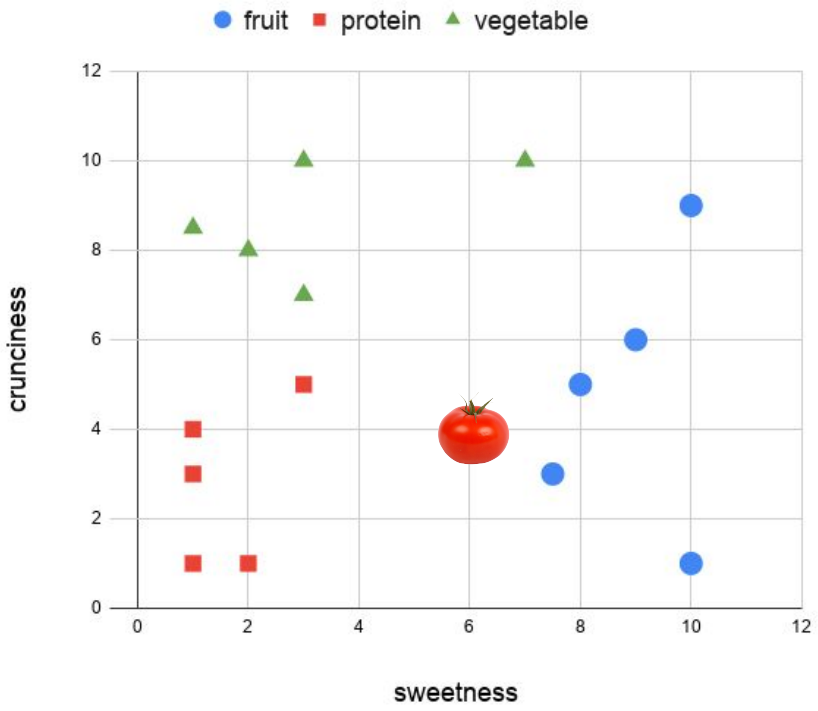
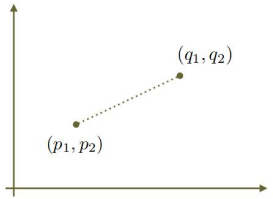
KNN - Exercise 1 - Solution



