DATA 245: Machine Learning

Homework - 3

Submitted By: Poojan Gagrani

SJSU ID: 016795285

Q1. 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: "money, money, money", "free money for free gambling fun", "gambling for fun", "machine learning for fun, fun, fun", "free machine learning" ?

	Bag-of-Words							
ID	MONEY	FREE	FOR	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

a) What target level would a nearest neighbor model using Euclidean distance return for the following email: "machine learning for free"?

Ans: The bag of words for the given query can be represented as following:

ID	MONEY	FREE	FOR	GAMBLING	FUN	MACHINE	LEARNING	SPAM
Query	0	1	1	0	0	1	1	?

Euclidean distance can be calculated by the following: **Euclidean Distance** = $\sqrt{\Sigma(a[i] - b[i])^2}$

Calculating the Euclidean Distance for ID = 1,

$ID_1 = \sqrt{9 + 1 + 1 + 0 + 0 + 1 + 1} = 3.605$

Similarly, Euclidean distance for all the ID is the following:

ID	MONEY	FREE	FOR	GAMBLING	FUN	MACHINE	LEARNING	EUCLIDEAN DISTANCE
1	9	1	1	0	0	1	1	3.605
2	1	1	0	1	1	1	1	2.449
3	0	1	0	1	1	1	1	2.236
4	0	1	0	0	9	0	0	3.162
5	0	0	1	0	0	0	0	1

By performing the Euclidean Distance, the query distance is the nearest neighbor to instance d5 = 1 for which the SPAM value is false. Hence, the model will predict **SPAM = false.**

b) What target level would a k-NN model with k=3 and using Euclidean distance return for the same query?

Ans: For k=3 based on the distance calculations performed in the above question the nearest neighbors are d5=1, d3=2.236, d2= 2.449 and two of them have SPAM = true. Hence, the 3-NN model will return the prediction of **SPAM = true**.

c) 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?

Ans: Weighted k-NN can be calculated by the following: $\textbf{Weighted K-NN} = 1/(\sqrt{\Sigma(a[i] - b[i])^2})^2$

Calculating the Weighted Distance for ID = 1,

$ID_1 = 1 / (3.605)^2 = 0.076$

Similarly, Weighted distance for all the ID is the following:

ID	Weights	SPAM
1	0.076	true
2	0.166	true
3	0.2	true
4	0.1	true
5	1	false

For SPAM = true target level is 0.0769 + 0.1667 + 0.2 = 0.443. The total weight for the SPAM = false target level is 0.1 + 1 = 1.1. Hence, the SPAM = false has the maximum weight, and thus the prediction returned by the model is **SPAM = false**.

d) What target level would a k-NN model with k = 3 and using Manhattan distance return for the same query?

Ans: k-NN using Manhattan distance can be calculated by the following: **Manhattan distance = Σ abs(a[i] - b[i])**

ID	MONEY	FREE	FOR	GAMBLING	FUN	MACHINE	LEARNING	MANHATTAN DISTANCE
1	3	1	1	0	0	1	1	7
2	1	1	0	1	1	1	1	6
3	0	1	0	1	1	1	1	5
4	0	1	0	0	3	0	0	4
5	0	0	1	0	0	0	0	1

The three nearest neighbors to the query are the instances d5, d4, d3 and two of them have SPAM = FALSE. Hence, the model will predict **SPAM = false**.

e) 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

cosine similarity return for the query?

