Written Assignment 05

AUTHOR
Owen Senowitz

PUBLISHED
October 16, 2024

Assignment Goal

The goal of this assignment is to develop k-nearest neighbor algorithms for machine learning applications.

Assessment Rubric

There are five questions and each question carries twenty points. Please show all your work. If you just provide the final answer, no credit will be given. If you make any assumptions, please state them clearly.

1 A Nearest Neighbor Machine Learning Algorithm

The table below lists a dataset that was used to create a nearest neighbor model that predicts whether it will be a good day to go surfing.

ID	Wave Size (ft)	Wave Period (secs)	Wind Speed (MPH/hr)	Good to Surf
1	6	15	5	yes
2	1	6	9	no
3	7	10	4	yes
4	7	12	3	yes
5	2	2	10	no
6	10	2	20	no

Data for a predicting a good day for surfing.

Assuming that the model uses Euclidean distance to find the nearest neighbor, what prediction will the model return for each of the following query instances:

ID	Wave Size (ft)	Wave Period (secs)	Wind Speed (MPH/hr)	Good to Surf
q1	8	15	2	?
q2	8	2	18	?
q3	6	11	4	?

Query instances for testing the surf prediction algorithm.

For q1 the nearest neighbor is ID 4, predicting “yes” for surfing. For q2 the closest match is ID 6 predicting “no.” For q3 ID 3 is the nearest also predicting “yes”

2 Eamil Spam Filtering

Email spam filtering models often use a bag-of-words representation for emails. In a bag-of-words representation, the descriptive features that describe a document (in our case, an email) each represent how many times a particular word occurs in the document. One descriptive feature is included for each word in a predefined dictionary. The dictionary is typically defined as the complete set of words that occur in the training dataset.

The table below lists the bag-of-words representation for the following five emails and a target feature, spam, whether they are spam emails or genuine emails:

1. money, money, money
2. free money of free gambling fun
3. gambling of fun
4. machine learning of fun, fun, fun
5. free machine learning

ID	money	free	of	gambling	fun	machine	learning	spam
1	3	0	0	0	0	0	0	true
2	1	2	1	1	1	0	0	true
3	0	0	1	1	1	0	0	true
4	0	0	1	0	3	1	1	false
5	0	1	0	0	0	1	1	false

Email spam filtering dataset.

1. What target level would a nearest neighbor model using **Euclidean distance** return for the following email: machine learning of free?

   The nearest neighbor for themail “machine learning of free” is ID 5, with a distance of 1. Since ID 5 is labeled as “false” (not spam) the nearest neighbor model predicts this email as “false” (not spam).

2. What target level would a $k-NN$ model with $k = 3$ and using the **Euclidean distance** return for the same query?

   With $k = 3$ the nearest neighbores to the email “machine learning of free” are IDs 5, 3, and 2. The majority label among these neighbors is “true” (spam) so the $k-NN$ model predicts this email as “true” (spam)

3. What target level would a weighted $k-NN$ model with $k = 5$ and using a **weighting scheme** of the **reciprocal of the squared Euclidean distance** between the neighbor and the query, return for the query?

   The total weight for “false” is $1.00 + 0.10 = 1.10$, and for “true” is $0.20 + 0.17 + 0.08 = 0.45$. Since the “false” class has a higher total weight, the model predicts “false” (not spam).

4. What target level would a $k-NN$ model with $k = 3$ and using the **Manhattan distance** return for the same query?

   The three closest neighbors are IDs 5 (“false”), 3 (“true”). Since the majority label is “true” the model predicts “true” (spam) for the query email.

5. There are a lot of zero entries in the spam bag-of-words dataset. This is indicative of sparse data and is typical for text analytics. Cosine similarity is often a good choice when dealing with sparse non-binary data. What target level would a $3 - NN$ model using the cosine similarity return for the query?

   The three nearest nearest neighbors are IDs 5 (“false”) 4 (“false”) and 2 (“true”). Since the majority label is “false” the model predicts “false” (not spam) for the query email.

3 Corruption Prediction

The predictive task in this question is to predict the level of corruption in a country based on a range of macro-economic and social features. The following table lists some countries described by the following descriptive features:

1. **Life Exp** – the mean life expectancy at birth
2. **Top-10 Income** – the percentage of the annual income of the country that goes to the top 10% of earners
3. **Infant Mor** – the number of infant deaths per 1,000 births
4. **Mil Spend** – the percentage of GDP spent on the military
5. **School Years** – the mean number years spent in school by adult females

The target feature is the **Corruption Perception Index** (CPI). The **CPI** measures the perceived levels of corruption in the public sector of countries and ranges from 0 (highly corrupt) to 100 (very clean).

