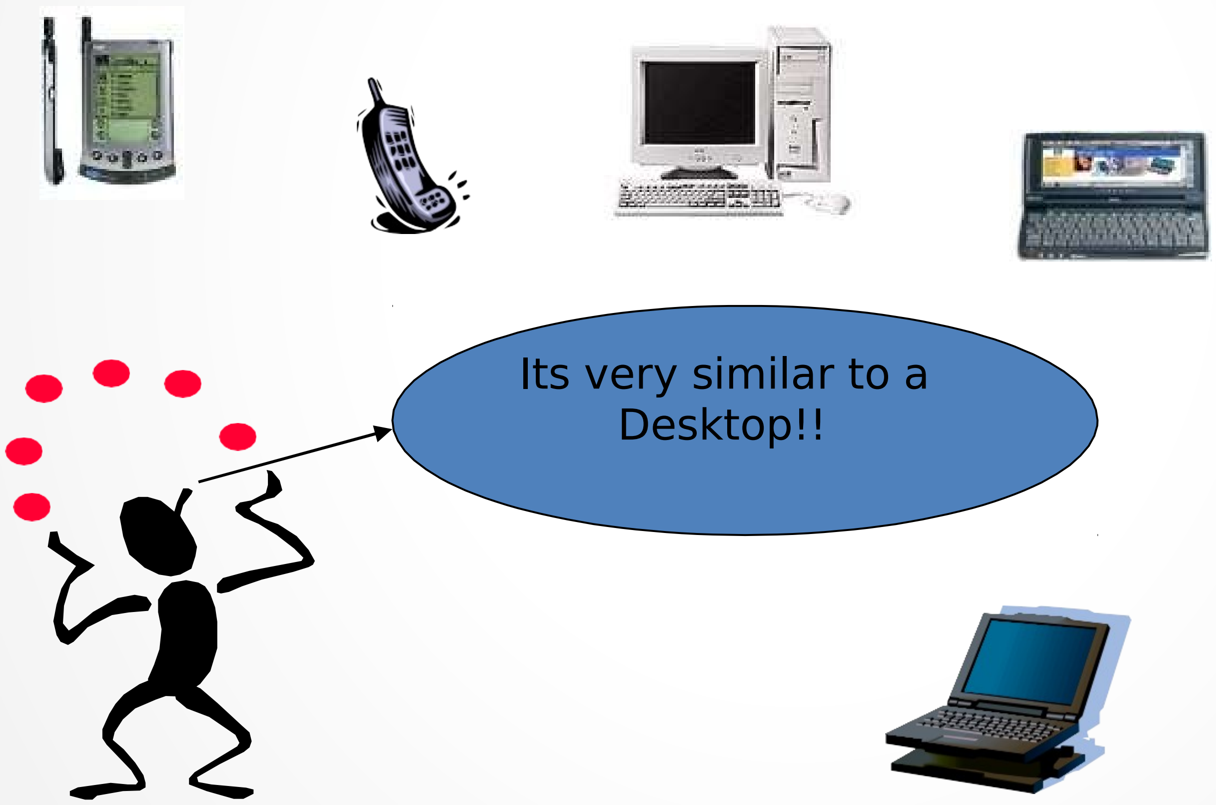
K nearest neighbor

Jony Sugianto
jony@evolvemachinelearners.com
0812-13086659
github.com/jonysugianto

Simple Analogy..

- Tell me about your friends(*who your neighbors are*) and *I will tell you who you are*.

Instance-based Learning



KNN – Different names

- K-Nearest Neighbors
- Memory-Based Reasoning
- Example-Based Reasoning
- Instance-Based Learning
- Lazy Learning

