kNN Classification

Problem 1: You are given the following dataset, with points of two different classes:

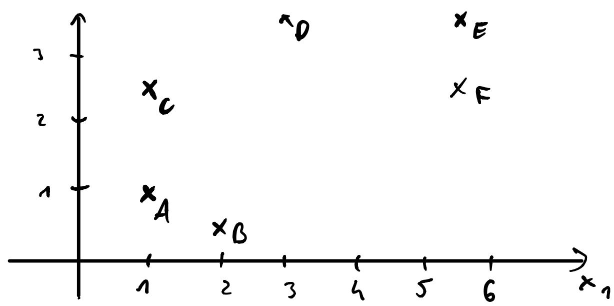
Name	x_1	x_2	class
A	1.0	1.0	1
B	2.0	0.5	1
C	1.0	2.5	1
D	3.0	3.5	2
E	5.5	3.5	2
F	5.5	2.5	2

We perform 1-NN classification with leave-one-out cross validation on the data in the plot.

a) Compute the distance between each point and its nearest neighbor using L_1-norm as distance measure.

b) Compute the distance between each point and its nearest neighbor using L_2-norm as distance measure.

c) What can you say about classification if you compare the two distance measures?



a) L_1-Norm: $\|x\|_1 = \sum_{i=1}^{n} |x_i|$ $\qquad d(x_1, x_2) = |x_1 - x_2| = \sum_{i=1}^{n} |x_{1i} - x_{2i}|$

$d_1(A,B) = 1.5 \qquad d_1(B,A) = 1.5 \qquad d_1(C,A) = 1.5$

$d_1(D,E) = 2.5 \qquad d_1(E,F) = 1 \qquad d_1(F,E) = 1$

b) L_2-Norm: $\|x\|_2 = \sqrt{\sum_{i=1}^{n} |x_i|^2}$ $\qquad d_2(x_1, x_2) = \sqrt{|x_1 - x_2|^2} = \sqrt{\sum_{i=1}^{n} |x_1 - x_2|^2}$

$d_2(A,B) \approx 1.12 \qquad d_2(B,A) \approx 1.12 \qquad d_2(C,A) \approx 1.5$

$d_2(D,C) \approx 2.2 \qquad d_2(E,F) = 1 \qquad d_2(F,E) = 1$

c) The classes are the same but for D. Here the L1 distance is smallest to point E whereas with L2 the closest point is C.

$\boxed{L_1}$

A

B: $|(1,1)-(2,0.5)| = 1+0.5 = 1.5$

C: $|(1,1)-(1,2.5)| = 0+1.5 = 1.5$

D: $|(1,1)-(3,0.5)| = 2+2.5 = 4.5$

E: $|(1,1)-(5.5,3.5)| = 4.5+2.5 = 7$

F: $|(1,1)-(5.5,2.5)| = 4.5+1.5 = 6$

B

C: $|(2,0.5)-(1,2.5)| = 1+2 = 3$

D: $|(2,0.5)-(3,3.5)| = 1+3 = 4$

E: $|(2,0.5)-(5.5,3.5)| = 3.5+3 = 6.5$

F: $|(2,0.5)-(5.5,2.5)| = 3.5+2 = 6.5$

C

D: $|(1,2.5)-(3,3.5)| = 2+1 = 3$

E: $|(1,2.5)-(5.5,3.5)| = 4.5+1 = 5.5$

F: $|(1,2.5)-(5.5,2.5)| = 4.5+0 = 4.5$

D

E: $|(3,3.5)-(5.5,3.5)| = 2.5+0 = 2.5$

F: $|(3,3.5)-(5.5,2.5)| = 2.5+1 = 3.5$

E

F: $|(5.5,3.5)-(5.5,2.5)| = 0+1 = 1$

$\boxed{L2}$

A

B: $|(1,1)-(2,0.5)|^2 = \sqrt{1^2+0.5^2} \approx 1.12$

C: $|(1,1)-(1,2.5)|^2 = \sqrt{0^2+1.5^2} \approx 1.5$

D: $|(1,1)-(3,3.5)|^2 = \sqrt{2^2+2.5^2} \approx 3.4$

E: $|(1,1)-(5.5,3.5)|^2 = \sqrt{4.5^2+2.5^2} \approx 5.2$

F: $|(1,1)-(5.5,2.5)|^2 = \sqrt{4.5^2+1.5^2} \approx 4.8$

B

C: $|(2,0.5)-(1,2.5)|^2 \approx 2.4$

D: $|(2,0.5)-(3,3.5)|^2 \approx 3.3$

E: $|(2,0.5)-(5.5,3.5)|^2 \approx 4.7$

F: $|(2,0.5)-(5.5,2.5)|^2 \approx 4.2$

C

D: $|(1,2.5)-(3,3.5)|^2 \approx 2.2$

E: $|(1,2.5)-(5.5,3.5)|^2 \approx 4.7$

F: $|(1,2.5)-(5.5,2.5)|^2 \approx 4.6$

D

E: $|(3,3.5)-(5.5,3.5)|^2 \approx 2.5$

F: $|(3,3.5)-(5.5,2.5)|^2 \approx 2.7$

E

F: $|(5.5,3.5)-(5.5,2.5)|^2 \approx 1$

Problem 2: Consider a dataset with 3 classes $\mathcal{C} = \{A, B, C\}$, with the following class distribution $N_A = 16, N_B = 32, N_C = 64$. We use unweighted k-NN classifier, and set k to be equal to the number of data points, i.e. $k = N_A + N_B + N_C =: N$.

a) What can we say about the prediction for a new point x_{new}?

b) How about if we use the weighted (by distance) version of k-Nearest Neighbors?