- Euclidean distance
  2-dimension

$$D(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2}$$

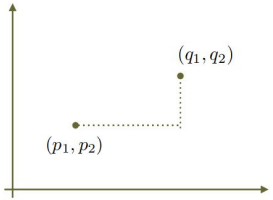


- N-dimension

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

- Manhattan distance
  2-dimension

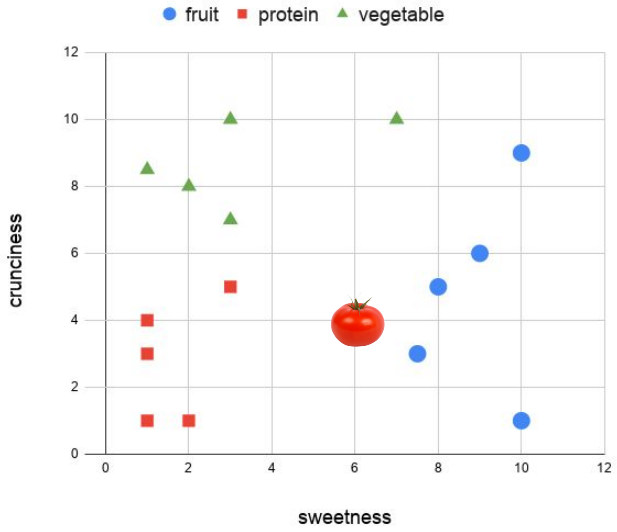
$$D(p,q) = |p_1 - q_1| + |p_2 - q_2|$$



- N-dimension

$$D(p,q) = |p_1 - q_1| + |p_2 - q_2| + \cdots + |p_n - q_n|$$

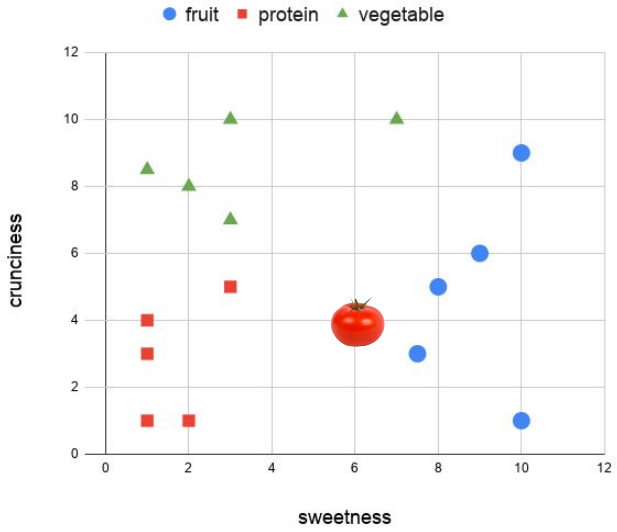
KNN - Exercise 1 - Solution



				from tomato (6,4)	
	sweetness	crunciness	class	L2 distances	L1 distances
apple	10	9	fruit	6,40	9,00
bacon	1	4	protein	5,00	5,00
banana	10	1	fruit	5,00	7,00
carrot	7	10	vegetable	6,08	7,00
celery	3	10	vegetable	6,71	9,00
cheese	1	1	protein	5,83	8,00
green bean	3	7	vegetable	4,24	6,00
grape	8	5	fruit	2,24	3,00
orange	7,5	3	fruit	1,80	2,50
pear	9	6	fruit	3,61	5,00
nuts	3	5	protein	3,16	4,00
shrimp	1	3	protein	5,10	6,00
fish	2	1	protein	5,00	7,00
lettuce	1	8,5	vegetable	6,73	9,50
cucumber	2	8	vegetable	5,66	8,00

L2: k=1 fruit, k=3 (fruit, fruit, protein): fruit
k=5 (fruit, fruit, protein, fruit, vegetable): fruit

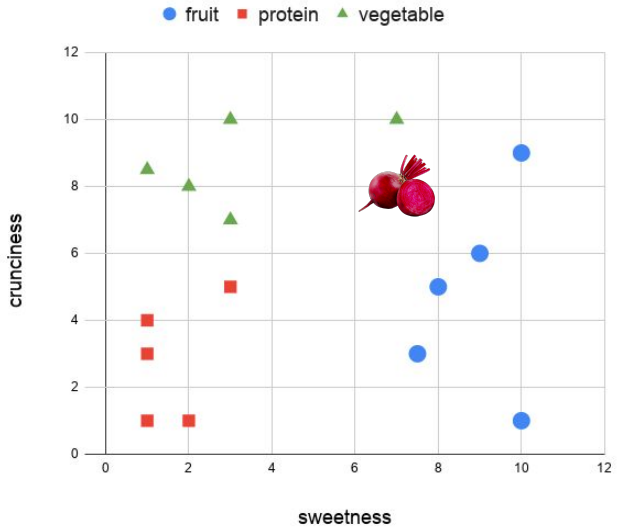
KNN - Exercise 1 - Solution



				from tomato (6,4)	
	sweetness	crunciness	class	L2 distances	L1 distances
apple	10	9	fruit	6,40	9,00
bacon	1	4	protein	5,00	5,00
banana	10	1	fruit	5,00	7,00
carrot	7	10	vegetable	6,08	7,00
celery	3	10	vegetable	6,71	9,00
cheese	1	1	protein	5,83	8,00
green bean	3	7	vegetable	4,24	6,00
grape	8	5	fruit	2,24	3,00
orange	7,5	3	fruit	1,80	2,50
pear	9	6	fruit	3,61	5,00
nuts	3	5	protein	3,16	4,00
shrimp	1	3	protein	5,10	6,00
fish	2	1	protein	5,00	7,00
lettuce	1	8,5	vegetable	6,73	9,50
cucumber	2	8	vegetable	5,66	8,00