Ans: In order to calculate the cosine similarity, we need to calculate the vector length for each instance and the query. The vector length can be calculated as following: **Vector length** $=\sqrt{\Sigma a[i]^2}$

ID	$a[i]^2$	Sum	Vector Length
1	9 0 0 0 0 0 0	9	3
2	1 4 1 1 1 0 0	8	2.828
3	0 0 1 1 1 0 0	3	1.732
4	0 0 1 0 9 1 1	12	3.464
5	0 1 0 0 0 1 1	3	1.732
Query	0 1 1 0 0 1 1	4	2

Now we need to calculate the dot product between eac query and instance which can be calculate by the following: $\mathbf{a.b = \Sigma(a[i] \times b[i])}$

Pair	(a[i] x b[i])	DOT PRODUCT
(a,b1)	0 0 0 0 0 0 0	0
(a,b2)	0 2 1 0 0 0 0	3
(a,b3)	0 0 1 0 0 0 0	1
(a,b4)	0 0 1 0 0 1 1	3
(a,b5)	0 1 0 0 0 1 1	3

Finally, we can calculate the cosine similarity for each query-instance pair by the following $\mathbf{COSINE(a,b) = a.b / \sqrt{\Sigma a[i]^2} \times \sqrt{\Sigma b[i]^2}}$

Pair	Cosine Similarity
(a,b1)	0/3x2 =0
(a,b2)	3/2.828 x 2 = 0.5303
(a,b3)	1/1.732 x 2 = 0.2887
(a,b4)	3/3.464 x 2 = 0.4330
(a,b5)	3/1.732 x 2 = 0.8660

While using Cosine Similarity index, the higher number the more similar the instances. For 3-NN models the most similar instances are d5, d2 and d4. Here, d5 and d4 have SPAM = false, and hence the prediction model will return **SPAM = false**.

Q2. 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 table below lists some countries described by the following descriptive feature?:

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

Ans: Given Russia as a country for this question. The descriptive features are as following:

COUNTRY ID	LIFE EXP.	TOP-10 INCOME	INFANT MORT.	MIL. SPEND	SCHOOL YEARS	CPI
Russia	67.62	31.68	10.00	3.87	12.90	?

Euclidean distance can be calculated by the following: **Euclidean Distance** = $\sqrt{\Sigma(a[i] - b[i])^2}$

Now we can calculate the Euclidean distance for Country ID = Afghanistan

COUNTRY ID	LIFE EXP.	TOP-10 INCOME	INFANT MORT.	MIL. SPEND	SCHOOL YEARS	CPI
Afghanistan	59.61	23.21	74.30	4.44	0.40	1.5171

Euclidean Distance = √64.16+71.74+4134.49+0.32+156.25 = √4426.96 = 66.541.

The following table shows the countries in the dataset with their respective CPI values by ascending order of Euclidean distance from Russia.

Id	CPI	Euclidean (a , bi)
Argentina	2.9961	9.7805
China	3.6356	10.7898
U.S.A	7.1357	12.6033
Egypt	2.8622	13.7217
Brazil	3.7741	14.7394
U.K.	7.7751	15.0621
Israel	5.8069	16.0014
Ireland	7.5360	16.0490
New Zealand	9.4627	16.3806
Canada	8.6725	17.2765
Australia	8.8442	18.1472
Germany	8.0461	18.2352
Sweden	9.2985	19.8056
Afghanistan	1.5171	66.5419
Haiti	1.7999	69.6705
Nigeria	2.4493	75.2712

Hence, we can see that the three nearest neighbors to Russia are Argentina, China and the U.S.A. Hence, the CPI value returned will be the average CPI score of these 3 neighbors i.e., **(2.9961 + 3.6356 +7.1357) / 3 = 4.5891**

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?

Ans: Weighted k-NN can be calculated by the following: **Weighted K-NN = $1/(\sqrt{\Sigma}(a[i] - b[i])^2)^2$**

Hence, the weights for each of the instances in the dataset are:

