KIM NGUYEN
PSID: 1255456
COSC 3337

HOMEWORK 1

PROBLEM SET 1

QUESTION 1:

Pick 3 machine learning application examples from the
Face detection and matching (e.g., iPhone X) Email spam detection Web search (e.g., DuckDuckGo, Bing, Google) Sports predictions Post once (e.g., sorting letters by zip codes) ATMs (e.g., reading checks) Credit card fraud Stock predictions Smart assistants (Apple Siri, Amazon Alexa, . . .)
Product recommendations (e.g., Net ix, Amazon) Self-driving cars (e.g., Uber, Tesla) Language translation (Google translate) Sentiment analysis Drug design Medical diagnoses and answer the following questions:

• What is the overall goal?
• How would an appropriate dataset look like?
• Which general machine learning category (supervised, unsupervised, reinforcement learning)
does this problem fit in?
• How would you evaluate the performance of your model (in very general, non technical terms)

Answer:

1. *Face detection and matching*

- **Goal:** A potential goal would be to learn from an face object how to classify it as the true object or not.
- **Dataset:** The dataset will be include a set of pictures of a single face we want to classify, with its many angle and pointer that used to validate later on.
- **Category:** Since we are working with classification – face detect as we want or not, this is supervised learning (prediction).
- **Measure Performance:** Build an algorithm based on specific set of face picture, then used it to validate number of face picture later to predict the new face detection is valid or not.

2. *Web Search (Google)*

- **Goal:** For a web search a potential goal will be based on the text type in of the question or related topic, keywords, return several most efficient related website to user.
- **Dataset:** A good data set will include every website - webpages that have same or similar or related to the text - keywords - or topics to the search.
- **Category:** A web search will include the prediction of web sites for a text applied, then it is supervised learning.
- **Measure Performance:** Based on the web search model and dataset, for a new text or question predict related website to user.

3. Stock prediction

- Goal: The ultimate goal will be predict stock price on market in the future.
- Dataset: The dataset will included the stock price and all its information on previous time until today, the more data will be the better for modeling.
- Category: Stock prediction will predict the stock price based on dataset, then it is a supervised learning.
- Measure Performance: Used the stock price dataset to perform model on it, the predict its prices in the future will be up or down, or even in a specific price range.

QUESTION 2:

If you think about the task of spam classification more thoroughly, do you think that the classification accuracy or misclassification error is a good error metric of how good an email classifier is?
What are potential pitfalls? (Hint: think about false positives [non-spam email classified as spam] and false negatives [spam email classified as non-spam]).

Answer:

For every cases of classification, both classification accuracy and misclassification error are important and is a good error metric, depends on the dataset we are working on.

- In case of classification accuracy, it is important as we measure how wells and fit the model we analyze from training data, that help leads to good measure and prediction for future data and conclusion.

- In case of misclassification error, it works similar to classification accuracy in case labeling how good the model performances. While depends on which dataset we are working with and its factor affection, we use potential pitfalls in case of false positives and false negatives to determined how should we use the model.

- As the example of spam email classified, we use potential pitfall to decide how well the model performance and how it can effect on future decisions. If the model have high misclassification in false positives, when non-spam email classified as spam, the model performance is not accurate well when we can tends to eliminate non-spam emails as it can be importance. But a high misclassification in false negative, when spam emails classified as non-spam, is less affective and can be use.

QUESTION 3:

In the exercise example of E 1), email spam classification was listed as an example of a supervised machine learning problem. List 2 examples of unsupervised learning tasks that would fall into the category of clustering. In one or more sentences, explain why you would describe these examples as clustering tasks and not supervised learning tasks. Select examples that are not already that are in the “Lecture note list” from E 1).

Answer:
Examples of unsupervised learning:

1. From a data set of 100 different Average American people as attributes, with their information in medical history as variables, included: age, weight, blood pressure, glucose, insulin, body mass index. Perform classification on the dataset to clustering them into 2 possible group of people that may have diabetes or not. From the classification identify which factors effective most to the results.
   While the problem is trying to grouping – clustering dataset into different group based on their information, then it is unsupervised learning task.

2. Given a dataset taken from student survey about their information in school, perform a classification task to clustering – group students into two categories: student who have GPA greater than 3.2 and other student who have GPA less than 3.2. The data set include 100 students with different variables in: age, race, high school grade, parent income, parent education history, childhood places …
   Base on the data set we will try to grouping – clustering data into two categories group, then this task is unsupervised learning task.

QUESTION 4 :

In the k-nearest neighbor (k-NN) algorithm, what computation happens at training and what computation happens at test time? Explain your answer in 1-2 sentences.

Answer:
In the K-Nearest Neighbor (KNN) algorithm:

- At training time, the algorithm will consist of data storing and exploring, also cleaning data noise. It includes storing dataset values, assign and organize data labels, then cleaning all missing variables.

- At testing time, the algorithm is use the dataset values to classified and assigned data labels with its corresponding, by find k-number nearest data values among training sets and assign labels with the most frequents label. The algorithm will consists of follow steps in defining and computation: calculate distance (Euclidean) between data values points, find closest neighbor points in number of k- assign and distance, the compute the highest frequent value to assign label for new value.

QUESTION 5:

Does (k-NN) work better or worse if we add more information by adding more feature variables (assuming the number of training examples is fixed)? Explain your reasoning.

Answer:

- The KNN classifier is a very intuitive method that classifies unlabeled examples based on their similarity to examples in the training set, by define the highest frequent labels among number of k nearest neighbor examples.

- We can conclude that KNN will work betters with small number of input variables and if they also on same scale, that they can choose the most accuracy labels. Therefore if we add more information by adding more features variables, KNN algorithm became worse and less efficient.

QUESTION 6:

If your dataset contains several noisy examples (or outliers), is it better to increase or decrease k? Explain your reasoning.

Answer:

- In case when out dataset contains several noisy examples or outliers, it is better to k-value. Because when KNN algorithm decide to assign the new value with the most frequent neighbor labels, while your data can be missing, it is better when you can choose higher number of k- value, which is more number of neighbor data examples, that help you to get more accurate variable labeling.

QUESTION 21:

Then, compare results with the results you got in E 22). Did you make the function faster? Yes or No? Explain why, in 1-2 sentences.

Answer:

- EXPLAIN: YES, the function run faster when rewrite the Euclidean distance function in *NumPy* using *np.sum* function.
- Since the Euclidean distance calculation from above is using a for loop to calculate, the lower Euclidean distance is calculate by a function at a single time, then the second function run a lot faster.

QUESTION 22:

Summarize your findings in 1-3 sentences.

Answer:

- The use of function *nsmallest* help to run the *KNNClasifier* implementation aspect a lot faster, while it is implemented a heap data structure, which is better and faster than use normal sorting aspect.