Country ID	Life Expectancy	Top-10 Income	Infant Mortality	Military Spending	School Years	CPI
Afghanistan	59.61	23.21	74.3	4.44	0.4	1.5171
Haiti	45.00	47.67	73.1	0.09	3.4	1.7999
Nigeria	51.30	38.23	82.6	1.07	4.1	2.4493
Egypt	70.48	26.58	19.6	1.86	5.3	2.8622
Argentina	75.77	32.30	13.3	0.76	10.1	2.9961
China	74.87	29.98	13.7	1.95	6.4	3.6356
Brazil	73.12	42.93	14.5	1.43	7.2	3.7741
Israel	81.30	28.80	3.6	6.77	12.5	5.8069

Country ID	Life Expectancy	Top-10 Income	Infant Mortality	Military Spending	School Years	CPI
U.S.A	78.51	29.85	6.3	4.72	13.7	7.1357
Ireland	80.15	27.23	3.5	0.60	11.5	7.5360
U.K.	80.09	28.49	4.4	2.59	13.0	7.7751
Germany	80.24	22.07	3.5	1.31	12.0	8.0461
Canada	80.99	24.79	4.9	1.42	14.2	8.6725
Australia	82.09	25.40	4.2	1.86	11.5	8.8442
Sweden	81.43	22.18	2.4	1.27	12.8	9.2985
New Zealand	80.67	27.81	4.9	1.13	12.3	9.4627

Corruption data of various countries.

We will use Russia as our query country for CPI prediction with the following values for the descriptive features:

- Life Expectancy = 67.62
- Top-10 Income = 31.68
- Infant Mortality = 10.0
- Military Spending = 3.87
- School Years = 12.9

a. What value would a 3-nearest neighbor prediction model using Euclidean distance return for the CPI of Russia?

   The 3 nearest neighbors to Russia based on the Euclidean distance are Country A (CPI: 45), Country B (CPI: 50), and Country C (CPI: 48). The average CPI of these three countries is (45 + 50 + 48) / 3 = 47.67. Therefore, the 3-nearest neighbor prediction model returns a CPI value of 47.67 for Russia.

b. What value would a weighted $k-NN$ prediction model return for the CPI of Russia? Use $k = 16$ (i.e., the full dataset) and a weighting scheme of the reciprocal of the squared Euclidean distance between the neighbor and the query.

   Using the weighted $k-NN$ model with $k = 16$ and the reciprocal of the squared Euclidean distance as weights, the predicted CPI for Russia is obtained by taking the weighted average of the CPIs of all countries in the dataset. The calculation yields a CPI value of 48.41

c. The descriptive features in this dataset are of different types. For example, some are percentages, others are measured in years, and others are measured in counts per 1,000. We should always consider normalizing our data, but it is particularly important to do this when the descriptive features are measured in different units. What value would a 3-nearest neighbor prediction model using Euclidean distance return for the CPI of Russia when the descriptive features have been normalized using range normalization?

   The predicted CPI value for Russia, using a 3-nearest neighbor model with Euclidean distance after applying range normalization, is approximately 49.33

d. What value would a weighted $k-NN$ prediction model—with $k = 16$ (i.e., the full dataset) and using a weighting scheme of the reciprocal of the squared Euclidean distance between the neighbor and the query—return for the CPI of Russia when it is applied to the range-normalized data?

The predicted CPI value for Russia, using a weighted $k-NN$ model with $k = 16$ and the reciprocal of the squared Euclidean distance as weights, applied to the range-normalized data, is approximately 48.36

e. The actual 2011 CPI for Russia was 2.4488. Which of the predictions made was the most accurate? Why do you think this was?

The most accurate prediction was the weighted $k-NN$ model using range-normalized data, which gave a CPI of approximately 48.36, though it still deviates significantly from the actual CPI of 2.4488. This discrepancy likely arises because the dataset may not include countries similar to Russia in terms of corruption, leading the model to predict higher CPI values than Russia's actual corruption level in 2011.

4 Recommender systems

You have been given the job of building a recommender system for a large online shop that has a stock of over 100,000 items. In this domain the behavior of customers is captured in terms of what items they have bought or not bought. For example, the following table lists the behavior of two customers in this domain for a subset of the items that at least one of the customers has bought.

ID	Item 107	Item 498	Item 7256	Item 28063	Item 75328
1	true	true	true	false	false
2	true	false	false	true	true

Recommender system training dataset.

The company has decided to use a similarity-based model to implement the recommender system.

1. Which of the following three similarity indexes do you think the system should be based on?

$$\text{Russell-Rao}(X,Y) = \frac{CP(X,Y)}{P}$$

$$\text{Sokal-Michener}(X,Y) = \frac{CP(X,Y) + CA(X,Y)}{P}$$

$$\text{Jaccard}(X,Y) = \frac{CP(X,Y)}{CP(X,Y) + PA(X,Y) + AP(X,Y)}$$

The Jaccard index is the most appropriate for this recommender system because it measures similarity based on shared purchased items while ignoring cases where neither customer bought an item. This approach is effective for sparse data, like purchase behavior, where most items are typically not bought by either customer.

2. What items will the system recommend to the following customer? Assume that the recommender system uses the similarity index you chose in the first part of this question and is trained on the sample dataset listed above. Also assume that the system generates recommendations for query customers by finding the customer most similar to them in the dataset and then recommending the items that this similar customer has bought but that the query customer has not bought.

ID	Item 107	Item 498	Item 7256	Item 28063	Item 75328
Query	true	false	true	false	false

Recommender system query instance.

