drawn as two attributes,

1, the acid durability

2, the strength that can withstand

Example: Product quality judgment

- Suppose we need to judge the quality of paper towels

(x_1, x_2)

Example: Product quality judgment

- Training Data

X1 acid durability (s)	X2 strength (kg/square meter)	Quality Y
7	7	Bad
7	4	Bad
3	4	Good
1	4	Good

Example: Product quality judgment

X1 acid durability (s)	X2 strength (kg/square meter)	Quality Y
7	7	Bad
7	4	Bad
3	4	Good
1	4	Good

$(x_1, x_2) = (3, 7)$

What is the quality of this paper tissue if K = 3?

Example: Product quality judgment

X1 acid durability (s)	X2 strength (kg/square meter)	Quality Y	Distance
7	7	Bad	$(7-3)^2+(7-7)^2=16$
7	4	Bad	$(7-3)^2+(4-7)^2=25$
3	4	Good	$(3-3)^2+(4-7)^2=9$
1	4	Good	$(1-3)^2+(4-7)^2=13$

$(x_1, x_2) = (3, 7)$

Example: Product quality judgment

X1 acid durability (s)	X2 strength (kg/square meter)	Quality Y	Distance
7	7	Bad	$(7-3)^2+(7-7)^2=16$
7	4	Bad	$(7-3)^2+(4-7)^2=25$
3	4	Good	$(3-3)^2+(4-7)^2=9$
1	4	Good	$(1-3)^2+(4-7)^2=13$

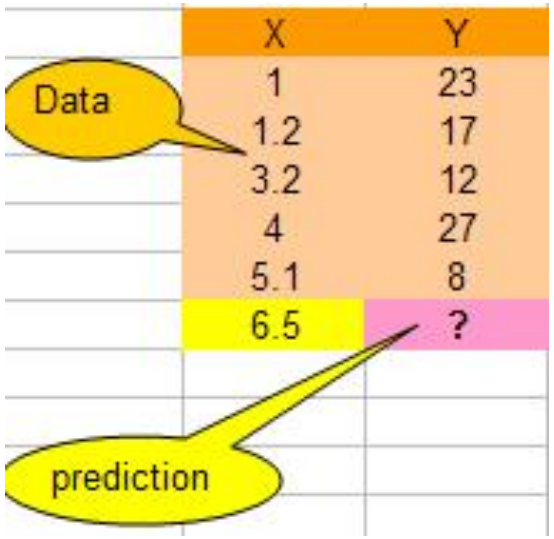
$(x_1, x_2) = (3, 7)$

K = 3

Good products have 2 votes, bad ones have 1 vote, and the final test (3, 7) is a qualified product

Example: KNN for prediction

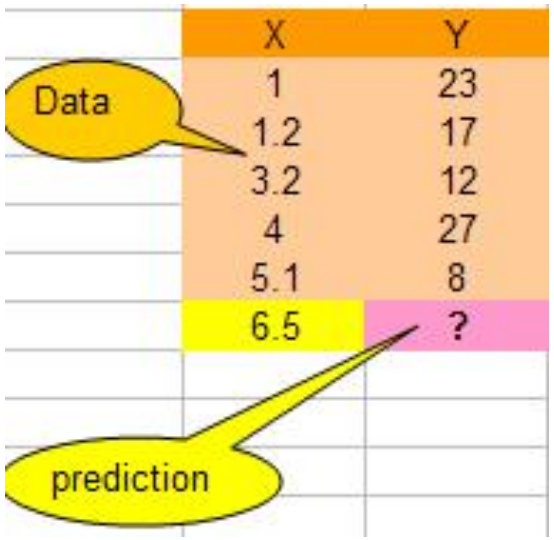
- Suppose we have the following set of data, assuming that X is the number of seconds that have elapsed, and the value of Y is a value that changes over time (you can think of it as stock value)



When the time is 6.5 seconds, what is the Y value if K =2 ?

Example: KNN for prediction

- Suppose we have the following set of data, assuming that X is the number of seconds that have elapsed, and the value of Y is a value that changes over time (you can think of it as stock value)

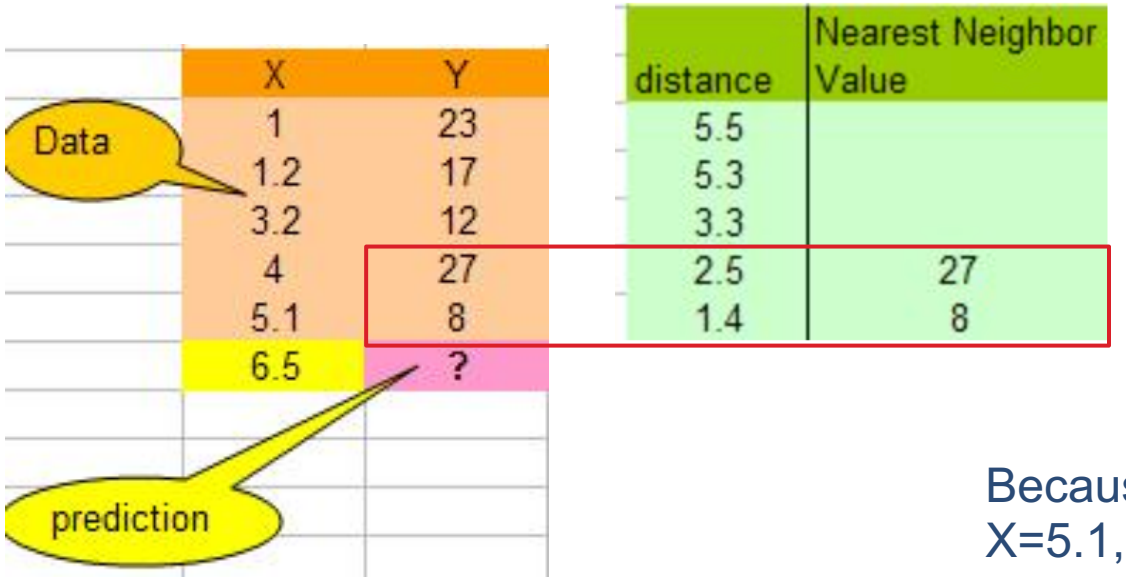


distance	Nearest Neighbor Value
5.5	
5.3	
3.3	
2.5	27
1.4	8

Calculate the distance from all X points to 6.5, such as:
X=5.1, the distance is | 6.5 – 5.1 | = 1.4,
X=1.2, then the distance is | 6.5 – 1.2 | = 5.3