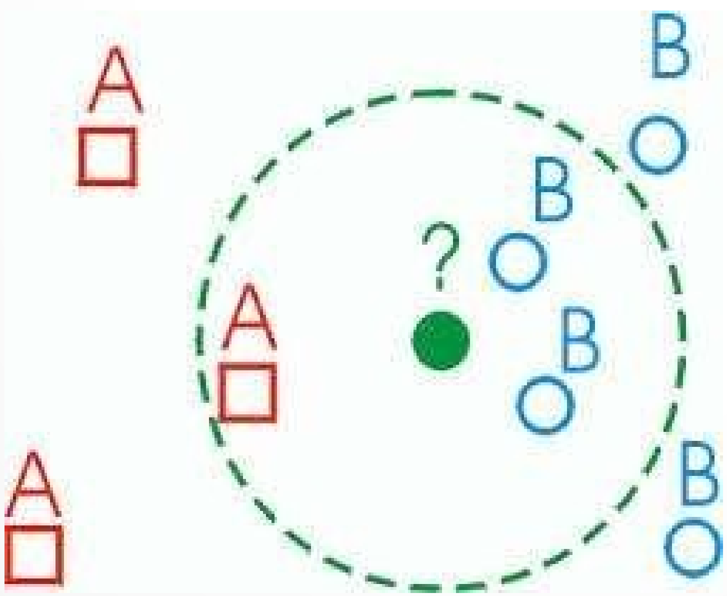
What is KNN?

