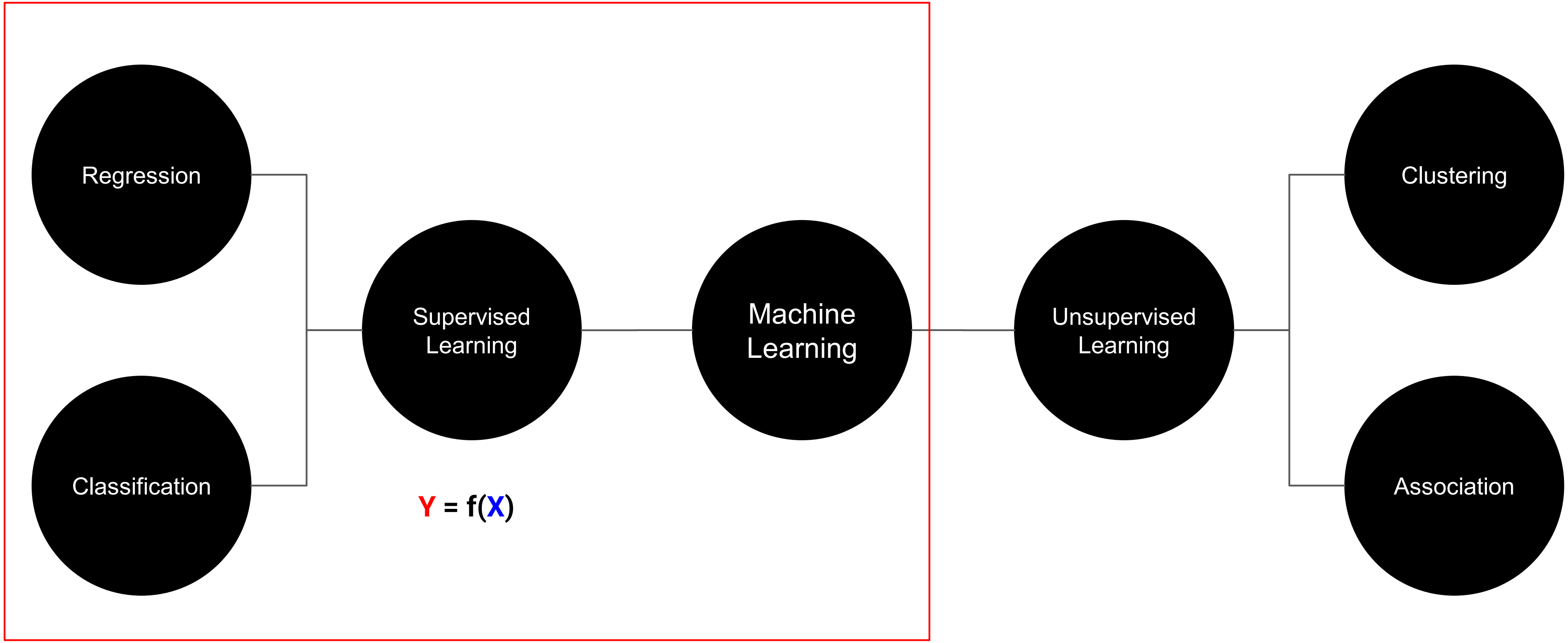
Data Prediction Model and Machine Learning

Online course #6

K-NN (Nearest Neighbors)



Regression

Classification

Supervised Learning

Machine Learning

Unsupervised Learning

Clustering

Association

$Y = f(X)$

K-NN

(Nearest Neighbours)

“Birds of a feather flock together”

類類相從

K-NN

(Nearest Neighbours)