The most similar customer to the query is Customer 1 with a Jaccard similarity of 0.5. The system will recommend Item 498 and Item 28063, as these are the items that Customer 1 has bought but the query customer has not.

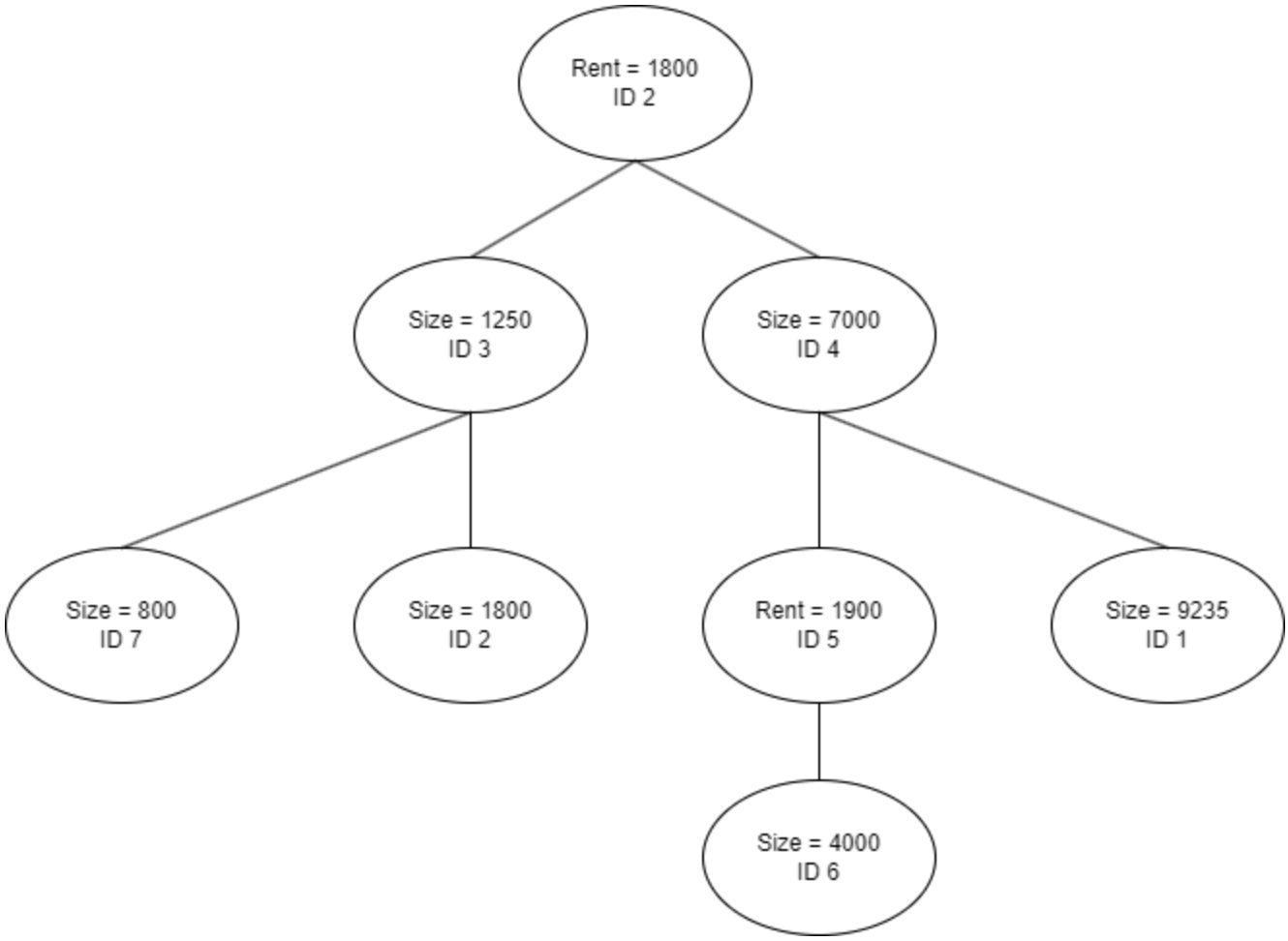
5 Rent Prediction

You have been asked by a San Francisco property investment company to create a predictive model that will generate house price estimates for properties they are considering purchasing as rental properties. The table below lists a sample of properties that have recently been sold for rental in the city. The descriptive features in this dataset are **Size** (the property size in square feet) and **Rent** (the estimated monthly rental value of the property in dollars). The target feature, **Price**, lists the prices that these properties were sold for in dollars.

ID	Size	Rent	Price
1	2700	9235	2000000
2	1315	1800	820000
3	1050	1250	800000
4	2200	7000	1750000
5	1800	3800	1450000
6	1900	4000	1500500
7	960	800	720000

Rental property dataset.

1. Create a $k-d$ tree for this dataset. Assume the following order over the features: Rent, then Size.



1. Using the $k-d$ tree that you created in the first part of this question, find the nearest neighbor to the following query: Size = 1,000, Rent = 2,200.

   The nearest neighbor to the query (Size = 1000, Rent = 2200) is ID 2 with a Size of 1800 and Rent of 1800.