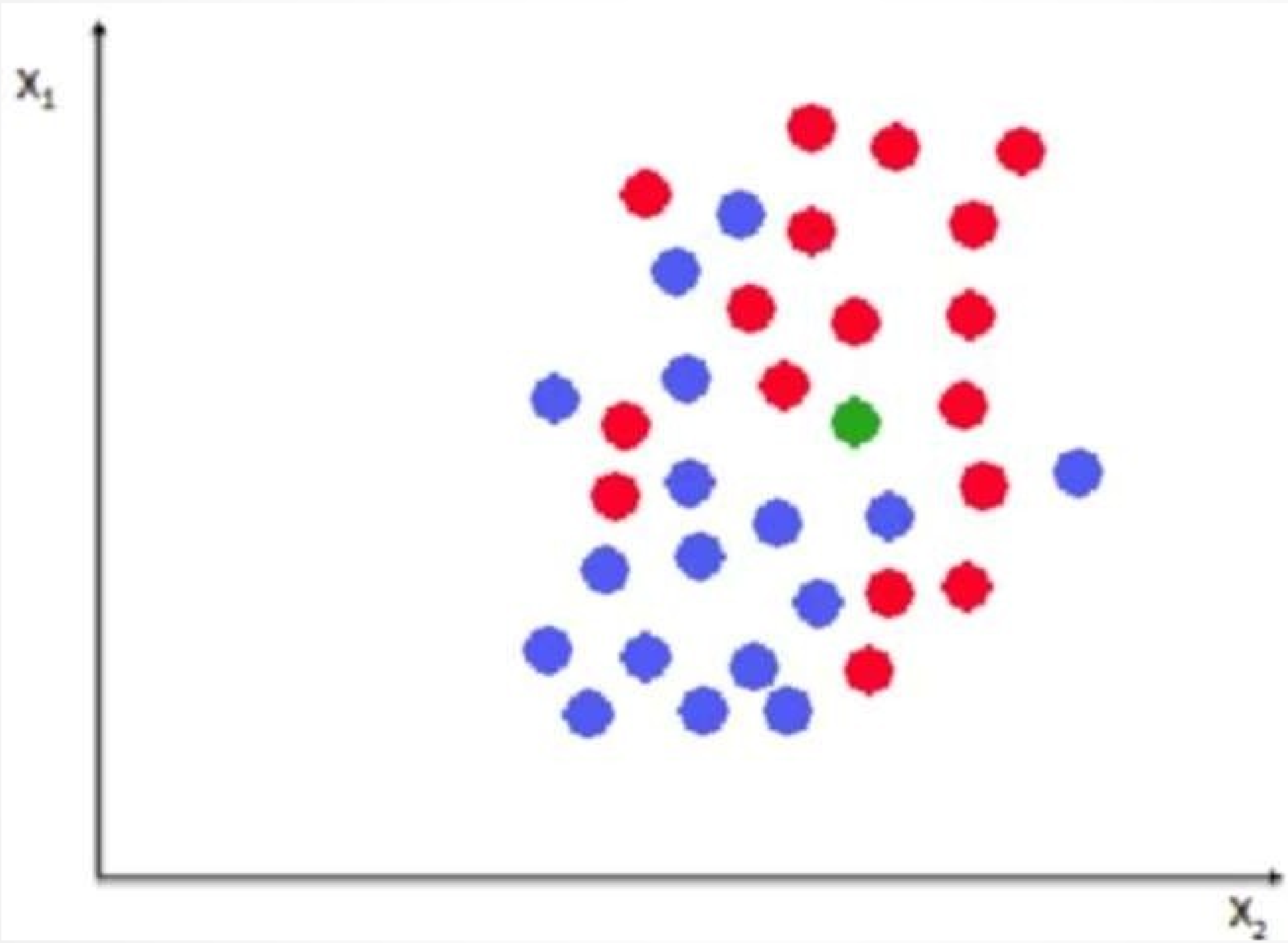
EVOLVE
MACHINE LEARNERS

- A powerful classification algorithm used in pattern recognition
- K nearest neighbors stores all available cases and classifies new cased based on a similarity measure (e.g. distance function)
- One of the top data mining algorithms used today
- A non parametric lazy learning algorithm(an instance based learning method)

KNN: Classification Approach

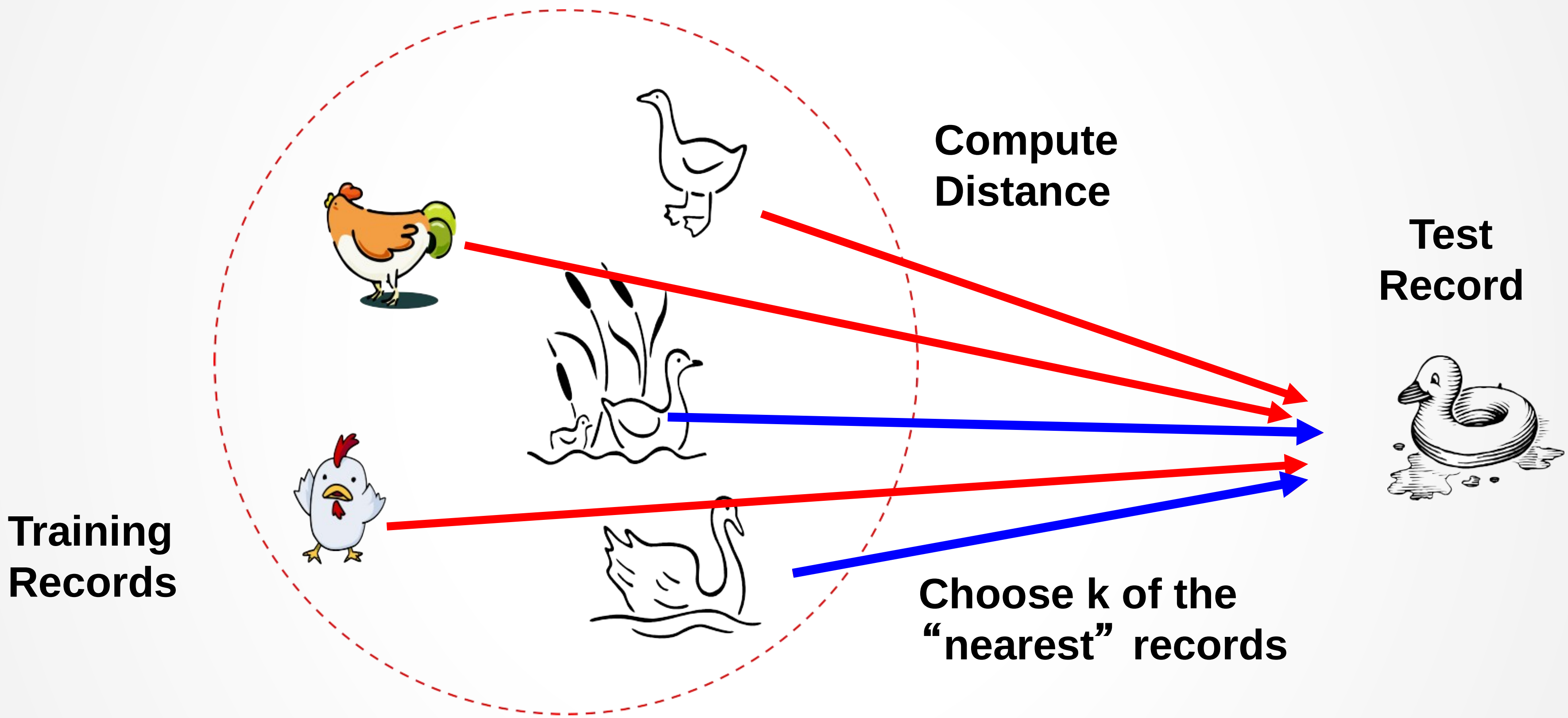
- An object (a new instance) is classified by a majority votes for its neighbor classes.
- The object is assigned to the most common class amongst its K nearest neighbors.(*measured by a distance function*)