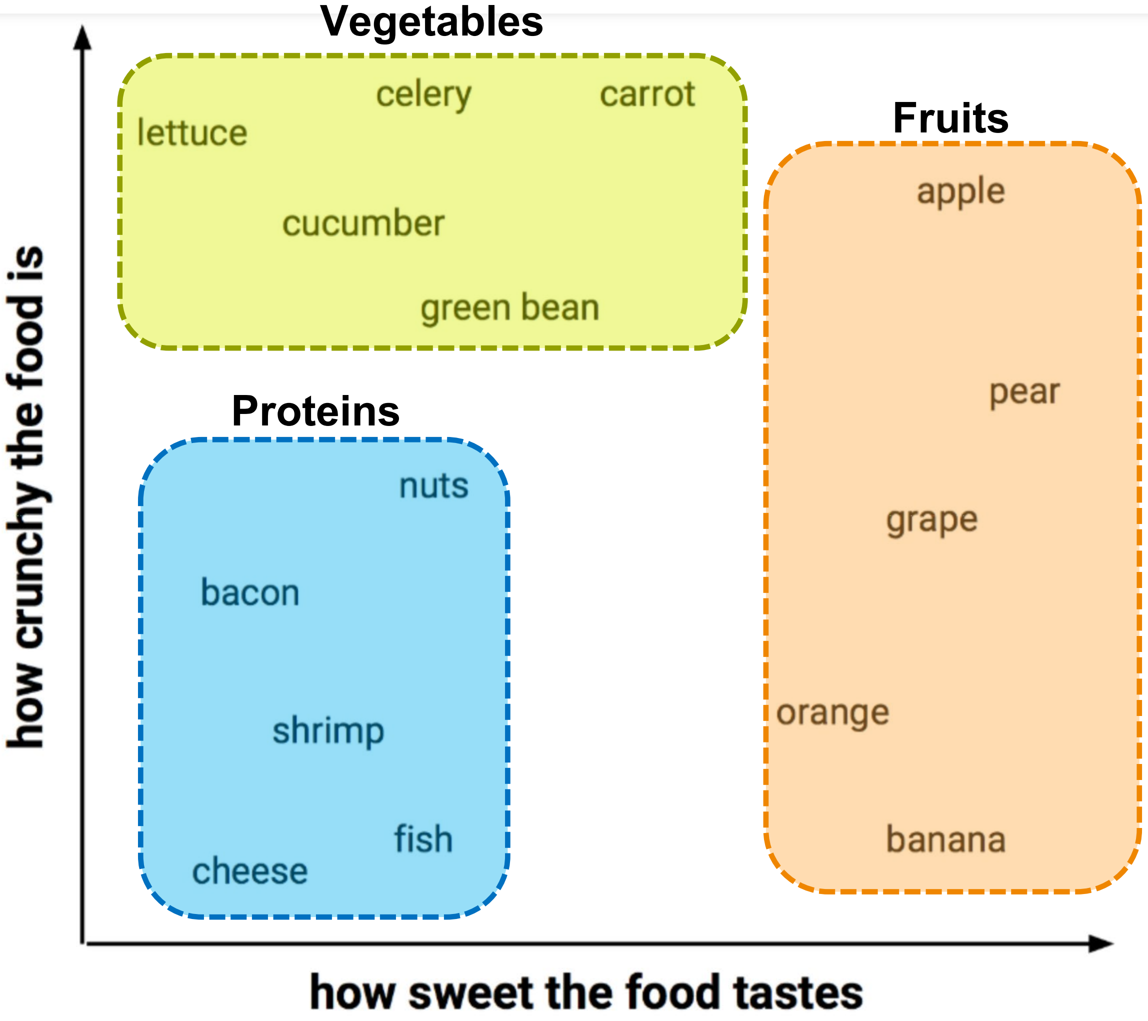
Blind testing

Ingredients	Sweet taste	Crunchy	Type
Apple	10	9	Fruit
Bacon	1	4	Protein
Banana	10	1	Fruit
Carrot	7	10	Vegetable
Salary	3	10	Vegetable
Cheese	1	1	Protein

K-NN

(Nearest Neighbours)