Id	Euclidean (a,bi)	CPI	Weight	Weight * CPI
Argentina	9.7805	2.9961	0.0105	0.0313
China	10.7898	3.6356	0.0086	0.0312
U.S.A	12.6033	7.1357	0.0063	0.0449
Egypt	13.7217	2.8622	0.0053	0.0152
Brazil	14.7394	3.7741	0.0046	0.0174
U.K.	15.0621	7.7751	0.0044	0.0343
Israel	16.0014	5.8069	0.0039	0.0227
Ireland	16.0490	7.5360	0.0039	0.0293
New Zealand	16.3806	9.4627	0.0037	0.0353
Canada	17.2765	8.6725	0.0034	0.0291
Australia	18.1472	8.8442	0.0030	0.0269
Germany	18.2352	8.0461	0.0030	0.0242
Sweden	19.8056	9.2985	0.0025	0.0237
Afghanistan	66.5419	1.5171	0.0002	0.0003
Haiti	69.6705	1.7999	0.0002	0.0004
Nigeria	75.2712	2.4493	0.0002	0.0004
Sum Weight			0.0637	
Sum weight * CPI				0.3666

Hence, the value returned by the model is: **0.3665 / 0.0637 = 5.7507**

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?

Ans: We can perform the normalization using the following formula: **Range Normalization = (xi – min(x)) / (max(x) – min(x))**

First we need to normalize the descriptive features and CPI

Country Id	Life Exp	Top 10 income	Infant Mort	Mil. spend	School Years	CPI
Afghanistan	0.3940	0.0445	0.8965	0.6507	0.0000	1.5171
Haiti	0.0000	1.0000	0.8815	0.0000	0.2174	1.7999
Nigeria	0.1698	0.6313	1.0000	0.1384	0.2681	2.4493
Egypt	0.6869	0.1762	0.2145	0.2652	0.3551	2.8622
Argentina	0.8296	0.3996	0.1359	0.0963	0.7029	2.9961
China	0.8053	0.3090	0.1409	0.2786	0.4348	3.6356
Brazil	0.7582	0.8148	0.1509	0.2004	0.4928	3.7741
Israel	0.9785	0.2629	0.0150	1.0000	0.8768	5.8069
U.S.A	0.9034	0.3039	0.0486	0.6922	0.9638	7.1357
Ireland	0.9477	0.2016	0.0137	0.0757	0.8043	7.5360
U.K	0.9459	0.2508	0.0249	0.3749	0.9130	7.7751
Germany	0.9501	0.0000	0.0137	0.1818	0.8406	8.0461
Canada	0.9702	0.1063	0.0312	0.1996	1.0000	8.6725
Australia	1.0000	0.1301	0.0224	0.2651	0.8043	8.8442
Sweden	0.9821	0.0043	0.0000	0.1760	0.8986	9.2985
New Zealand	0.9617	0.2242	0.0312	0.1547	0.8623	9.4627

We also need to normalize the descriptive feature for Russia as well

Country Id	Life Exp	Top 10 income	Infant Mort	Mil. spend	School Years	CPI
Russia	0.6099	0.3754	0.0948	0.5658	0.9058	?

Now, we need to calculate the Euclidean distance between Russia and other countries in dataset,

Id	Euclidean (q,di)	CPI
Egypt	0.00004	2.8622
Brazil	0.00048	3.7741
China	0.00146	3.6356
Afghanistan	0.00217	1.5171
Argentina	0.00233	2.9961
U.S.A	0.00742	7.1357
U.K	0.01275	7.7751
Ireland	0.01302	7.5360
Germany	0.01531	8.0461
New Zealand	0.01531	9.4627
Israel	0.01685	5.8069
Sweden	0.01847	9.2985
Australia	0.01847	8.8442
Nigeria	0.02316	2.4493
Canada	0.03753	8.6725
Haiti	0.13837	1.7999

We can see that the USA, U.K and Argentina are the three nearest neighbors to Russia. Hence, the CPI value returned by the model: **7.1357 + 7.7751 + 2.9961 / 3 = 5.9689**

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?