L1: k=1 fruit, k=3 (fruit, fruit, protein): fruit
k=5 (fruit, fruit, protein, fruit, protein): fruit

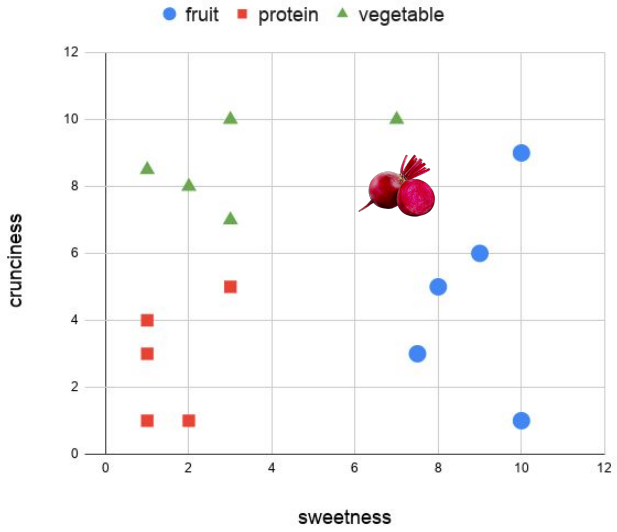
KNN - Exercise 2 - Solution



	sweetness	crunciness	class	beetroots (6.5,8) L2 distances
apple	10	9	fruit	3,640
bacon	1	4	protein	6,801
banana	10	1	fruit	7,826
carrot	7	10	vegetable	2,062
celery	3	10	vegetable	4,031
cheese	1	1	protein	8,902
green bean	3	7	vegetable	3,640
grape	8	5	fruit	3,354
orange	7,5	3	fruit	5,099
pear	9	6	fruit	3,202
nuts	3	5	protein	4,610
shrimp	1	3	protein	7,433
fish	2	1	protein	8,322
lettuce	1	8,5	vegetable	5,523
cucumber	2	8	vegetable	4,500

L2: k=5 (vegetable, fruit, fruit, vegetable, fruit): fruit….uhm...

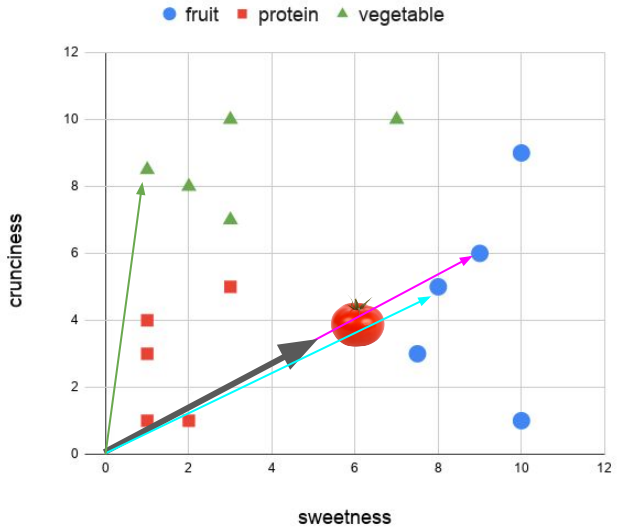
KNN - Exercise 2 - Solution



	sweetness	crunciness	class	beetroots (6.5,8) L2 distances
apple	10	9	fruit	3,640
bacon	1	4	protein	6,801
banana	10	1	fruit	7,826
carrot	7	10	vegetable	2,062
celery	3	10	vegetable	4,031
cheese	1	1	protein	8,902
green bean	3	7	vegetable	3,640
grape	8	5	fruit	3,354
orange	7,5	3	fruit	5,099
pear	9	6	fruit	3,202
nuts	3	5	protein	4,610
shrimp	1	3	protein	7,433
fish	2	1	protein	8,322
lettuce	1	8,5	vegetable	5,523
cucumber	2	8	vegetable	4,500