X1
X2

Distance Measure



Distance measure for Continuous Variables

Distance functions

Euclidean

$$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$$

Manhattan

$$\sum_{i=1}^{k}|x_i - y_i|$$

Minkowski

$$\left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{1/q}$$

When is the difference?

Distance Between Neighbors

EVOLVE MACHINE LEARNERS

- Calculate the distance between new example (E) and all examples in the training set.

- *Euclidean* distance between two examples.
  - $X = [x_1, x_2, x_3, .., x_n]$
  - $Y = [y_1, y_2, y_3, ..., y_n]$
  - The Euclidean distance between *X* and *Y* is defined as:

$$D(X, Y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Non Numeric Data

- Feature values are not always numbers
- Example
  - Boolean values: Yes or no, presence or absence of an attribute
  - Categories: Colors, educational attainment, gender
- How do these values factor into the computation of distance?

Dealing with non numerical data

- Boolean values => convert to 0 or 1
  - Applies to yes-no/presence-absence attributes
- Non-binary characterizations
  - Use natural progression when applicable; e.g., educational attainment: GS, HS, College, MS, PHD => 1,2,3,4,5
  - Assign arbitrary numbers but be careful about distances; e.g., color: red, yellow, blue => 1,2,3
- How about unavailable data?
  (0 value not always the answer)

Nominal/Categorical Data

- Distance works naturally with numerical attributes.
- Binary value categorical data attributes can be regarded as 1 or 0.

Hamming Distance

$$D_H = \sum_{i=1}^{k} |x_i - y_i|$$

$$x = y \Rightarrow D = 0$$

$$x \neq y \Rightarrow D = 1$$

X	Y	Distance
Male	Male	0
Male	Female	1

Distance Between Neighbors

EVOLVE
MACHINE LEARNERS

- Calculate the distance between new example (E) and all examples in the training set.

- *Euclidean* distance between two examples.
  - $X = [x_1, x_2, x_3, .., x_n]$
  - $Y = [y_1, y_2, y_3, ..., y_n]$

  - The Euclidean distance between *X* and *Y* is defined as:

$$D(X,Y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

K-Nearest Neighbor Algorithm

EVOLVE
MACHINE LEARNERS

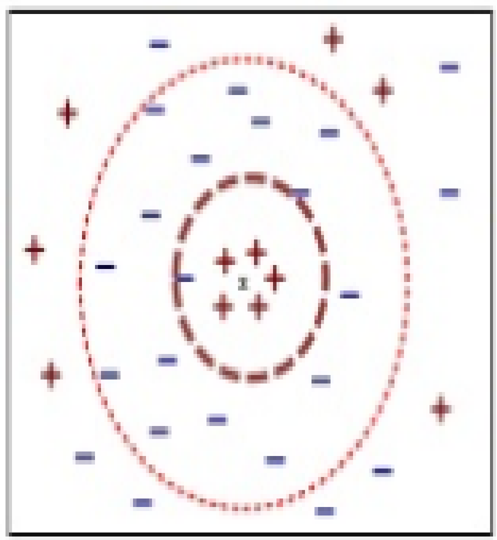
- All the instances correspond to points in an n-dimensional feature space.
- Each instance is represented with a set of numerical attributes.
- Each of the training data consists of a set of vectors and a class label associated with each vector.
- Classification is done by comparing feature vectors of different K nearest points.
- Select the K-nearest examples to E in the training set.
- Assign E to the most common class among its K-nearest neighbors.