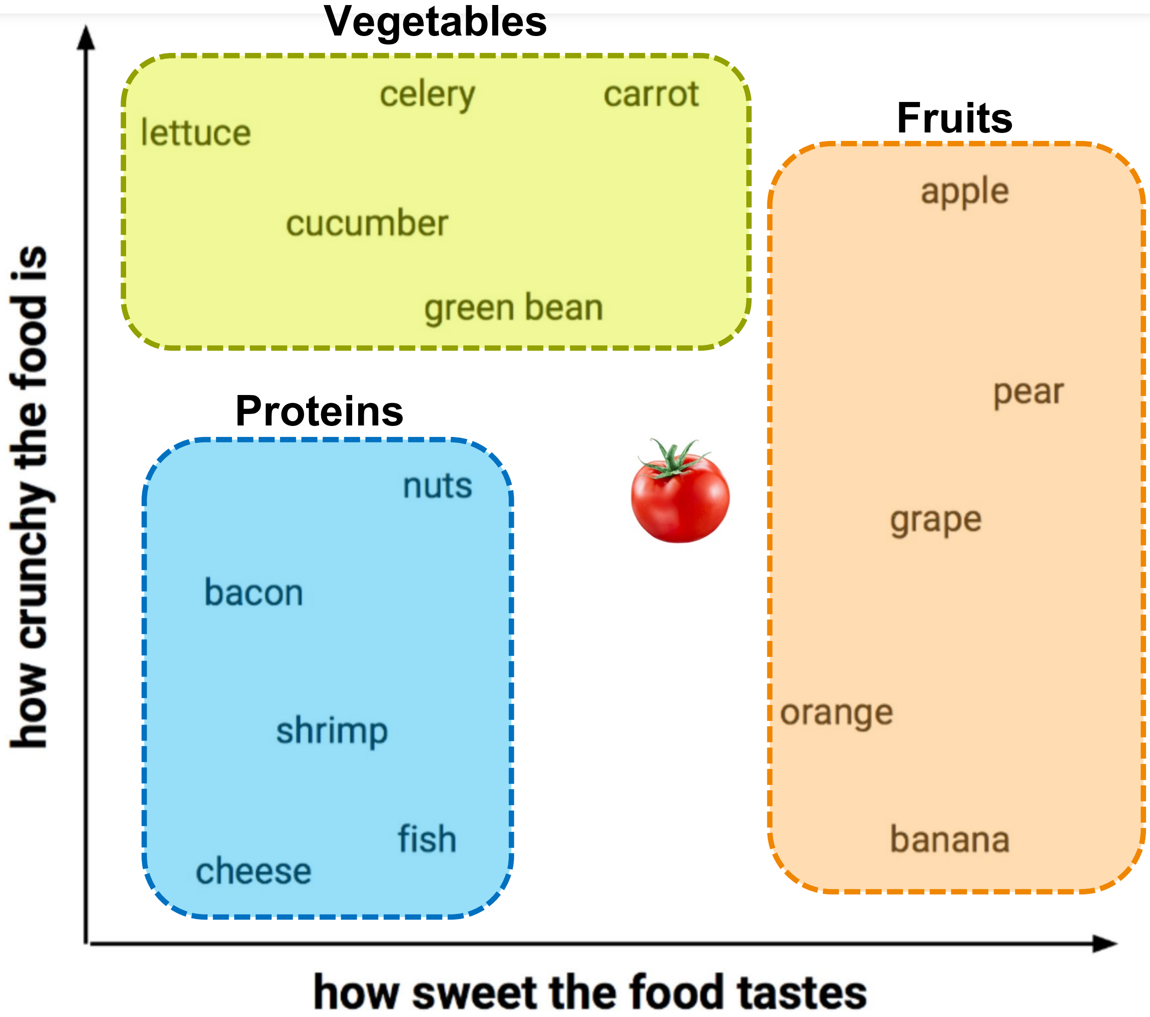
- **Vege**: Crunchy but not sweet
- **Fruit**: Mostly sweet
- **Protein**: not so crunchy and not sweet as well



Is Tomato Fruit or Vegetable?

K-NN

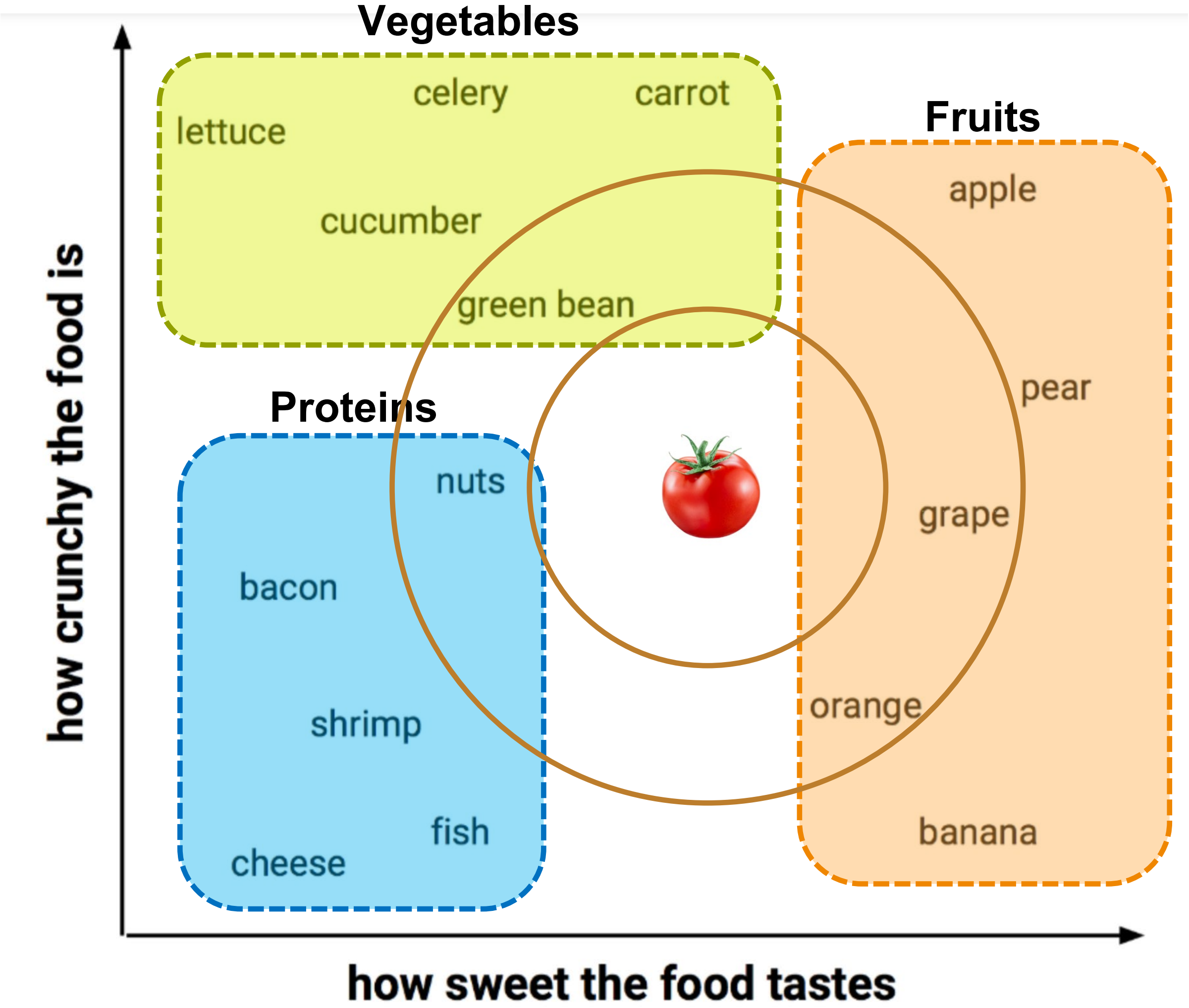
(Nearest Neighbours)