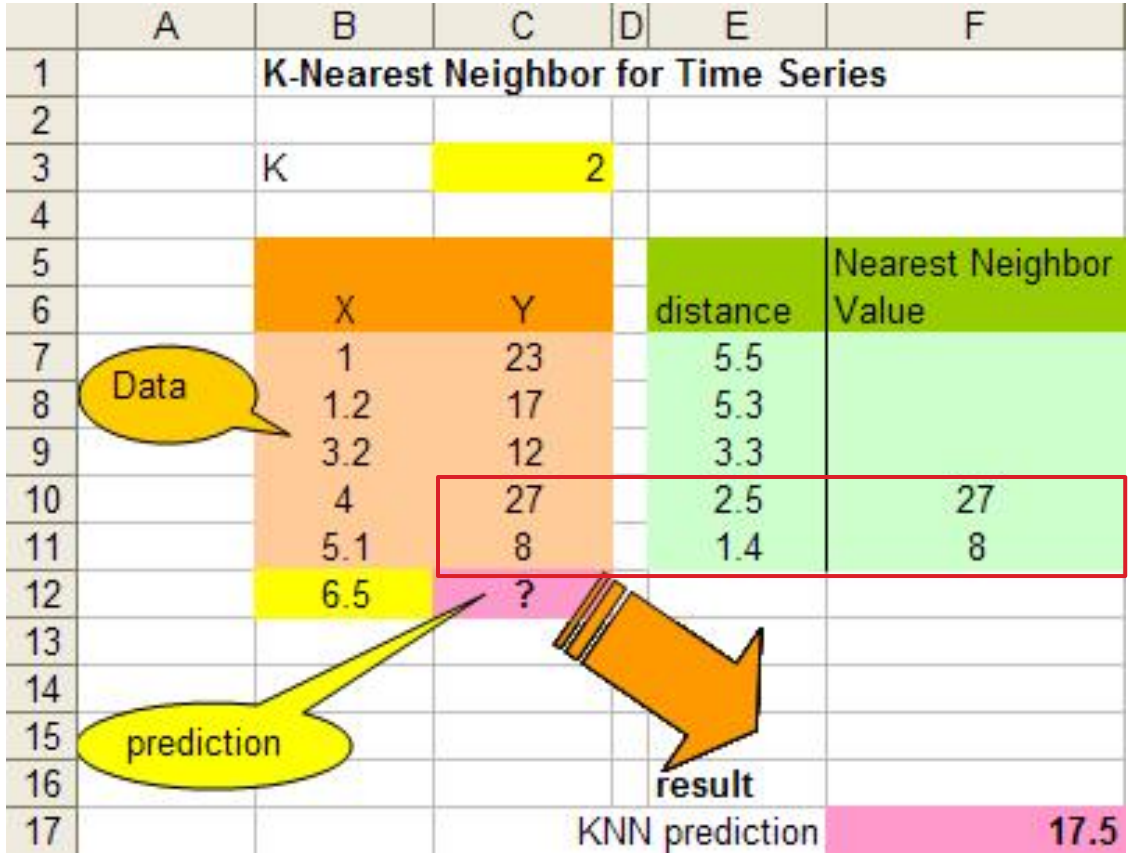
Example: KNN for prediction

- Suppose we have the following set of data, assuming that X is the number of seconds that have elapsed, and the value of Y is a value that changes over time (you can think of it as stock value)



Because K=2, we get the closest points with X=4 and X=5.1, and the values of Y are 27 and 8 respectively

Example: KNN for prediction

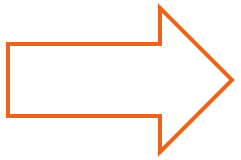


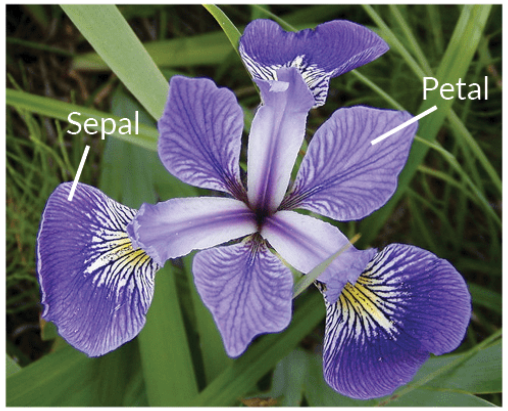
Simply use the average to calculate:

$$\frac{27+8}{2}=17.5$$

KNN Practice

- Predict the type of a given iris





Iris Versicolor

Iris Setosa

Iris Virginica

https://github.com/kevinsuo/CS7357/blob/master/knn/show-data.py

KNN Practice

- Install package scikit-learn
  - $ sudo pip install scikit-learn

- Include various necessary packages

```
import matplotlib.pyplot as plt
import numpy as np
from sklearn import neighbors
from sklearn import datasets
```

- Load data set

```
iris = datasets.load_iris()
X = iris.data[:, 2:4]
y = iris.target
```

//Take only the last two dimensions in the feature space

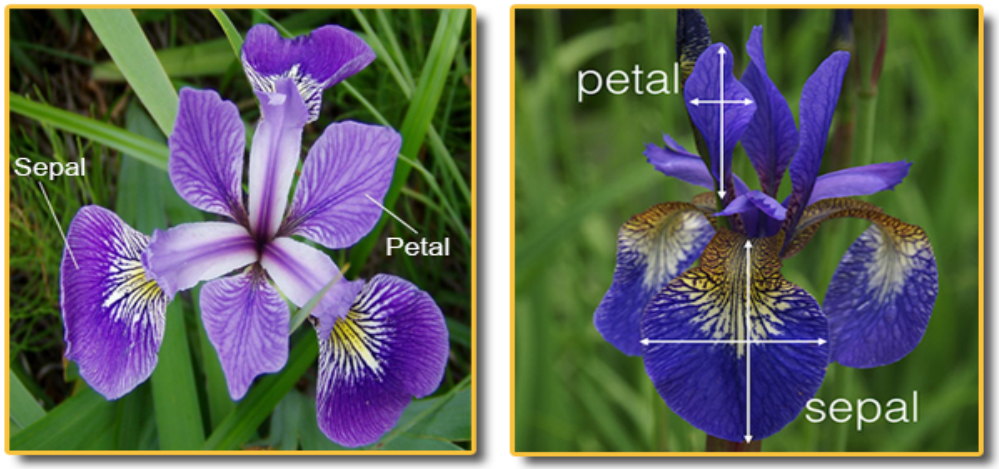
Iris flower data set

- https://archive.ics.uci.edu/ml/datasets/iris

- The data set contains 150 data samples, divided into 3 categories, each with 50 data

- 3 categories: Setosa, Versicolour, Virginica

- Each data contains 4 attributes: sepal length, sepal width, petal length, petal width



KNN Practice

- Show the distribution of data

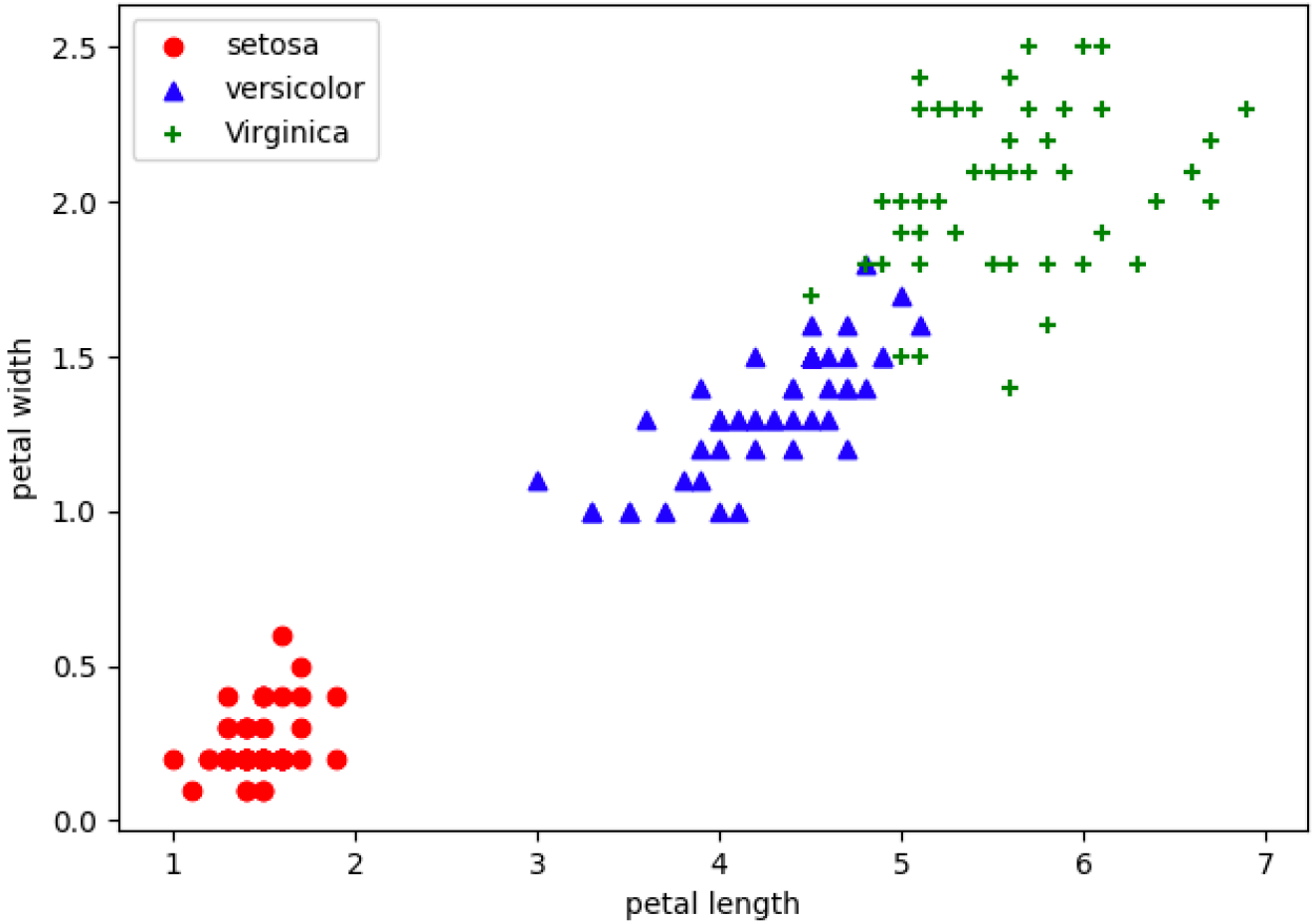
```
color = ("red", "blue", "green")
plt.scatter(X[:50, 0], X[:50, 1], c = color[0], marker='o', label='setosa')
plt.scatter(X[50:100, 0], X[50:100, 1], c = color[1], marker='^', label='versicolor')
plt.scatter(X[100:, 0], X[100:, 1], c = color[2], marker='+', label='Virginica')

plt.xlabel('petal length')
plt.ylabel('petal width')
plt.legend(loc=2)
plt.show()
```