3-KNN: Example(1)

Customer	Age	Income	No. credit cards	Class	Distance from John
George	35	35K	3	No	Sqrt((35-37)**2+(35-50)**2+(3-2)**2)=15.16
Rachel	22	50K	2	Yes	Sqrt((22-37)**2+(50-50)**2+(2-2)**2)=15
Steve	63	200K	1	No	152.23
Tom	59	170K	1	No	122
Anne	25	40K	4	Yes	15.74
John	37	50K	2	?	Yes

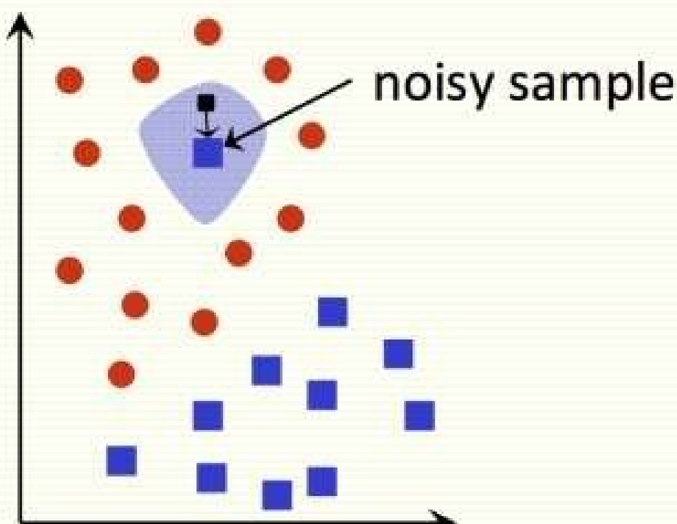
How to choose K?

- If K is too small it is sensitive to noise points.
- Larger K works well. But too large K may include majority points from other classes.



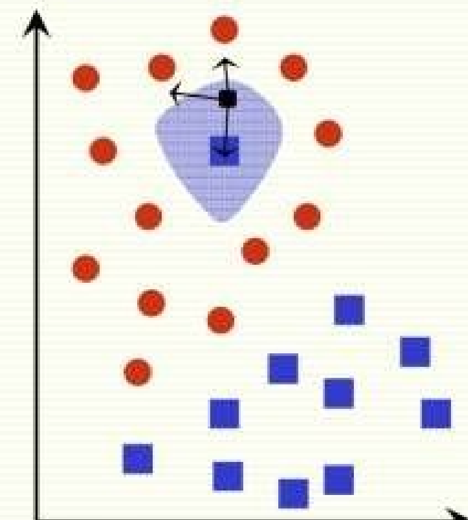
- Rule of thumb is K < sqrt(n), n is number of examples.