K-NN

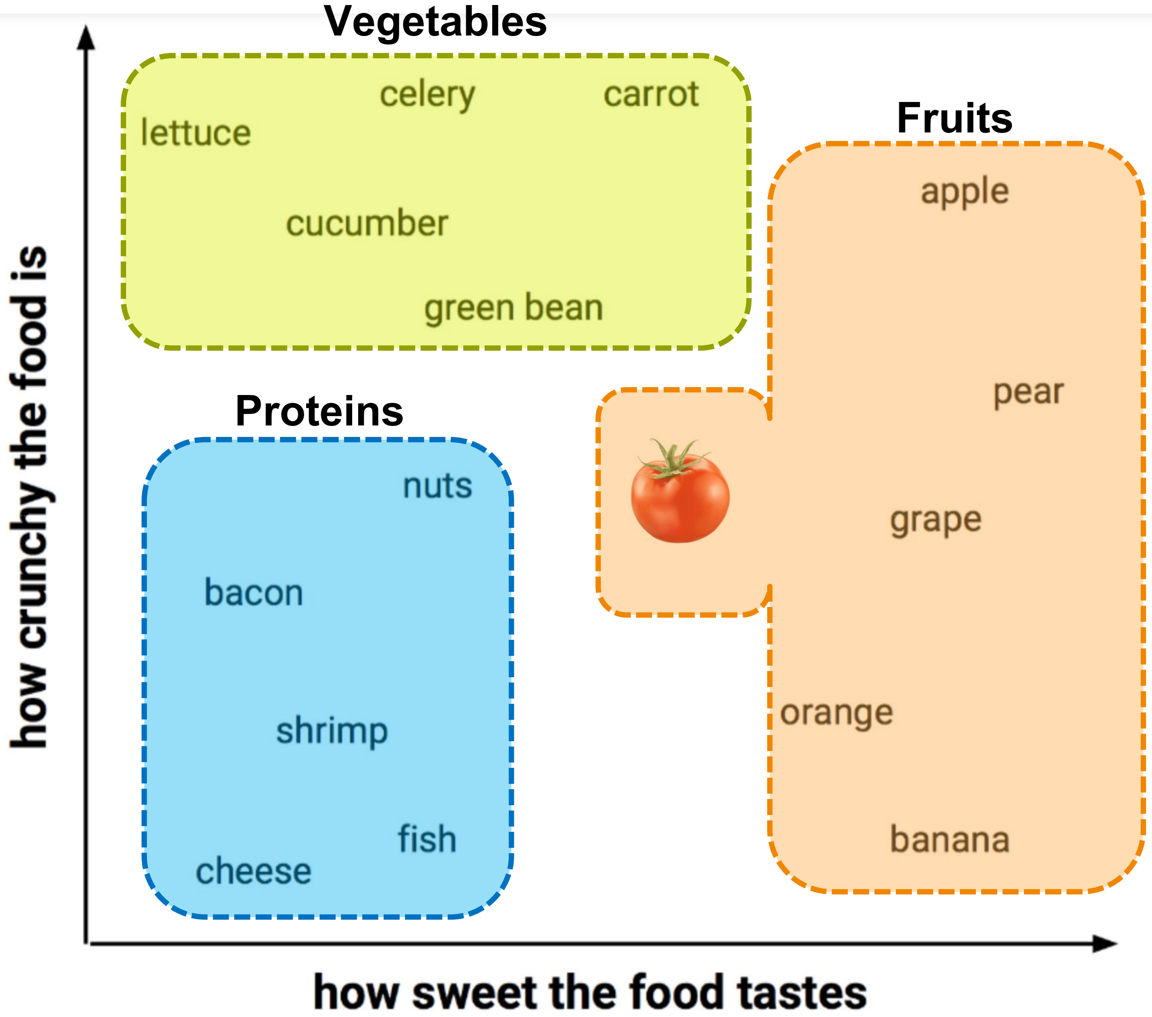
(Nearest Neighbours)