L2: k=5 (vegetable, fruit, fruit, vegetable, fruit)
average dist. v = 2,851 average dist. f = 3,399 ➡️ vegetable

KNN - Exercise 3 - Solution



	sweetness	crunciness	class	from tomato (6,4) cosine similarity
apple	10	9	fruit	0,9895
bacon	1	4	protein	0,7399
banana	10	1	fruit	0,8831
carrot	7	10	vegetable	0,9316
celery	3	10	vegetable	0,7704
cheese	1	1	protein	0,9806
green bean	3	7	vegetable	0,8376
grape	8	5	fruit	0,9996
orange	7,5	3	fruit	0,9785
pear	9	6	fruit	1,0000
nuts	3	5	protein	0,9037
shrimp	1	3	protein	0,7894
fish	2	1	protein	0,9923
lettuce	1	8,5	vegetable	0,6481
cucumber	2	8	vegetable	0,7399

Maximal cosine similarity. Nearest neighbour assigns class **fruit**

KNN - Exercise 4 - Solution

Suppose you want to build a nearest neighbors classifier to predict whether a beverage is a coffee or a tea using two features: the volume of the liquid (in milliliters) and the caffeine content (in grams). You collect the following data:

volume (ml)	caffeine (g)	label	**L2 from (120; 0,013)**
238	0,026	tea	118,00
100	0,011	tea	20,00
120	0,040	coffee	0,03
237	0,095	coffee	117,00

1. What is the label for a test point with Volume = 120, Caffeine = 0.013? (k=1, L2 distance)
   Coffee

2. Why your correct answer may still be wrong?
   One feature (Volume) dominate the distance

3. How would you fix the problem?

KNN - Exercise 5 - Solution

Suppose you want to build a nearest neighbors classifier to predict whether a beverage is a coffee or a tea using two features: the volume of the liquid (in milliliters) and the caffeine content (in grams). You collect the following data:

volume (ml)	caffeine (g)	label	**L2 from (120; 0,013)**
238	0,026	tea	118,00
100	0,011	tea	20,00
120	0,040	coffee	0,03
237	0,095	coffee	117,00

g / ml
0,00011
0,00011
0,00033
0,00040

Query: (120; 0,013)
corresponds to 0,00011 g/ml

1. What is the label for a test point with Volume = 120, Caffeine = 0.013? (k=1, L2 distance)
   Coffee

2. Why your correct answer may still be wrong?
   One feature (Volume) dominate the distance

3. How would you fix the problem?

KNN - Exercise 5 - Solution

Suppose you want to build a nearest neighbors classifier to predict whether a beverage is a coffee or a tea using two features: the volume of the liquid (in milliliters) and the caffeine content (in grams). You collect the following data:

volume (ml)	caffeine (g)	label	L2 from (120; 0,013)
238	0,026	tea	118,00
100	0,011	tea	20,00
120	0,040	coffee	0,03
237	0,095	coffee	117,00

173,75	0,043	mean
74,0647	0,0366	stdv

			L2 from (-0,726; -0,819)
0,867484	-0,464058		1,63
-0,995751	-0,873522		0,28
-0,725717	-0,081893		0,74
0,853983	1,419473		2,74

1. What is the label for a test point with Volume = 120, Caffeine = 0.013? (k=1, L2 distance)
   Coffee

2. Why your correct answer may still be wrong?
   One feature (Volume) dominate the distance

3. How would you fix the problem?
   Rescale the features to zero mean and unit variance (Z-score normalization)

KNN vs Perceptron - Test Solutions

KNN
assumption: similar/close samples have the same label

the 'learning' process stores….all the training samples

are there issues related to data dimensionality? yes

is it a multiclass classifier? yes

generalization guarantee: when the number of samples (n) goes to ∞ ….the theoretical error (risk) remains lower than twice the bias error (risk)