KNN Practice

https://github.com/kevinsuo/CS7357/blob/master/knn/show-data.py

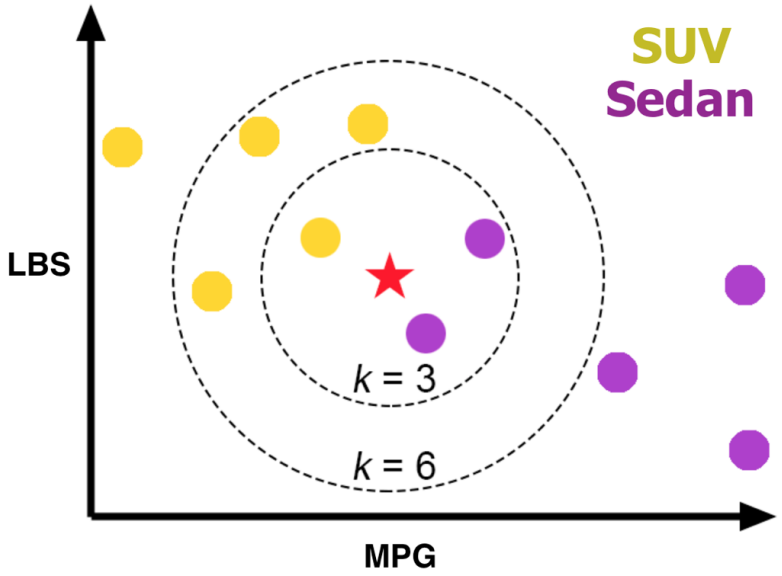
- Show the distribution of data



https://github.com/kevinsuo/CS7357/blob/master/knn/knn-example.py

KNN Practice

- Two different NN classifiers are provided in scikit-learn: KNeighborsClassifier and RadiusNeighborsClassifier
  - KNeighborsClassifier: will receive an integer k entered by the user, and then search for k nearest neighbors (and then vote), here k = 5 by default
  - RadiusNeighborsClassifier: A floating point number r will be required as the radius, and then points within this radius will be voted

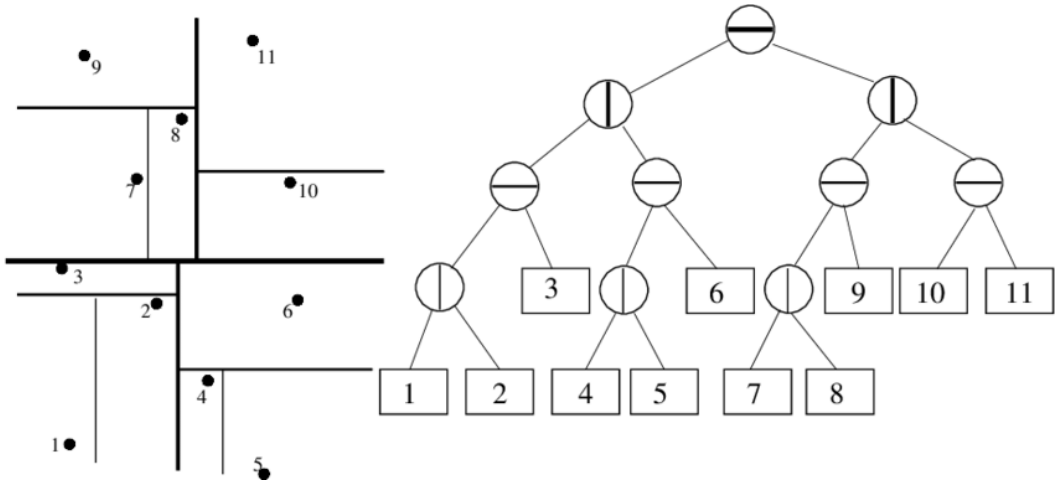


KNN Practice

- Use KNeighborsClassifier to build a kNN model for the iris dataset. Store the training instance in a Tree. The Tree structure is mainly for faster calculation of NN.

```
clf = neighbors.KNeighborsClassifier(n_neighbors = 5,
                      algorithm = 'kd_tree', weights='uniform')
clf.fit(X, y)
```

- About k-d tree
  - https://en.wikipedia.org/wiki/K-d_tree



KNN Practice

- Using the established model, predict which category the two new data points should be classified into