K= 4



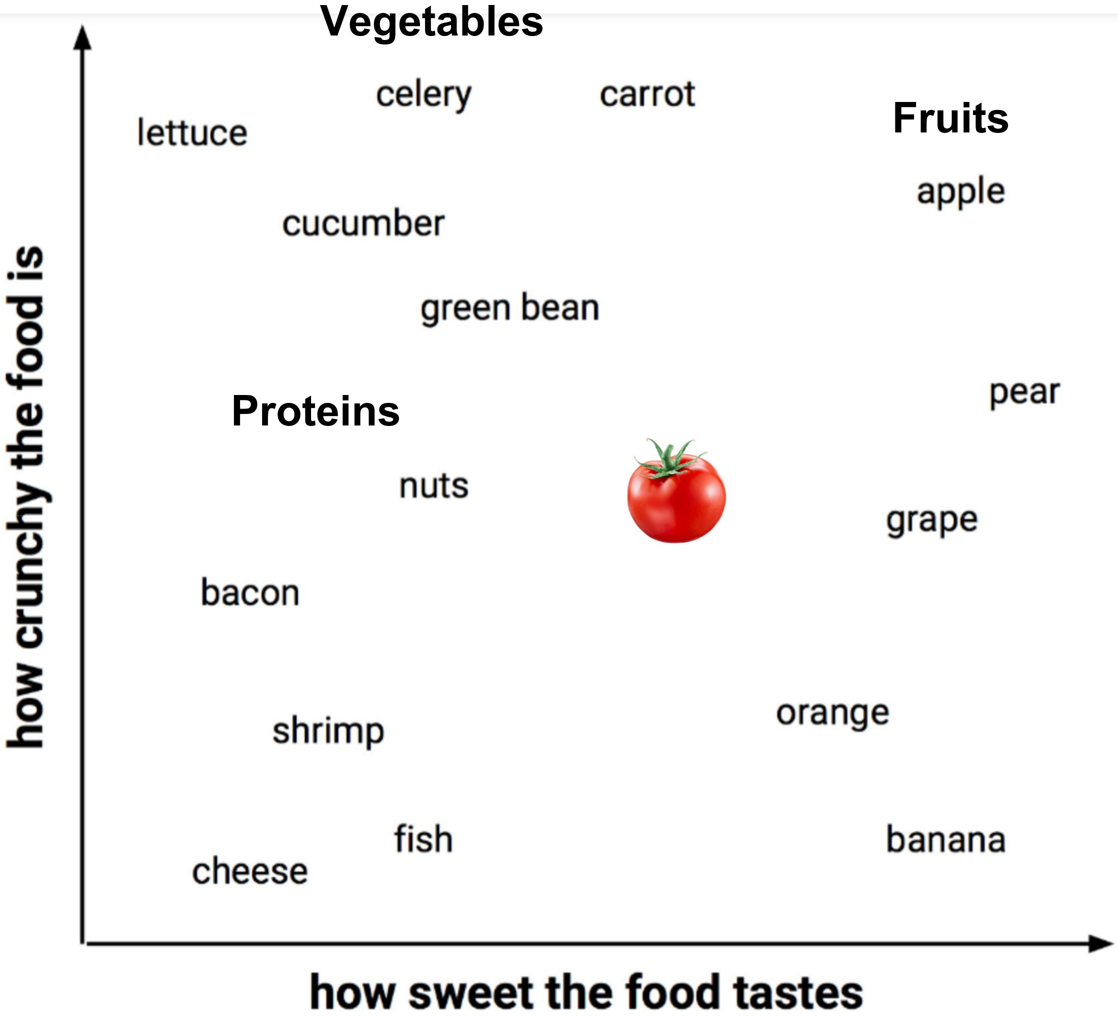
K-NN

(Nearest Neighbours)