$N = 16 + 32 + 64 = 112 = k$

a) $p(y = A | x, N) = \frac{16}{112}$

$p(y = B | x, N) = \frac{32}{112}$

$p(y = C | x, N) \quad \frac{64}{112}$ ← k will be classified as C since it is the largest class.

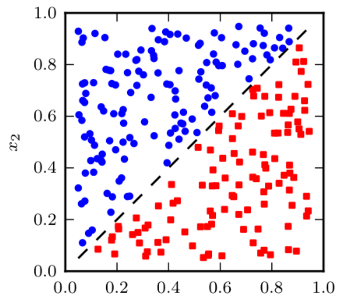
b) $Z = \sum_{i \in N_k(x)} \frac{1}{d(x, x_i)}$ $\qquad p(y = c | x, N) = \frac{1}{Z} \sum_{i \in N_k(x)} \frac{1}{d(x, x_i)} \mathbb{I}(y_i = c)$

$p(y = A | x, N)$

x_{new} will be predicted as C as well we consider all 112 nearest neighbors. The distance does not play a role when $k = N$.

Classification is based on the weights. Since those are not given we cannot make any class label prediction

Problem 3: The plot below shows data of two classes that can easily be separated by a single (diagonal) line. Does there exist a decision tree of depth 1 that classifies this dataset with 100% accuracy? Justify your answer.



The data cannot be split diagonally but only vertically and horizontally.
A diagonal cut touches both the x_1 and x_2 dimension.
With this data, a cut can only be accomplished with a zig-zag pattern

Problem 4: You are developing a model to classify games at which machine learning will beat the world champion within five years. The following table contains the data you have collected.

No.	x_1 (Team or Individual)	x_2 (Mental or Physical)	x_3 (Skill or Chance)	y (Win or Lose)
1	T	M	S	W
2	I	M 4	S	W 4
3	T 5	P	S 5	W
4	I	P	C	W
5	T	P	C	L
6	I	M	C	L
7	T	M	S	L
8	I	P	S	L
9	T	P	C	L
10	I	P	C	L

a) Calculate the entropy $i_H(y)$ of the class labels y.

b) Build the optimal decision tree of depth 1 using entropy as the impurity measure.

$i_H(t) = -\sum_{c_i \in C} \pi_{c_i} \log \pi_{c_i}$ $\quad (\lim_{x \to 0^+} x \log x = 0) \quad$ $\pi_c = p(y=c \mid t)$

$C = \{W, L\}$ $\quad p(y=W) = \frac{4}{10}$ $\quad p(y=L) = \frac{6}{10}$

a) $i_H(y) = p(y=W) \log p(y=W) + p(y=L) \log p(y=L)$

$= -\left(\frac{4}{10} \cdot \log \frac{4}{10} + \frac{6}{10} \cdot \log \frac{6}{10}\right)$

≈ 0.673012

b) Split on x_1:

$p(y=W \mid x_1 = T) = \frac{2}{5}$ $\quad p(y=W \mid x_1 = I) = \frac{2}{5}$

$i_H(x_1 = T) = i_H(x_1 = I) = \frac{2}{5} \cdot \log \frac{2}{5} + \frac{3}{5} \cdot \log \frac{3}{5} \approx 0.673$

$\Delta_{x_1} = 1 - \frac{5}{10} \cdot 0.673 - \frac{5}{10} \cdot 0.673 \approx 0.327$

Split on x_2:

$p(y=W \mid x_2 = M) = \frac{2}{4}$ $\quad p(y=W \mid x_2 = P) = \frac{2}{6}$

$i_H(x_2 = M) = \frac{2}{4} \log \frac{2}{4} + \frac{2}{4} \log \frac{2}{4} \approx 0.693$

$i_H(x_2 = P) = \frac{2}{6} \log \frac{2}{6} + \frac{4}{6} \log \frac{4}{6} \approx 0.637$

$\Delta_{x_2} = 1 - \frac{4}{10} \cdot 0.693 - \frac{6}{10} \cdot 0.637 = 0.340$

Split on x_3:

$p(y=W \mid x_3=S) = \frac{3}{5}$ $\qquad$ $p(y=W \mid x_3=C) = \frac{1}{5}$

$i_H(x_3=S) = \frac{3}{5} \cdot \log \frac{3}{5} + \frac{2}{5} \cdot \log \frac{2}{5} = i_H(x_1=T) \approx 0.673$

$i_H(x_3=C) = \frac{1}{5} \cdot \log \frac{1}{5} + \frac{4}{5} \cdot \log \frac{4}{5} \approx 0.5$

$\Delta_{x_3} = 1 - \frac{4}{10} \cdot 0.673 - \frac{6}{10} \cdot 0.5 \approx 0.431$

Split on x_3 (highest value)!

Programming Task

Problem 5: Load the notebook `exercise_02_notebook.ipynb` from Piazza. Fill in the missing code and run the notebook. Convert the evaluated notebook to HTML using nbconvert and add it to your submission.

Note: We suggest that you use Anaconda for installing Python and Jupyter, as well as for managing packages. We recommend that you use Python 3.

For more information on Jupyter notebooks, consult the Jupyter documentation. Instructions for converting the Jupyter notebooks to PDF are provided within the notebook.