Machine Learning

Tutorial 3 (Similarity-Based Learning)

Exercise 1:

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 for free gambling fun*"
3. "*gambling for fun*"
4. "*machine learning for fun, fun, fun*"
5. "*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 label would a nearest neighbor model using **Euclidean distance** return for the following email: "*machine learning for free*"?
(b) What target label would a *k*-NN model with $k = 3$ and using **Euclidean distance** return for the same query?
(c) What target label 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?
(d) What target label would a *k*-NN model with $k = 3$ and using **Manhattan distance** return for the same query?
(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 label would a 3-NN model using cosine similarity return for the query?

Exercise 2:

You have been given the job of building a recommender system for a large online shop that has a stock of over 100000 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.

	ITEM	ITEM	ITEM	ITEM	ITEM
ID	107	498	7256	28063	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?

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

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

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

(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.

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

Exercise 3:

You are working as an assistant biologist to Charles Darwin on the Beagle voyage. You are at the Galapagos 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.

ID	BIRTHS LIVE YOUNG	LAYS EGGS	FEEDS OFFSPRING OWN MILK	WARM-BLOODED	COLD-BLOODED	LAND AND WATER BASED	HAS HAIR	HAS FEATHERS	CLASS
1	true	false	true	true	false	false	true	false	mammal
2	false	true	false	false	true	true	false	false	amphibian
3	true	false	true	true	false	false	true	false	mammal
4	false	true	false	true	false	true	false	true	bird

The descriptive features of the mysterious newly discovered animal are as follows:

ID	BIRTHS LIVE YOUNG	LAYS EGGS	FEEDS OFFSPRING OWN MILK	WARM-BLOODED	COLD-BLOODED	LAND AND WATER BASED	HAS HAIR	HAS FEATHERS	CLASS
Query	false	true	false	false	false	true	false	false	?

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

Exercise 4:

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	2,700	9,235	2,000,000
2	1,315	1,800	820,000
3	1,050	1,250	800,000
4	2,200	7,000	1,750,000
5	1,800	3,800	1,450,500
6	1,900	4,000	1,500,500
7	960	800	720,000

(a) Create a ***k-d* tree** for this dataset. Assume the following order over the features: RENT then SIZE.
(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 = 1000, RENT = 2200.

Exercise 5:

A data analyst building a k-nearest neighbor model for a continuous prediction problem is considering appropriate values to use for k.

(a) Initially the analyst uses a simple average of the target variables for the *k* nearest neighbors in order to make a new prediction. After experimenting with values for *k* in the range 0-10, it occurs to the analyst that they might get very good results if they set *k* to the total number of instances in the training set. Do you think that the analyst is likely to get good results using this value for *k*?
(b) If the analyst was using a distance weighted averaging function rather than a simple average for his or her predictions, would this have made the analyst's idea any more useful?

Exercise 6:

A lecturer is about to leave for the airport to go on vacation when they find a script for a student they forgot to mark. They don't have time to manually grade the script before the flight, so they decide to use a k-nearest neighbor model to grade it instead. The model is designed to award a grade to a student on the basis of how similar they are to other students in the module in terms of their grades on other modules. The following table describes a set of students in terms of their grades out of 100 on two other modules (MODULE 1 and MODULE 2) and the GRADE they got in the lecturer's module: first-class honors, second-class honors, pass, or fail.

ID	MODULE 1	MODULE 2	GRADE
1	55	85	first
2	45	30	fail
3	40	20	fail
4	35	35	fail
5	55	75	pass
6	50	95	second

(a) Looking up the results on the other modules of the student whose script hasn't been corrected, the lecturer finds that the student got the following marks: MODULE 1=60, and MODULE 2=85. Assuming that the k-nearest neighbor model uses k=1 and Euclidean distance as its similarity metric, what GRADE would the model assign the student?

(b) Reviewing the spread of marks for the other two modules, the lecturer notices that there is a larger variance across students in the marks for Module 2 than there is for Module 1. So, the lecturer decides to update the k-nearest neighbor model to use the Mahalanobis distance instead of Euclidean distance as its similarity measure. Assuming that the inverse covariance matrix for the Module 1 and Module 2 results is

$$\sum^{-1} = \begin{bmatrix} 0.046 & -0.009 \\ -0.009 & 0.003 \end{bmatrix}$$

- what GRADE would the *k*-nearest neighbor model assign the student?