K-NN

(Nearest Neighbours)

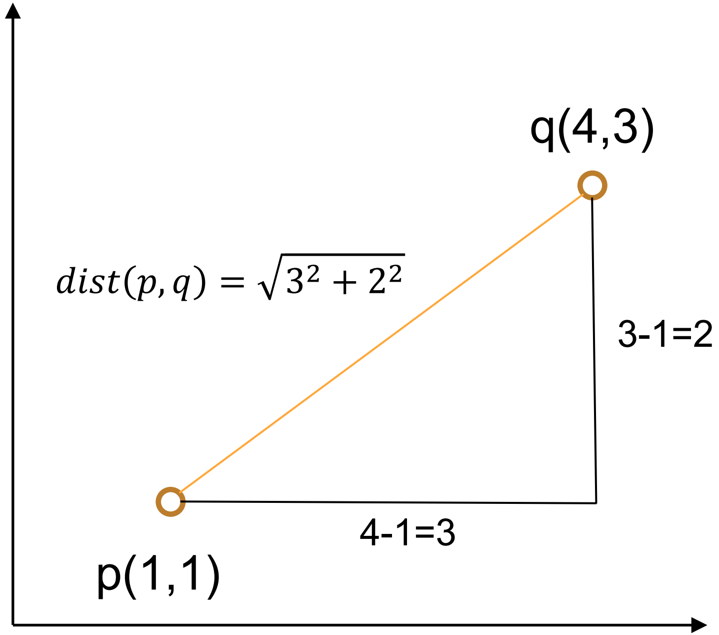


How to measure the distance to the nearest neighbours?

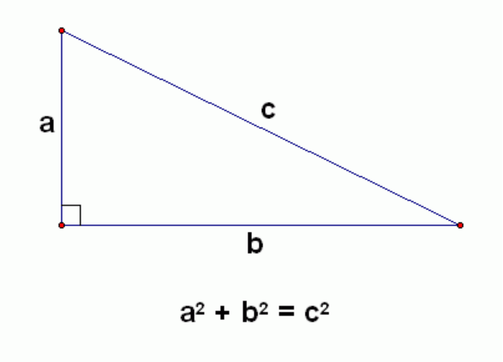
(A degree of similarity)

Euclidian distance

(feat. Pythagoras rule)