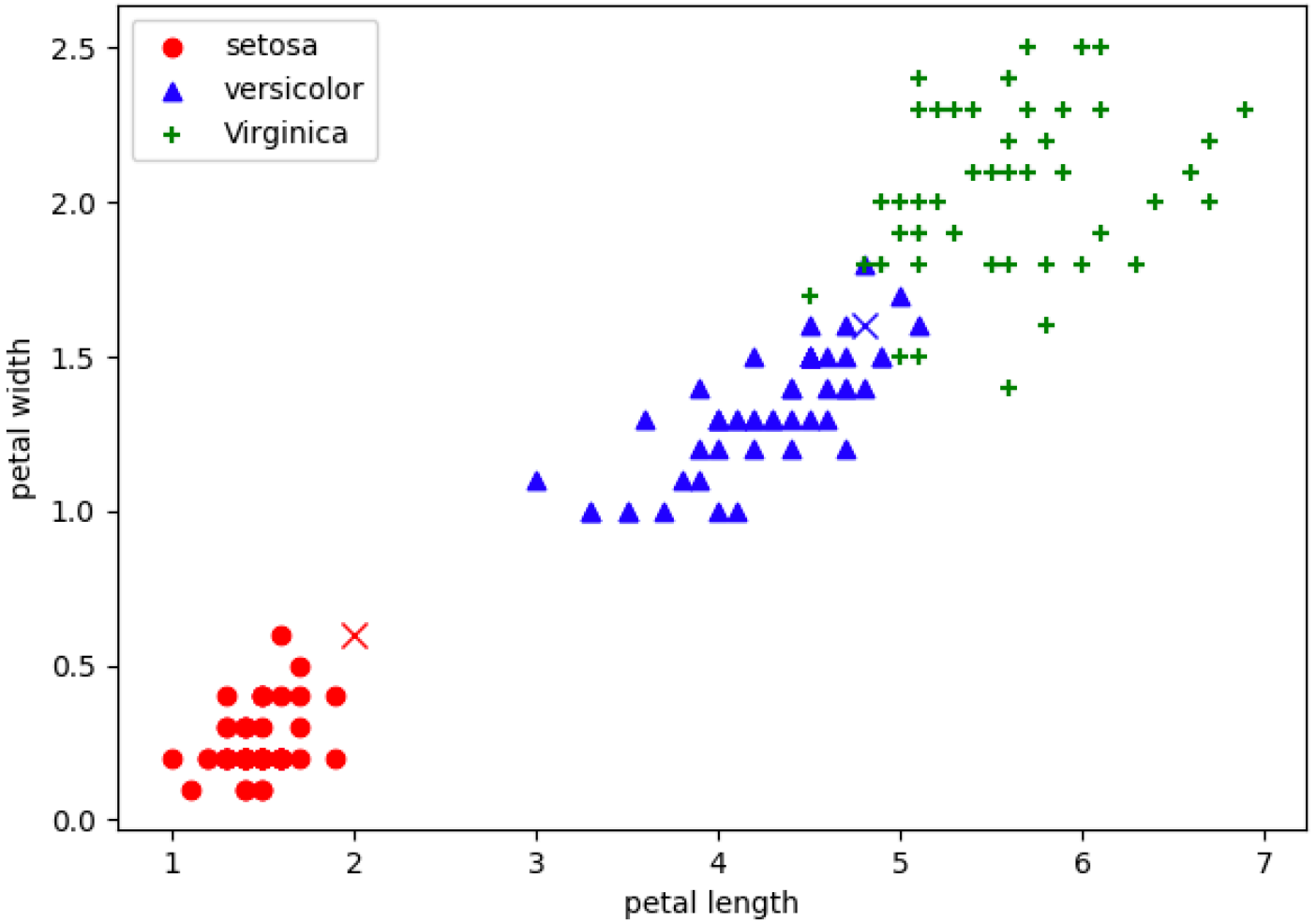
```
new_data = [[4.8, 1.6], [2, 0.6]]
result = clf.predict(new_data)
```

- Draw a figure

```
plt.plot(new_data[0][0], new_data[0][1], c = color[result[0]], marker='x', ms= 8)
plt.plot(new_data[1][0], new_data[1][1], c = color[result[1]], marker='x', ms= 8)
plt.show()
```

https://github.com/kevinsuo/CS7357/blob/master/knn/knn-example.py

KNN Practice



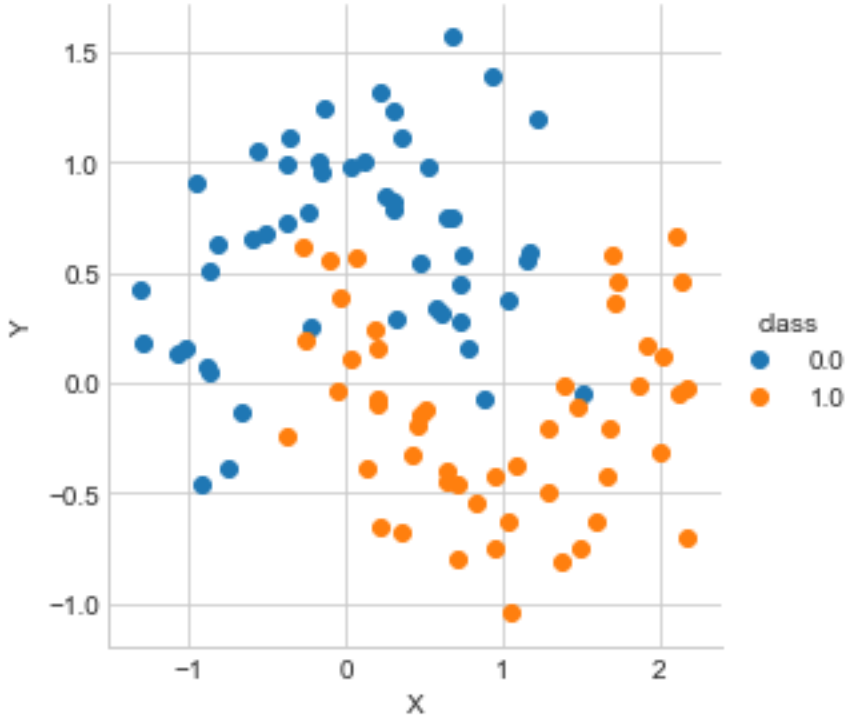
https://github.com/kevinsuo/CS7357/blob/master/knn/knn.ipynb

https://github.com/kevinsuo/CS7357/blob/master/knn/knn.py

KNN Practice

- pip install pandas
- pip install mlxtend

- U-Shaped dataset

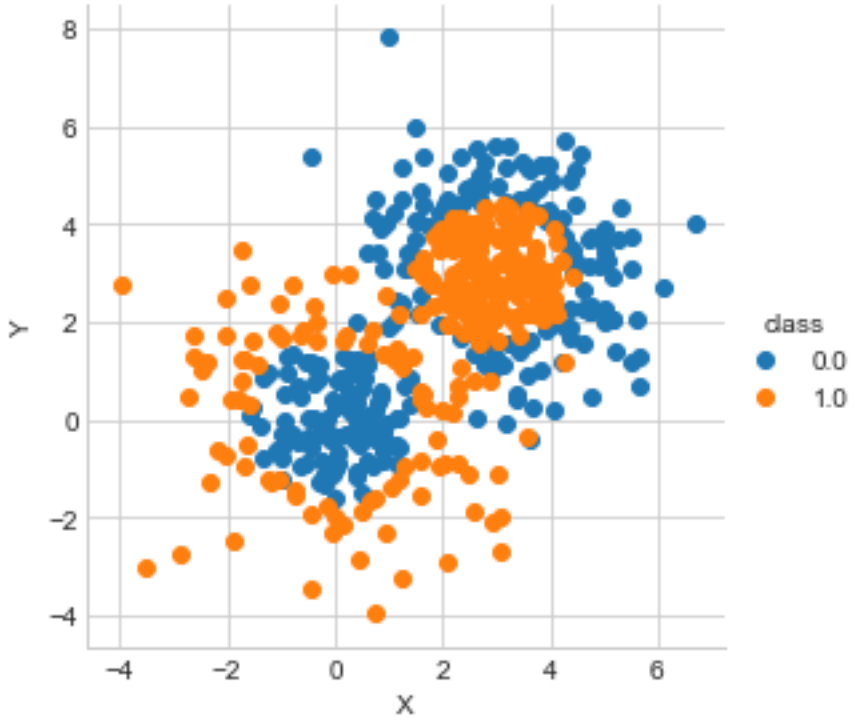


https://github.com/kevinsuo/CS7357/blob/master/knn/knn.ipynb

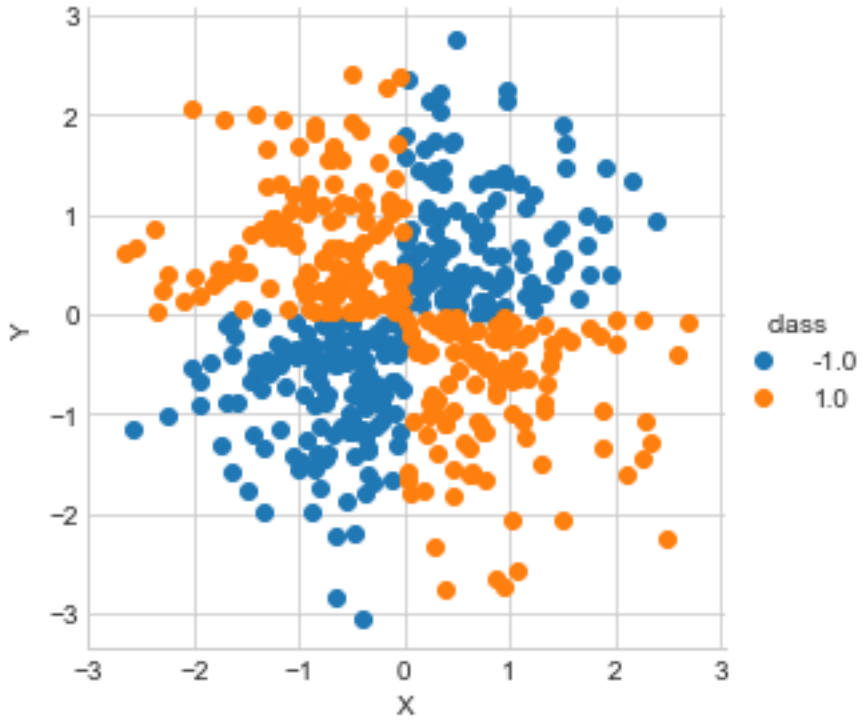
https://github.com/kevinsuo/CS7357/blob/master/knn/knn.py