1 NN

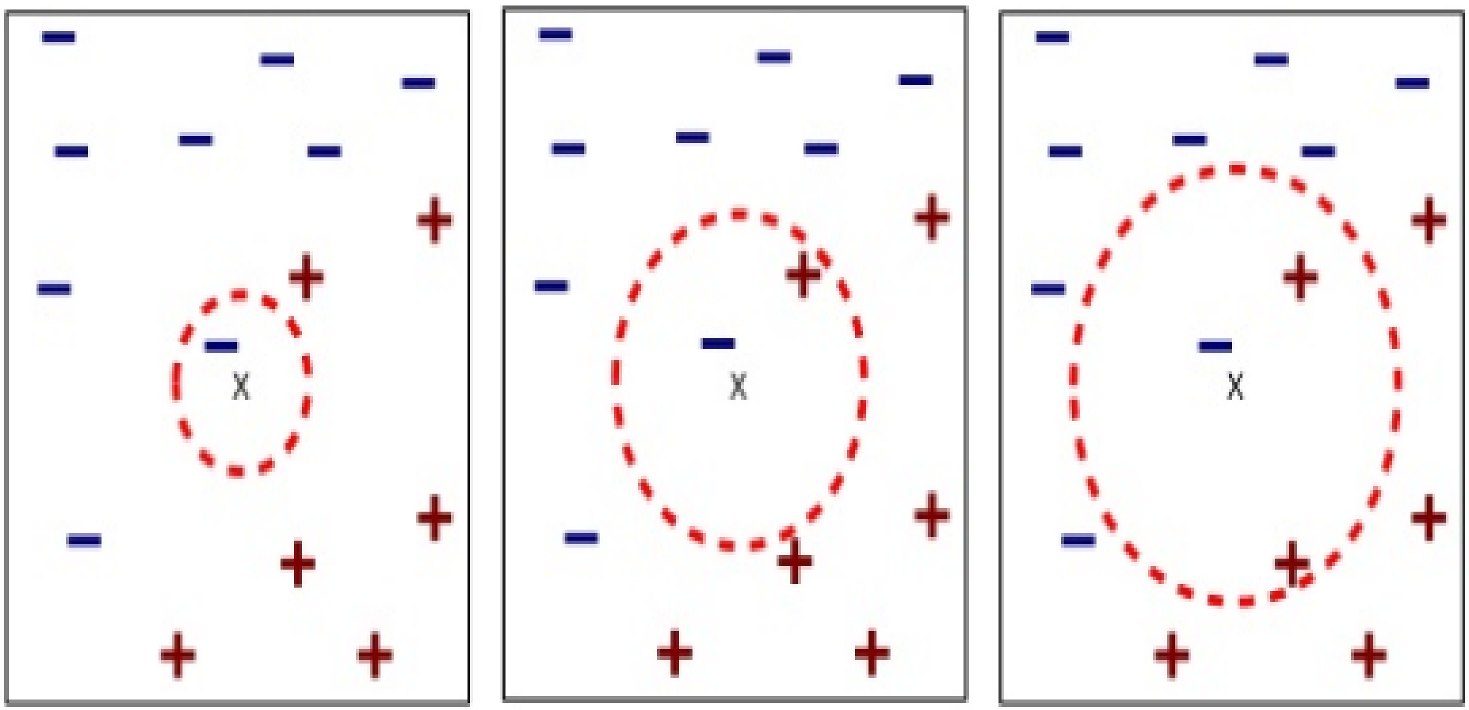


every example in the blue shaded area will be misclassified as the blue class

3 NN



every example in the blue shaded area will be classified correctly as the red class



(a) 1-nearest neighbor (b) 2-nearest neighbor (c) 3-nearest neighbor

K-nearest neighbors of a record x are data points that have the k smallest distance to x

KNN Feature Weighting

- Scale each feature by its importance for classification

$$D(a,b) = \sqrt{\sum_k \boxed{w_k} (a_k - b_k)^2}$$

- Can use our prior knowledge about which features are more important
- Can learn the weights wk using cross-validation (to be covered later)