$$dist(p,q) = \sqrt{3^2 + 2^2}$$

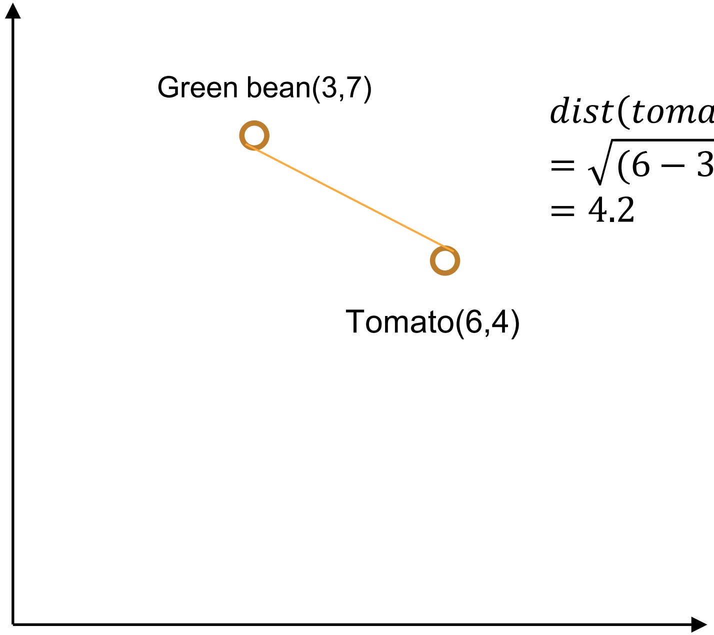


$$a^2 + b^2 = c^2$$

$$dist(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2}$$

Euclidian distance

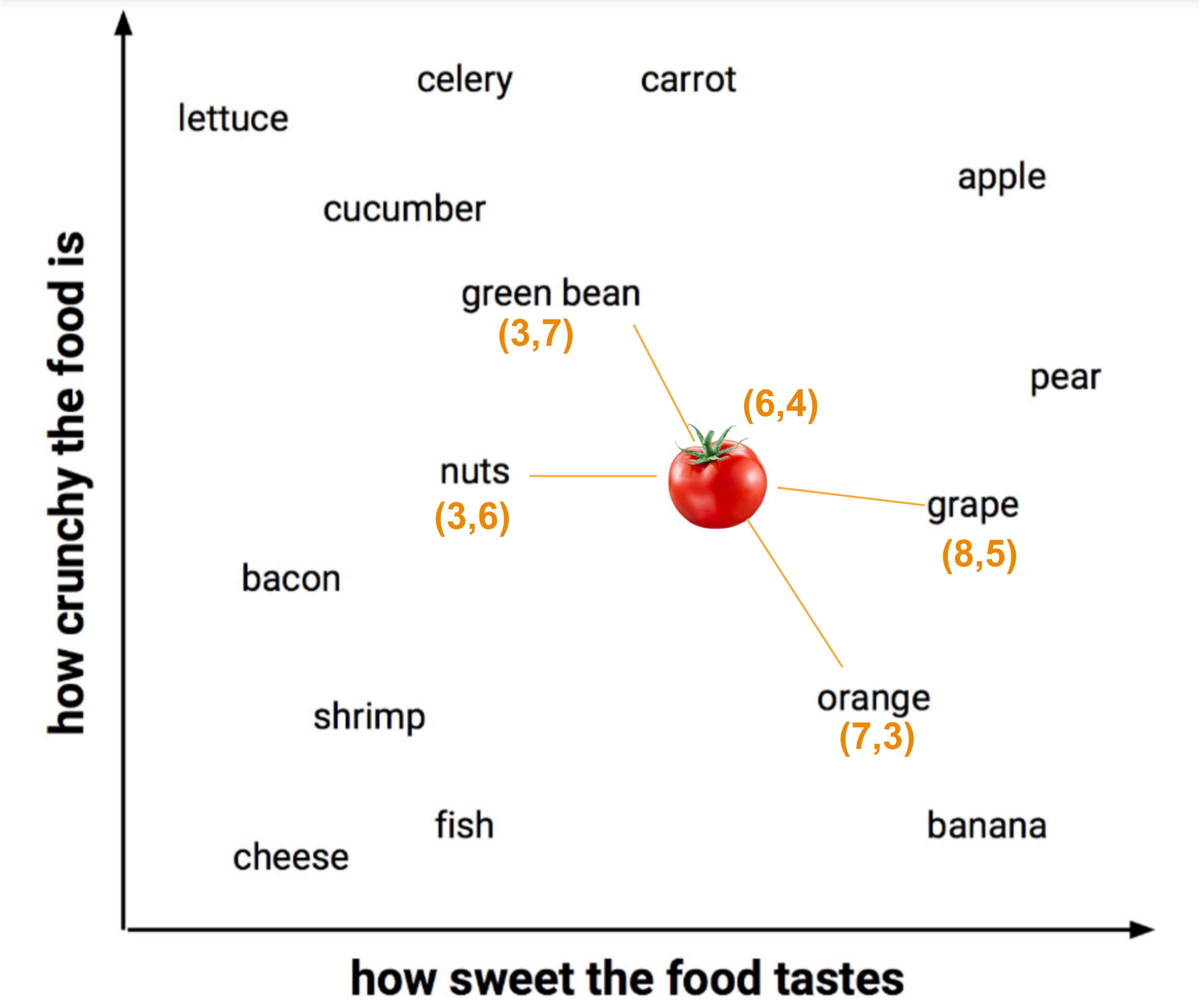
(feat. Pythagoras rule)



$$dist(tomato, green\_bean)$$
$$= \sqrt{(6-3)^2 + (4-7)^2}$$
$$= 4.2$$

$$dist(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2}$$

Distance to Tomato



Euclidian distance

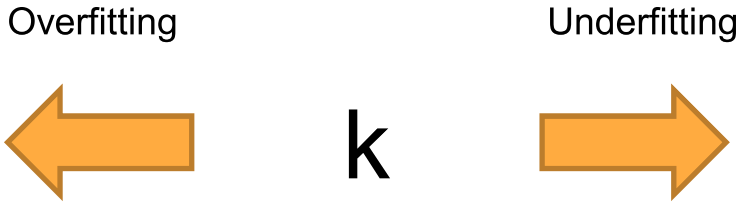
(feat. Pythagoras rule)

Ingredients	Sweat	Crunchy	Type	Distance to Tomato
Grape	8	5	Fruit	$\sqrt{(6-8)^2+(4-5)^2}= 2.2$
Green bean	3	7	Vegetable	$\sqrt{(6-3)^2+(4-7)^2}= 4.2$
Nuts	3	6	Protein	$\sqrt{(6-3)^2+(4-6)^2}= 3.6$
Orange	7	3	Fruit	$\sqrt{(6-7)^2+(4-3)^2}= 1.4$

- 1NN
- 3NN

How to choose the number of neighbours (k) ?