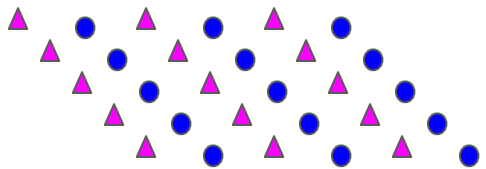
Perceptron
assumption: the data are linearly separable (i.e. there exist one hyperplane separating positive and negative samples without errors)

the learning process stores….the vector [w,b]

are there issues related to data dimensionality? no

is it a multiclass classifier? no - only binary problems

can we prove a guarantee when n goes to ∞ ….? no



Perceptron - Exercise 2 - Solution

$$y_i(\mathbf{w}^\top \mathbf{x}_i) > 0 \iff \mathbf{x}_i \text{ is classified correctly}$$

$$(4,3,6)^T \in \mathcal{N}, \quad (2,-2,3)^T \in \mathcal{P}, \quad (1,0,-3)^T \in \mathcal{P}, \quad (4,2,3)^T \in \mathcal{N}$$

pattern	output	classification	update	new weight vector
				$(1,0,0,0)^T$
$(1,4,3,6)^T \in \mathcal{N}$	$f_{step}(1)$	false positive	$-(1,4,3,6)^T$	$(0,-4,-3,-6)^T$
$(1,2,-2,3)^T \in \mathcal{P}$	$f_{step}(-20)$	false negative	$+(1,2,-2,3)^T$	$(1,-2,-5,-3)^T$
$(1,1,0,-3)^T \in \mathcal{P}$	$f_{step}(8)$	true positive	unchanged	unchanged
$(1,4,2,3)^T \in \mathcal{N}$	$f_{step}(-26)$	true negative	unchanged	unchanged
$(1,4,3,6)^T \in \mathcal{N}$	$f_{step}(-40)$	true negative	unchanged	unchanged
$(1,2,-2,3)^T \in \mathcal{P}$	$f_{step}(-2)$	false negative	$+(1,2,-2,3)^T$	$(2,0,-7,0)^T$
$(1,1,0,-3)^T \in \mathcal{P}$	$f_{step}(2)$	true positive	unchanged	unchanged
$(1,4,2,3)^T \in \mathcal{N}$	$f_{step}(-12)$	true negative	unchanged	unchanged
$(1,4,3,6)^T \in \mathcal{N}$	$f_{step}(-19)$	true negative	unchanged	unchanged
$(1,2,-2,3)^T \in \mathcal{P}$	$f_{step}(16)$	true positive	unchanged	unchanged

finished, weight vector $(2,0,-7,0)^T$ classifies all patterns correctly

Perceptron - Exercise 3 - Solution

$$(1,1)^T \in \mathcal{P}, \quad (1,0)^T \in \mathcal{N}, \quad (0,0)^T \in \mathcal{P}, \quad (0,1)^T \in \mathcal{N}$$

pattern	output	classification	update	new weight vector
				$(1,0,0)^T$
$(1,1,1)^T \in \mathcal{P}$	$f_{step}(1)$	true positive	unchanged	unchanged
$(1,1,0)^T \in \mathcal{N}$	$f_{step}(1)$	false positive	$-(1,1,0)^T$	$(0,-1,0)^T$
$(1,0,0)^T \in \mathcal{P}$	$f_{step}(0)$	true positive	unchanged	unchanged
$(1,0,1)^T \in \mathcal{N}$	$f_{step}(0)$	false positive	$-(1,0,1)^T$	$(-1,-1,-1)^T$
$(1,1,1)^T \in \mathcal{P}$	$f_{step}(-3)$	false negative	$+(1,1,1)^T$	$(0,0,0)^T$
$(1,1,0)^T \in \mathcal{N}$	$f_{step}(0)$	false positive	$-(1,1,0)^T$	$(-1,-1,0)^T$
$(1,0,0)^T \in \mathcal{P}$	$f_{step}(-1)$	false negative	$+(1,0,0)^T$	$(0,-1,0)^T$

finished, weight vector $(0,-1,0)^T$ occurs twice