Ans: Weighted k-NN can be calculated by the following: **Weighted K-NN = $1/(\sqrt{\Sigma(a[i] - b[i])^2})^2$**

Hence, the weights for each of the instances in the dataset are:

Id	Euclidean (q,di)	CPI	Weight	Weight*CPI
Afghanistan	1.2850	1.5171	0.6056	0.9187
Haiti	1.4733	1.7999	0.4607	0.8292
Nigeria	1.2910	2.4493	0.6000	1.4695
Egypt	0.6733	2.8622	2.2059	6.3137
Argentina	0.5495	2.9961	3.3118	9.9224
China	0.5936	3.6356	2.8380	10.3178
Brazil	0.7193	3.7741	1.9328	7.2945
Israel	0.5910	5.8069	2.8630	16.6251
U.S.A	0.3434	7.1357	8.4801	60.5114
Ireland	0.6329	7.5360	2.4965	18.8136
U.K	0.4130	7.7751	5.8627	45.5776
Germany	0.6429	8.0461	2.4194	19.4667
Canada	0.5834	8.6725	2.9381	25.4806
Australia	0.5676	8.8442	3.1040	27.5223
Sweden	0.6623	9.2985	2.2798	21.1987

New Zealand	0.5704	9.4627	3.0736	29.0845
Sum Weight			37.0439	
Sum Weight * CPI				240.8177

Hence the value returned by the model = **240.8177 / 37.0439 = 6.5634**

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?

Ans: Based on the actual CPI the closest prediction made was based on the normalized data using the weighted k-NN model, 2.8764. The primary reason behind this could be:

1) In this example it shows how important it is to normalize data. Since the data ranges in this dataset are very different, normalizing them is very important.
2) Here the dataset size is small. For a dataset of this size, using three nearest neighbors is likely to underfit slightly for this small dataset. Hence using weighted distances will mitigate this.

Q3. You have been given the job of building a recommender system for a large online 276 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

a) The company has decided to use a similarity-based model to implement the recommender system. Which of the following three similarity indexes do you think the system should be based on?

Ans: The given problem statement contains a large quantity of items, customers may not be aware of some of the products, resulting in a high percentage of features that are either missing or having “false” value in the the dataset. In these situations, it is ideal to

use a metric that disregards co-absences, as it can provide more accurate results. Thus the **Jaccard similarity index** is a suitable choice to implement the recommendation system.

b) 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?

Ans: The similarity between q and d1 can be identified by calculating co-presence (CP), co-absence (CA), presence-absence (PA), and absence presence (AP).

$$Simj(q,d1) = 2/(2+1+0) = 2/3 = 0.66$$

$$Simj\ (q,s2) = 1/(\ 1+2+1) = 1/4\ = 0.25$$

Based on the Jaccard similarity index, the query d1 is closer to the query d2 because except for ITEM 498 s1, the customer query has the same occurrence for the other items. Therefore, the model will recommend the query customer **ITEM 498**.

Q4. You are working as an assistant biologist to Charles Darwin on the Beagle voyage. You are at the Galápagos Islands, and you have just discovered a new animal that has not yet been classified. Mr. Darwin has asked you to classify the animal using a nearest neighbor approach, and he has supplied you the following dataset of already classified animals?

a) A good measure of distance between two instances with categorical features is the overlap metric (also known as the hamming distance), which simply counts the number of descriptive features that have different values. Using this measure of distance, compute the distances between the mystery animal and each of the animals in the animal dataset?

Ans:

ID	F1	F2	F3	F4	F5	F6	F7	F8	CLASS
1	T	F	T	T	F	F	T	F	Mammal
2	F	T	F	F	T	T	F	F	Amphibian
3	T	F	T	T	F	F	T	F	Mammal
4	F	T	F	T	F	T	F	T	Bird
q	F	T	F	F	F	T	F	F	

Based on the learning distance

ID1 = 6

ID2 = 1

ID3 = 6

ID4 = 2

b) If you used a 1-NN model, what class would be assigned to the mystery animal?

Ans: If we used 1NN, k=1 the nearest neighbors for the query would be d2 with the learning distance of 1 and the d2 belongs to the Amphibian class. So we can say that the query will predict the animal is **Amphibian**.

c) If you used a 4-NN model, what class would be assigned to the mystery animal? Would this be a good value for k for this dataset?