KNN Practice

- Two set concentric circles dataset



- XOR dataset

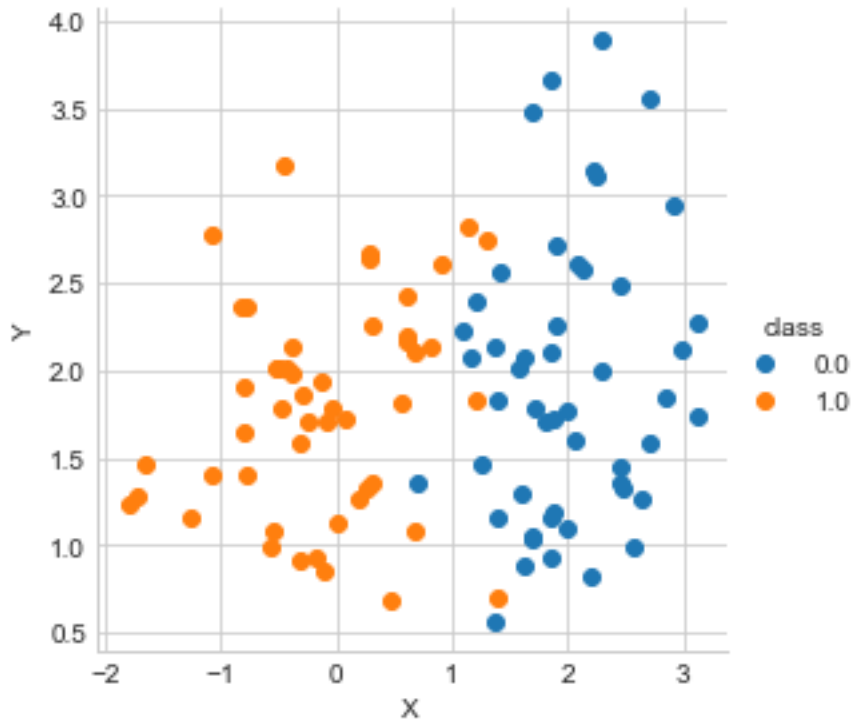


https://github.com/kevinsuo/CS7357/blob/master/knn/knn.ipynb

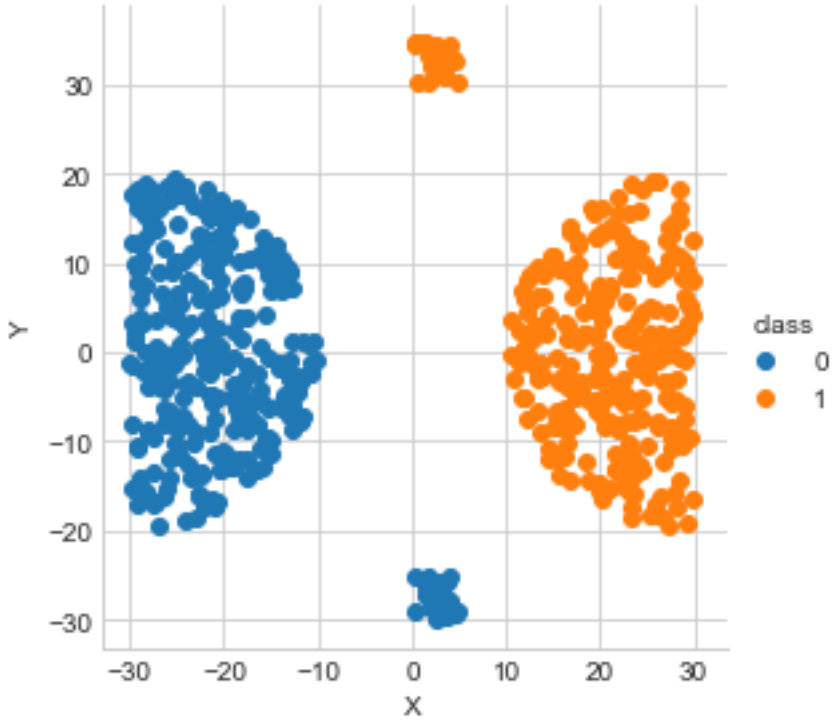
https://github.com/kevinsuo/CS7357/blob/master/knn/knn.py