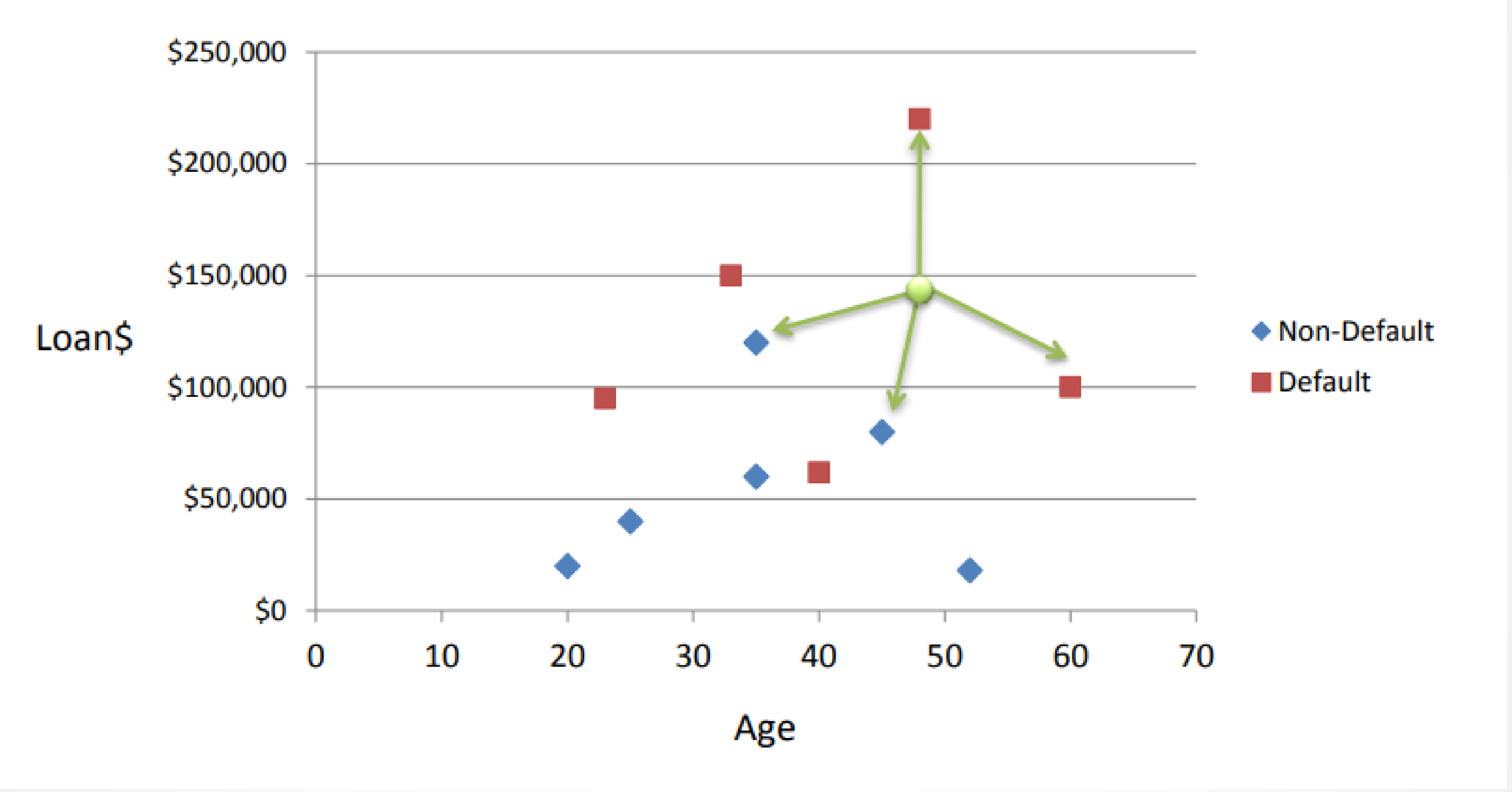
Feature Normalization

- Distance between neighbors could be dominated by some attributes with relatively large numbers. e.g., income of customers in our previous example.

$$a_i = \frac{v_i - \min v_i}{\max v_i - \min v_i}$$

- Arises when two features are in different scales.
- Important to normalize those features.
  – Mapping values to numbers between 0 – 1.

KNN Classification



KNN Classification – Distance

Age	Loan	Default	Distance
25	$40,000	N	102000
35	$60,000	N	82000
45	$80,000	N	62000
20	$20,000	N	122000
35	$120,000	N	22000
52	$18,000	N	124000
23	$95,000	Y	47000
40	$62,000	Y	80000
60	$100,000	Y	42000
48	$220,000	Y	78000
33	$150,000	Y	8000
48	**$142,000**	**?**	

Euclidean Distance

$$D = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

KNN Classification – Standardized Distance

Age	Loan	Default	Distance
0.125	0.11	N	0.7652
0.375	0.21	N	0.5200
0.625	0.31	N	0.3160
0	0.01	N	0.9245
0.375	0.50	N	0.3428
0.8	0.00	N	0.6220
0.075	0.38	Y	0.6669
0.5	0.22	Y	0.4437
1	0.41	Y	0.3650
0.7	1.00	Y	0.3861
0.325	0.65	Y	0.3771
0.7	**0.61**	**?**	