Ans: If we use the 4NN model we will consider all the ID's and the majority among the 4 learning distances is 6 and the class is Mammal. Hence, the mystery animal will be assigned a **Mammal**. It's **not a good value** of k because it includes all the instances for data & targets that are far from the target feature which impacts on the outcome of the query.

Q5. 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?

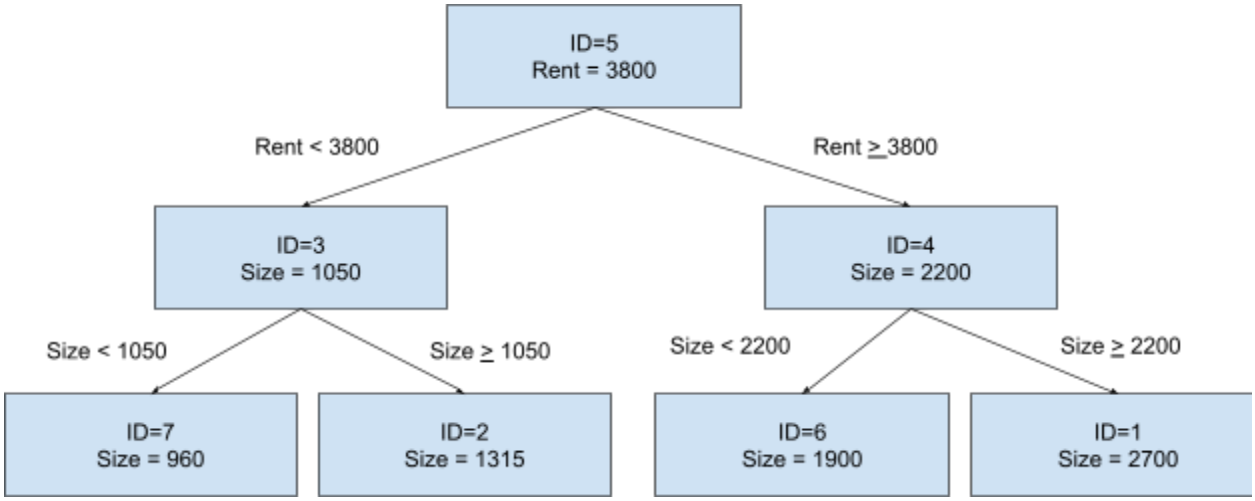
a) Create a k-d tree for this dataset. Assume the following order over the features: RENT then SIZE?

Ans: We need to prioritize the rent over size. Hence we need to sort the list in the ascending order of the rent price.

ID	Size	Rent	Price
7	960	800	720000

3	1050	1250	800000
2	1315	1800	820000
5	1800	3800	1450500
6	1900	4000	1500500
4	2200	7000	1750000
1	2700	9235	2000000

The median lies at d5, which will be our root node. Next Id 3 and 4 are the medians and will be our next root node.



b) 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?

Ans: Given,

query: SIZE = 1,000, RENT = 2,200

We will start with the root node as we were splitting based on "RENT", so the query will be "RENT". We will start with ID = 7 (d7).

Therefore, Euclidean distance (q, d7) = $\sqrt{(1000-960)^2}$ - 1+$(2200-800)^2$ = **1400.571**

Now we will move to the parent node which is instance ID = 3 (d3)

Therefore, Euclidean distance (q, d3) = $\sqrt{(1000-1050)^2}$ - 1+$(2200-1250)^2$ = **951.314**

Since the value of Euclidean distance is smaller than the value of leaf node i.e., d7 so it will move to other leaf node i.e., ID = 2 (d2)

Therefore, Euclidean distance (q, d2) = $\sqrt{(1000-1315)^2}$ - 1+$(2200-1800)^2$ = **509.141**

Since it has visited all the nodes, now it will move to the root node. Among all the instance **Euclidean distance between instance ID=2 and query is minimum**. This instance will become the nearest neighbor query which indicates that the value of property would be $820000