KNN Practice

- Linearly separable dataset



- Outliers dataset

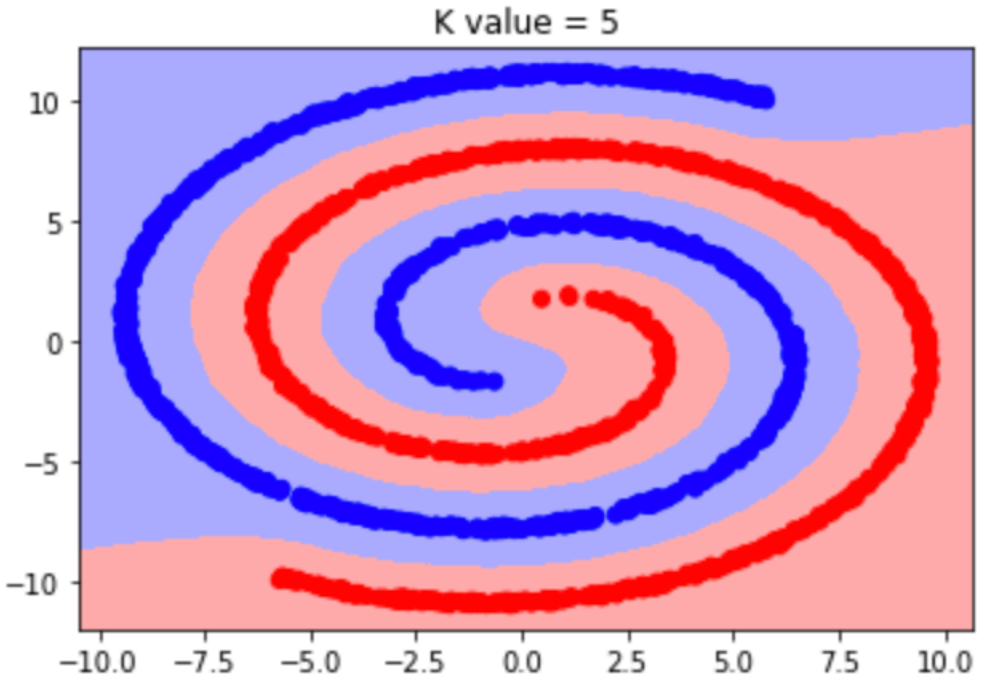


https://github.com/kevinsuo/CS7357/blob/master/knn/knn.ipynb

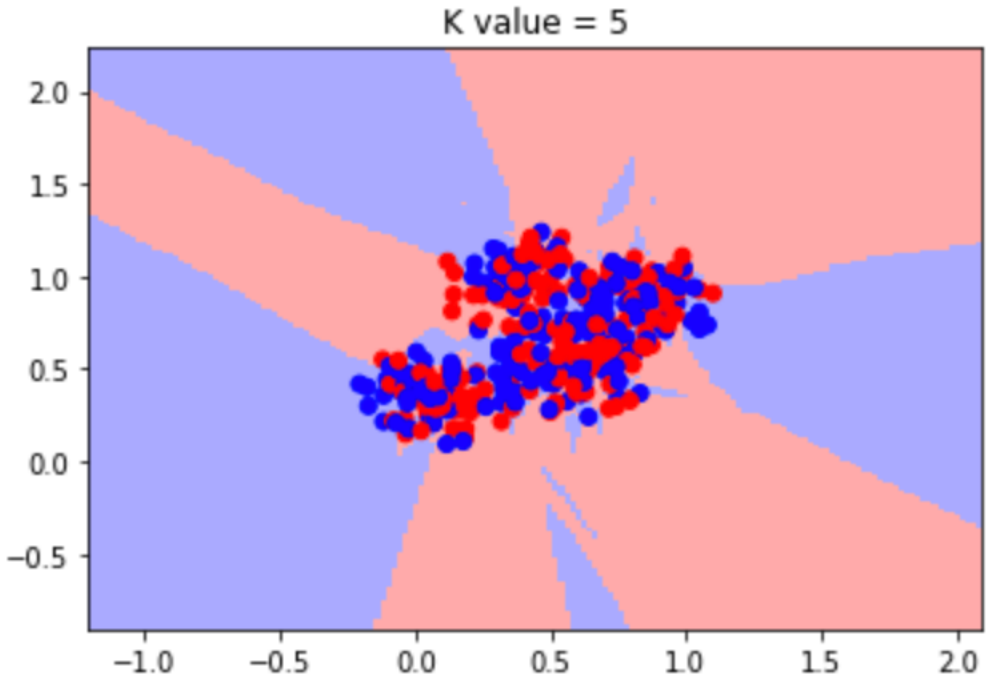
https://github.com/kevinsuo/CS7357/blob/master/knn/knn.py

KNN Practice

- Two spirals dataset



- random dataset



KNN Conclusion

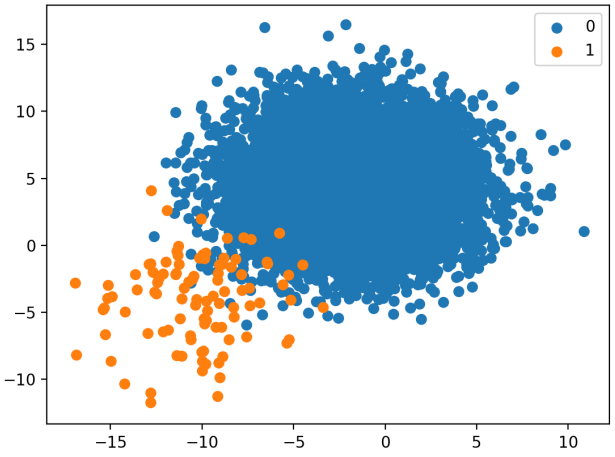
- Advantages:
  - The theory is very simple
  - Fast training (basically not learning in fact)
  - High accuracy

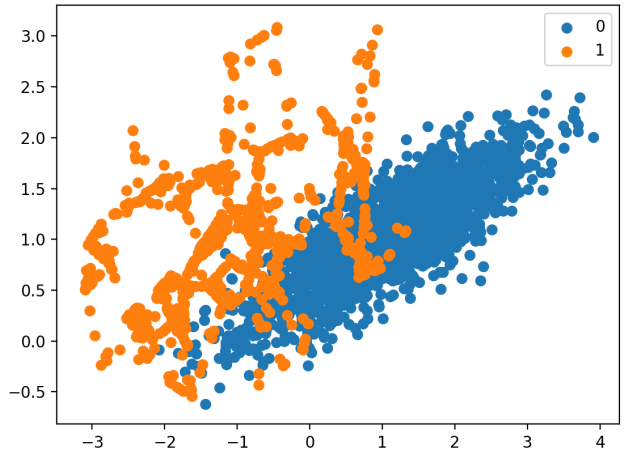
KNN Conclusion

- Disadvantages:
  - The amount of calculation is large (calculate the distance between the sample and the neighboring point).
  - The KNN model relies on the training data itself as a parameter and consumes a lot of memory.
  - When the sample is unbalanced, the prediction accuracy of the rare category is low.
  - Poor interpretability, it is difficult to form clear rules

KNN Conclusion

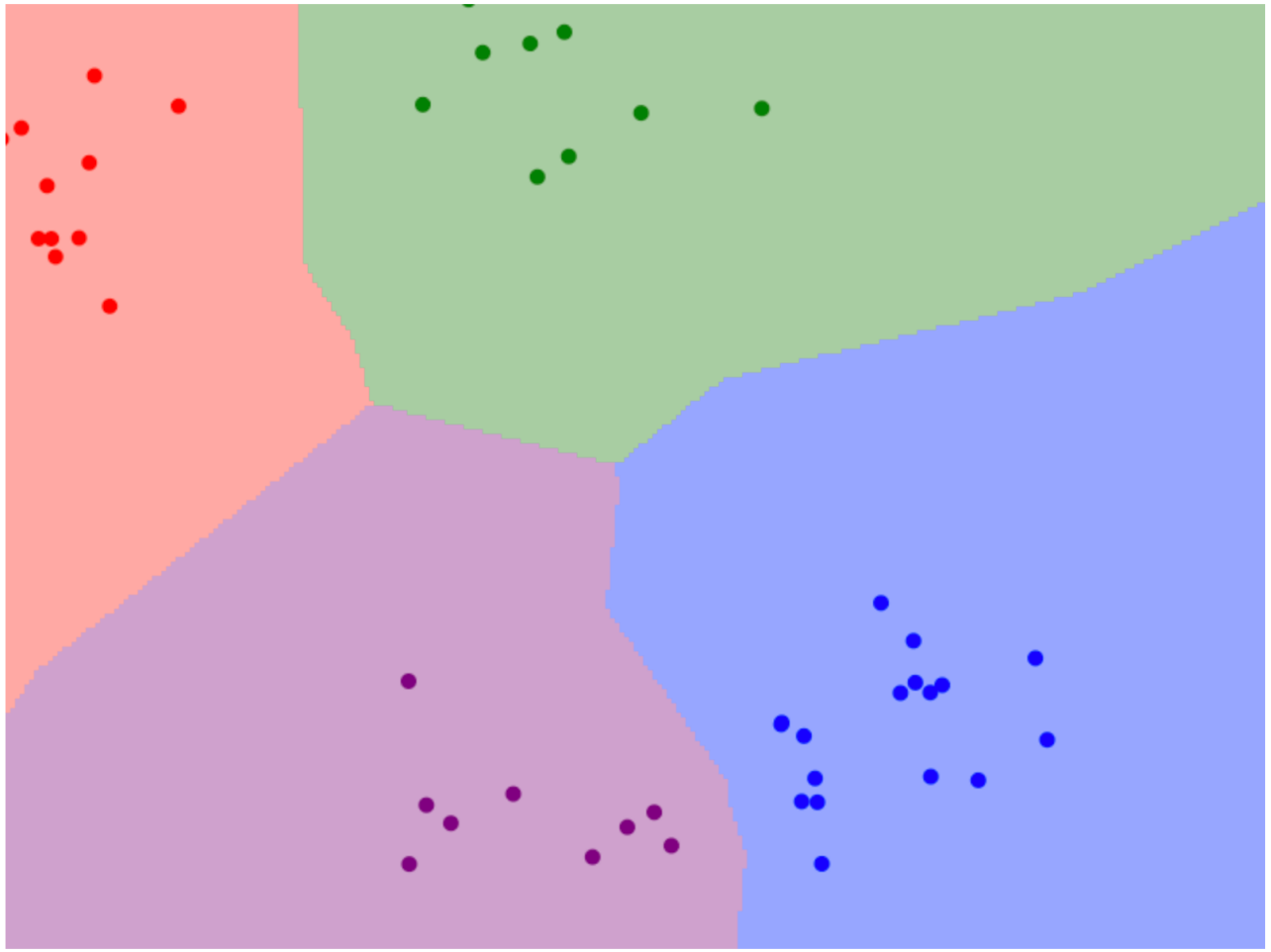
- Unbalanced data:
  - refers to classification problems where we have unequal instances for different classes