Standardized Variable

$$X_s = \frac{X - Min}{Max - Min}$$

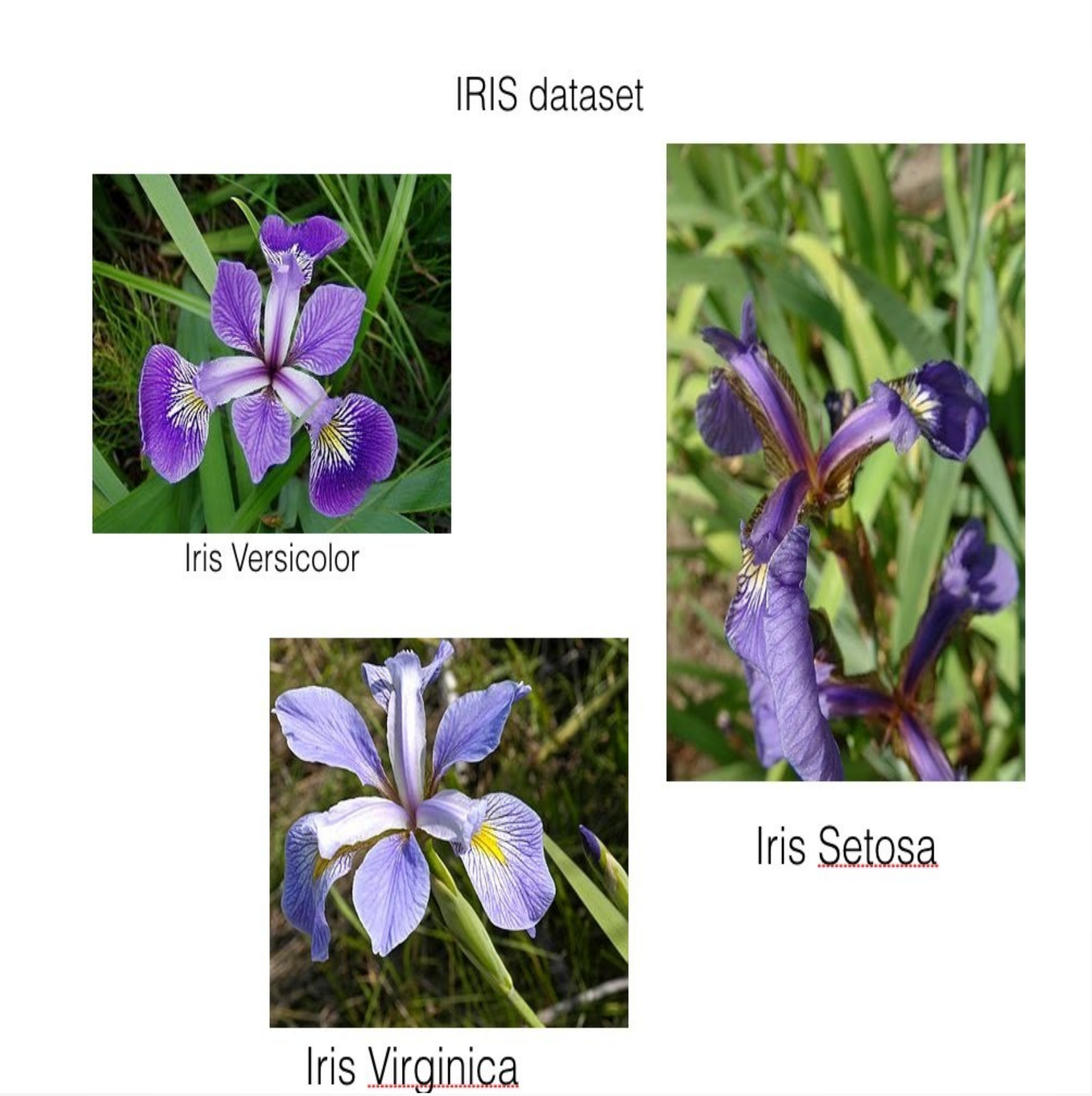
Strengths of KNN

- Very simple and intuitive.
- Can be applied to the data from any distribution.
- Good classification if the number of samples is large enough.

Weaknesses of KNN

- Takes more time to classify a new example.
  - need to calculate and compare distance from new example to all other examples.
- Choosing k may be tricky.
- Need large number of samples for accuracy.

Iris Flower Dataset



Iris Flower Dataset

EVOLVE
MACHINE LEARNERS

Fisher's *Iris* Data

Sepal length	Sepal width	Petal length	Petal width	Species
5.1	3.5	1.4	0.2	*I. setosa*
4.9	3.0	1.4	0.2	*I. setosa*
4.7	3.2	1.3	0.2	*I. setosa*
4.6	3.1	1.5	0.2	*I. setosa*
5.0	3.6	1.4	0.2	*I. setosa*
5.4	3.9	1.7	0.4	*I. setosa*
4.6	3.4	1.4	0.3	*I. setosa*
5.0	3.4	1.5	0.2	*I. setosa*
4.4	2.9	1.4	0.2	*I. setosa*
4.9	3.1	1.5	0.1	*I. setosa*
5.4	3.7	1.5	0.2	*I. setosa*
4.8	3.4	1.6	0.2	*I. setosa*
4.8	3.0	1.4	0.1	*I. setosa*

Sort ascending

Mnist Dataset