Name: _____

ML Problem Solving 1 01/14/2025

Problem 1. Assume, you work with a taxi system. The dataset consists of various columns that capture different aspects of the clients. Classify every column type (binary, numerical, categorical: ordinal and nominal, textual, datetime, geospatial, audio, etc.)

How many trips during a day	Average expenses on taxi	How often premium was used	Age	Agreed to upgrade to premium
1	6	15	27	yes
2	7	2	54	no
...	...	...	...	...

Problem 2. Consider the data about a new user with the following information available:

How many trips during a day	Average expenses on taxi	How often premium was used	Age	Agreed to upgrade to premium
3	10	0	25	

Based on the *compactness principle,* use ℓ_1-distance to predict if this user is likely to upgrade the trip. In other words, calculate the distance between these data and each record in the table from Problem 1. Based on the shortest distance, predict whether the user is likely to agree to an upgrade.

Problem 3. Consider now slightly different data (in one cell):

How many trips during a day	Average expenses on taxi	How often premium was used	Age	Agreed to upgrade to premium
1	6	100	27	yes
2	7	2	54	no
...	...	...	...	...

How the prediction for the passenger from Problem 2 will be changed? Does it seem reasonable? If not, how could you change the distance function to get more reasonable prediction?

Problem 1. Apply the kNN method with 2 neighbors to the following data.

x	0.45	-0.1	2	0.3	-0.5	0.7	0
y	$+1$	-1	$+1$	$+1$	-1	-1	$+1$

Report the accuracy.

Hint: Order the data by x first. Consider the current point unlabeled. In cases of ambiguity choose the closest point, if the points (with different labels) are equidistant, return 0 as a refusal of classification.