KNN playground

http://vision.stanford.edu/teaching/cs231n-demos/knn/



Metric

L1 L2

Num Neighbors (K)

1 2 3 4 5 6 7

Num classes

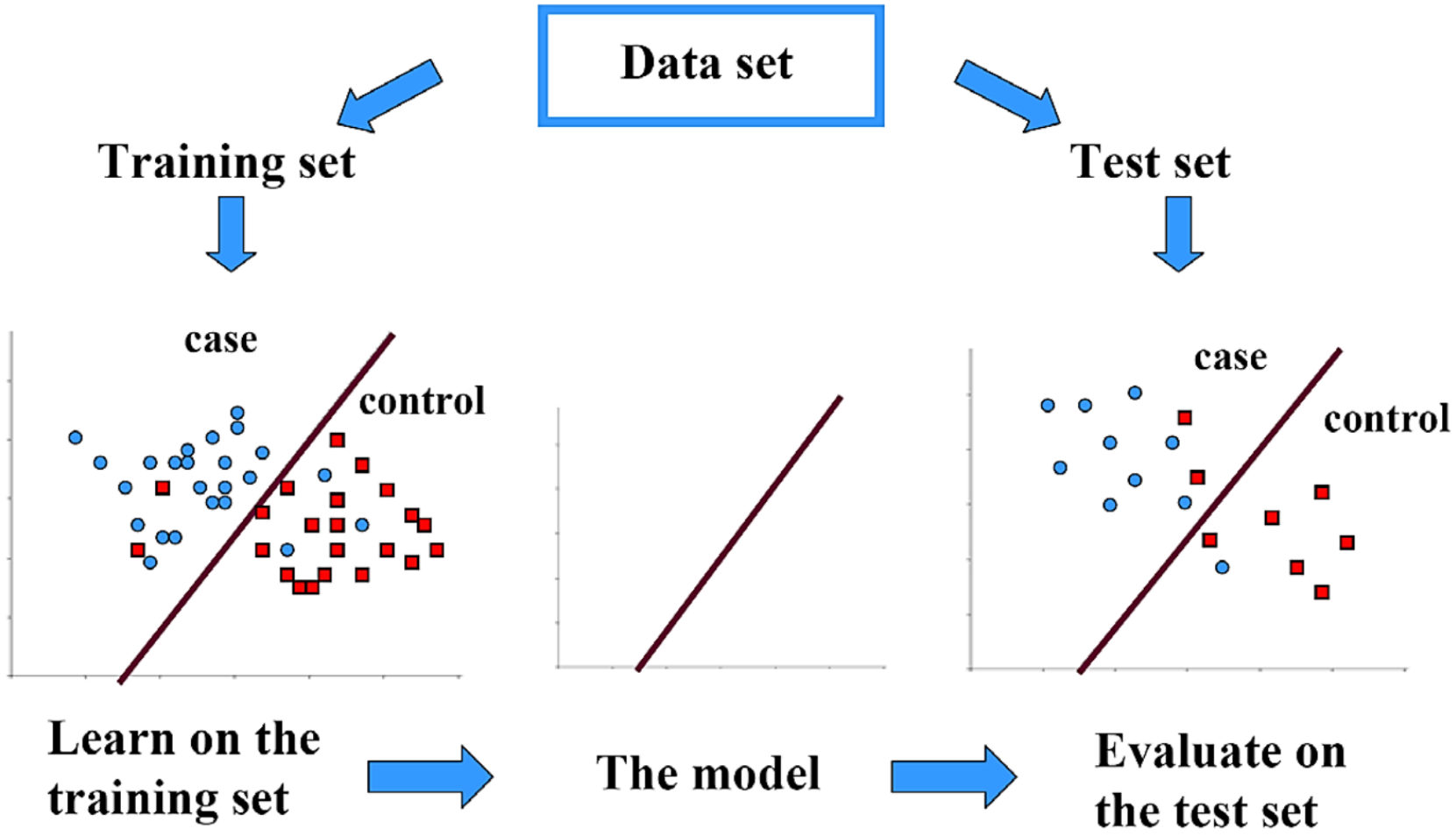
2 3 4 5

Num points

20 30 40 50 60

Evaluation

Evaluation for Classification



Evaluation Metrics

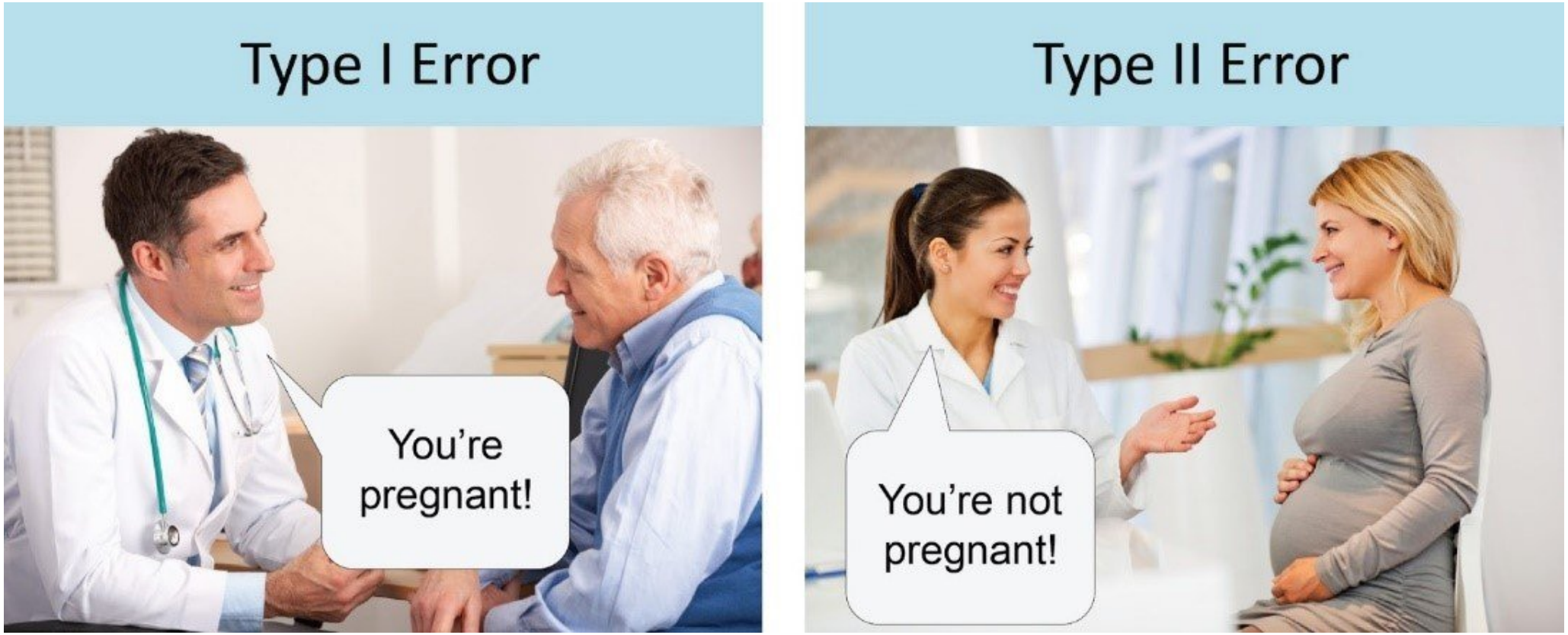
- Confusion Matrix: shows performance of an algorithm, especially predictive capability

