Kennesaw State University

CS 7357 Neutral Nets and Deep Learning

Project 1

Description

In this project, we will implement KNN using Python. You are given three data sets*:

data type	have a label
training set	Yes
validation set	Yes
test set	No

**A typical division is that the training set accounts for 50% of the total sample, and the others account for 25%, and all of them are randomly selected from the sample.*

Dataset:

Document number	The sentence words	Emotion
Train 1	I buy an apple phone	happy
Train 2	I eat the gig apple	happy
Train 3	The Apple products are too expensive	sad
Train 4	My friend has an apple	?

(1) use KNN for classification problems, here we select Manhattan distance model. When calculating distance (x_i, x_j) for a data x_i, do not include the distance to itself as it's always 0.

(1.1) You can use one-hot matrix to represent the sentences. Please use all the data sets to create the dictionary.

(1.2) Adjust the K value ($3 \leq k \leq \sqrt{N}$, here N denotes the size of data set), measure the KNN predict accuracy on both training set, and validation set. On the verification set, find the value of K with highest accuracy.

Example:

```
k = 5 Validation set correct rate: 0.27009646302250806
k = 7 Validation set correct rate: 0.2604501607717042
k = 9 Validation set correct rate: 0.26688102893890675
k = 11 Validation set correct rate: 0.29260450160771706
```

P.S., Distance formula:
Manhattan distance: In a plane with p1 at (x1, y1) and p2 at (x2, y2), it is $|x_1 - x_2| + |y_1 - y_2|$.
Euclidean distance: In a plane with p1 at (x1, y1) and p2 at (x2, y2), it is $\sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$.

(2) Apply the KNN model obtained in step 1 on the test set, and save the output result as "my_result.csv".

Example:

my_result

Words (split by space)	label
senator carl krueger thinks ipods can kill you	joy
who is prince frederic von anhalt	joy
prestige has magic touch	joy
study female seals picky about mates	joy
no e book for harry potter vii	joy
blair apologises over friendly fire inquest	fear

Submission

You have to submit the followings to D2L:

1. MS word file
   - Describe what you have done for the homework assignment.
2. Python source code file(s)
   - Must be well organized (comments, indentation, …)
   - You need to upload the "original python file (*.py)" and also its "PDF" version.
     o For the PDF file, you can just convert the source file to PDF. One way is to print the source file and save to "PDF".

You have to submit the files SEPERATELY. DO NOT compress into a ZIP file.