How to choose the number of neighbours (k) ?

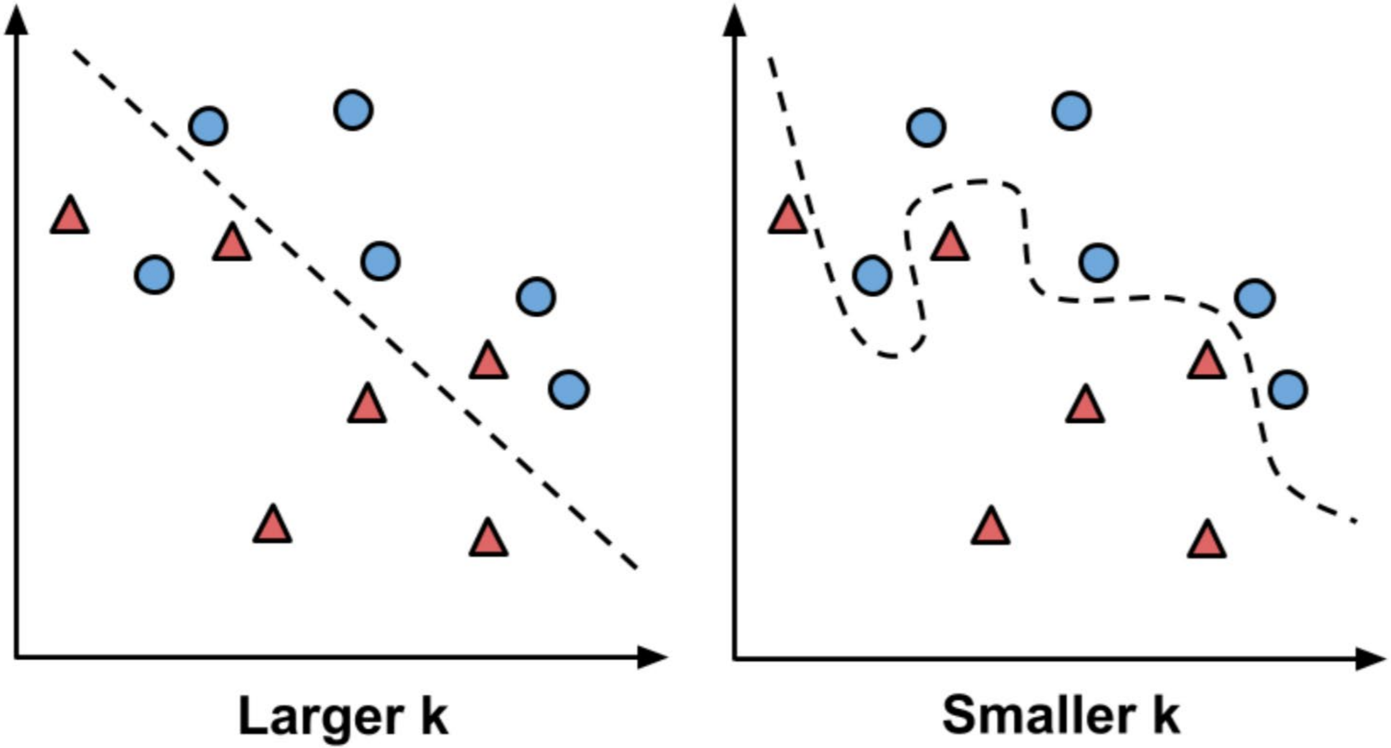


Overfitting

k

Underfitting

How to choose the number of neighbours (k) ?



Larger k

Smaller k

Feature standardization

1. Min-max normalization

$$X_{new} = \frac{X - \min(X)}{\max(X) - \min(X)}$$

2. Z-score standardization

$$X_{new} = \frac{X - \mu}{\sigma} = \frac{X - Mean(X)}{SD(X)}$$

Pros and Cons of the k-NN classifier

Advantages	Disadvantages
• Simple and efficient • No assumption on distribution of the underlying data • Fast training	• No model: difficult to understand the relationship between IVs and DV • Need to choose the right ‘k’ • Slow classification • Additional processing is required for nominal features and missing data