		Predicted Class	
Actual Class		Class = YES	Class = No
	Class = Yes	True Positive	False Negative
	Class = No	False Positive	True Negative

Type II error

Type I error

Type I and II error



Evaluation Metrics

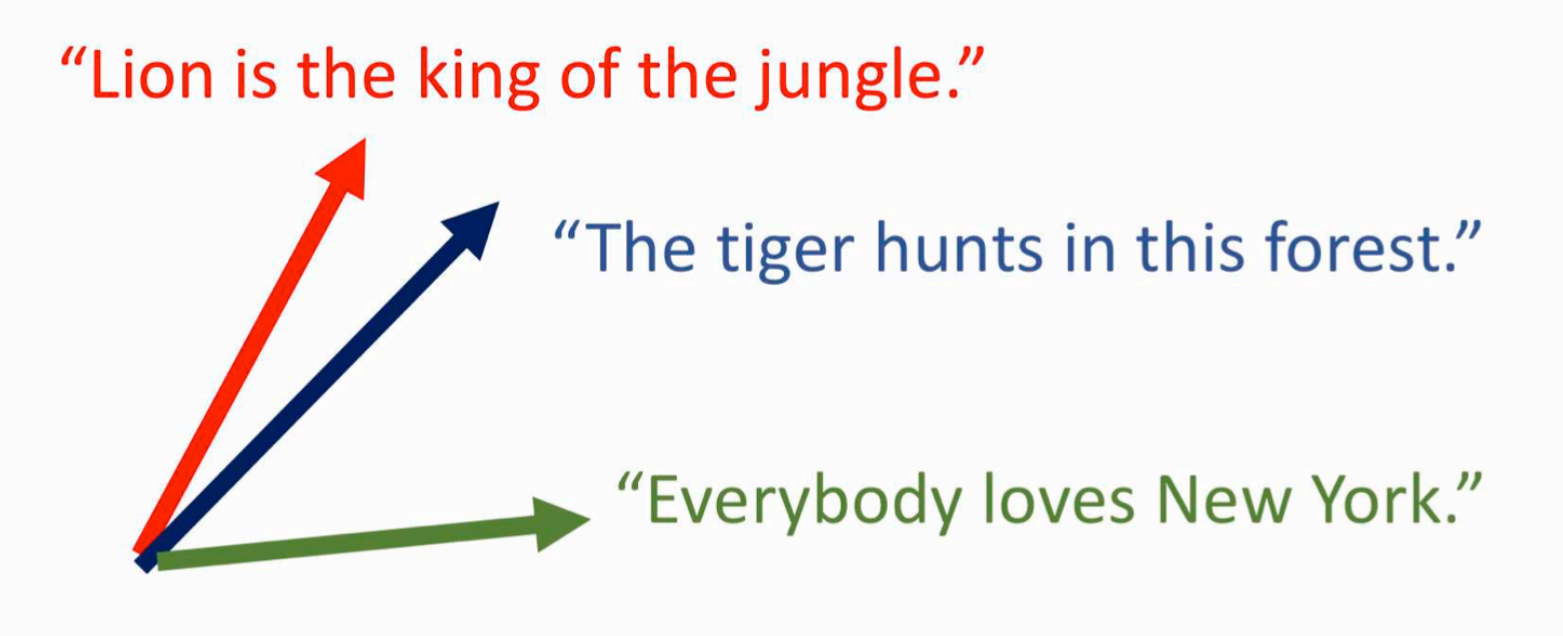
- Sensitivity or True Positive Rate (TPR)
  - TP/(TP+FN)
- Specificity or True Negative Rate (TNR)
  - TN/(FP+TN)
- Precision or Positive Predictive Value (PPV)
  - TP/(TP+FP)
- Negative Predictive Value (NPV)
  - TN/(TN+FN)
- Accuracy
  - (TP+TN)/(TP+FP+TN+FN)

Project 1

Project 1

- Text data set processing – encoding

- Why do we need text encoding? – Computability



Project 1

- Several basic steps of text encoding
  - 1. Divide the text into independent vocabulary
  - 2. Build a non-repetitive vocabulary
  - 3. Create a vector representation of the text based on the text and vocabulary

Project 1

Document number	The sentence words
Train 1	I buy an apple phone
Train 2	I eat a gig apple
Train 3	The apple are too expensive

non-repetitive vocabulary

I	buy	an	apple	phone	eat	a	gig	the	too	expensive

Project 1

- One-hot matrix:
  - A vector is used to represent a sentence, and the length of the vector is the size of the vocabulary. 1 means the corresponding word exists, 0 means it does not exist.

Document number	The sentence words
Train 1	I buy an apple phone

I	buy	an	apple	phone	eat	a	gig	the	too	expensive

I buy an apple phone =

1	1	1	1	1	0	0	0	0	0	0

Project 1

- Dataset:

data type	have a label	function
training set	Yes	Used to train the model or determine the model parameters, such as the determination of the weights in KNN
validation set	Yes	Used to determine the network structure/parameter or control the complexity of the model and modify the model.
test set	No	Used to test the performance of the optimal model.

Project 1

- Dataset:

Document number	The sentence words	Emotion
Train 1	I buy an apple phone	happy
Train 2	I eat the gig apple	happy
Train 3	The Apple products are too expensive	sad
Train 4	My friend has an apple	?

Project 1

- KNN classification problems
  - Input: original text
  - Output: class label (happy, sadness…)
  - Classification principle: majority voting principle

Document number	The sentence words	emotion
train 1	I buy an apple phone	happy
train 2	I eat the big apple	happy
train 3	The apple products are too expensive	sadnesss
test 1	My friend has an apple	?

Project 1

- Step 1: feature representation of the data set

Document number	The sentence words	emotion
train 1	I buy an apple phone	happy
train 2	I eat the big apple	happy
train 3	The apple products are too expensive	sadnesss
test 1	My friend has an apple	?

One-hot matrix:

Document number	I	buy	an	apple	...	friend	has	emotion
train 1	1	1	1	1	...	0	0	happy
train 2	1	0	0	1	...	0	0	happy
train 3	0	0	0	1	...	0	0	sadness
test 1	0	0	1	1	...	1	1	?

Project 1

- Step 2: similarity calculation
  - To calculate the Euclidean distance between test1 and each train, other distance measures can also be used)

$$d(train1, test1) = \sqrt{(1-0)^2 + (1-0)^2 + \cdots + (0-1)^2} = \sqrt{6};$$

$$d(train2, test1) = \sqrt{(1-0)^2 + (1-0)^2 + \cdots + (0-1)^2} = \sqrt{8};$$

$$d(train3, test1) = \sqrt{(0-0)^2 + (0-0)^2 + \cdots + (0-1)^2} = \sqrt{9};$$

If k=1, the label of test1 is the label happy of Train1;
if k=3, the label of test1 is train1, train2, and train3, which is the label happy.

Project 1

- Output on testset:

my_result

Words (split by space)	label
senator carl krueger thinks ipods can kill you	joy
who is prince frederic von anhalt	joy
prestige has magic touch	joy
study female seals picky about mates	joy
no e book for harry potter vii	joy
blair apologises over friendly fire inquest	fear
vegetables may boost brain power in older adults	surprise
afghan forces retake town that was overrun by taliban	fear
skip the showers male sweat turns women on study says	surprise
made in china irks some burberry shoppers	joy
britain to restrict immigrants from new eu members	joy
canadian breakthrough offers hope on autism	joy
russia to strengthen its military muscle	fear
alzheimer s drugs offer no help study finds	joy
uk police slammed over terror raid	fear