Assignment 1

Deadline: 27.11.2022

1. What are the characteristics of randam forest classification model? What is the difference between random forest model? Give the pseudo code of it. Explain the code.
2. What is transfer learning? Give a model and explain the model.
3. Explain support vector machine model in details. What are the advantages and disadvantages of it.
4. Explain fasttext classification model in details. What are the advantages and disadvantages of it.
5. What are the techniques used for class imbalance problem. Give specific techniques used in literatüre and explain each of them.

CSE454 Data Mining
Homework 1

Ahmet Tuğkan AYHAN
1901042692

Question 1

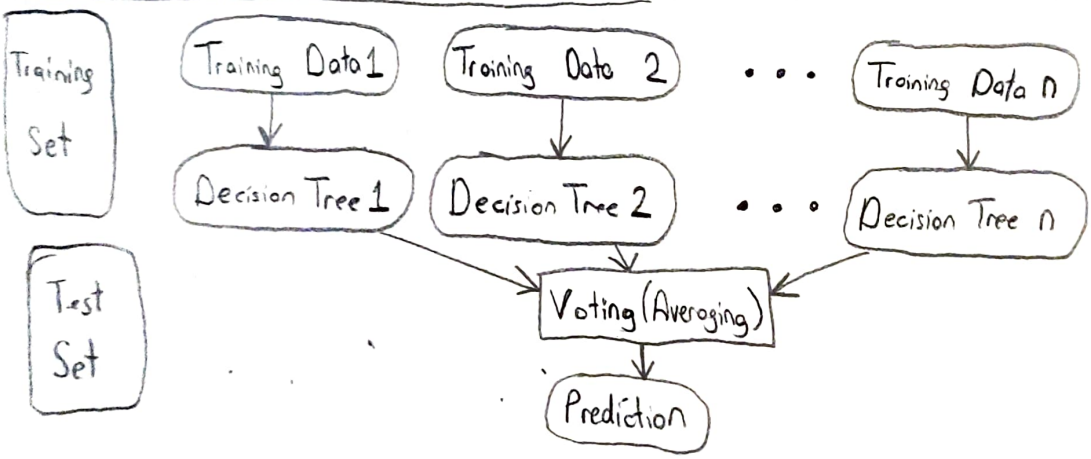
Characteristics of random forest classification model

* It has a very high accuracy rate (generally)
* It's efficient with large datasets
* In classification, it provides an estimate of important variables.
* Generated models can be reused.
* The output of the random forest classification model, is the class selected by most trees.

Differences from random forest regression model

* In classification, the output is the class selected by most trees. But in regression model, the result is the mean or average prediction of the individual trees.
* Classification model works with data having discrete labels (classes)
* Regression model works with data having a numeric or continious output. These are cannot be defined by classes.

Pseudo Code of Random Forest Classification Model (I will explain it with a demonstration)



Explanation

* Random forest is a two step algorithm.
* First we create the random forest by combining n decision tree. Then we make predictions for each tree created in first step.
* Algorithm works like this :
  - → Dataset is given to the random forest classifier
  - → The dataset is divided into subsets
  - → These subsets are given to each decision tree
  - → In training, each decision tree produced a prediction result
  - → According to the majority of the results, the Random Forest Classifier Model predicts the final decision.

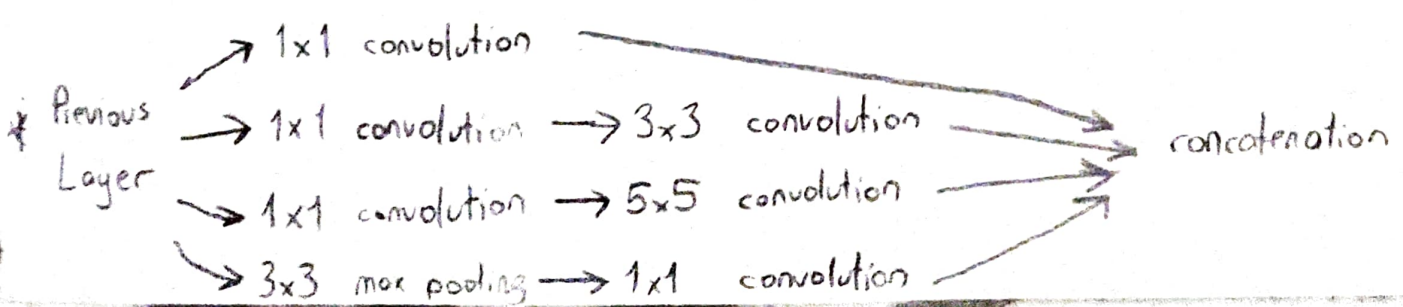
Question 2

* Transfer Learning is a machine learning method.
* In this method, a model developed for a task is used as a starting point for a model in another task.
* By doing so, model doesn't start learning from scratch which is time and energy consuming.

Traditional Machine Learning	Transfer Learning
Dataset 1 → Model 1	Dataset 1 → Model 1
Dataset 2 → Model 2	↓ Knowledge
	Dataset 2 → Model 2

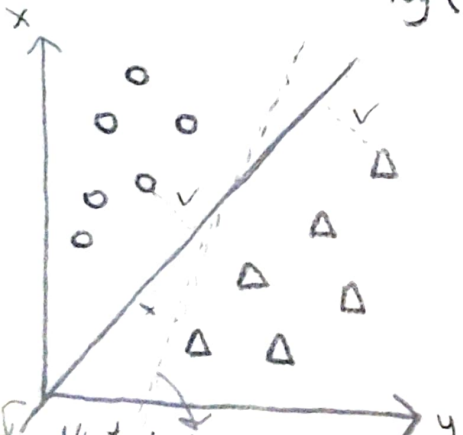
Inception Model

* It is a convolutional neural network.
* It helps classfying the different types of objects on the images
* It is also named as GoogLeNet.
* For training process, it uses ImageNet dataset
* There are 3 types of inception model:
  1. Inception v1 (Naïve version)
  2. Inception v2
  3. Inception v3
* First (naïve) version performs convolution on a layer (input).
* There are 3 different size of filters (1x1, 3x3 and 5x5)
* After performing convolution and max pooling, the output layers are concatenated and passed to the next model



Question 3

* Support Vector Machine (SVM) is a supervised machine learning model.
* It uses classification algorithms.
* It is used for two group classification problems.
* If we explain it with an example, let's say our data has two features x and y. Also we have two tags; dog and cat.
* SVM tells us if the given (x,y) pairs are belong to cat or dog.
* It does that by using a hyperplane.
* Hyperplane is a line which also called decision boundry.
* Anything that falls to one side of it is called cat and anything that falls to other side is called dog.
* The hyperplane is a line (but not necesseraly) whose distance to the nearest element of each tag (cat and dog) is the largest.



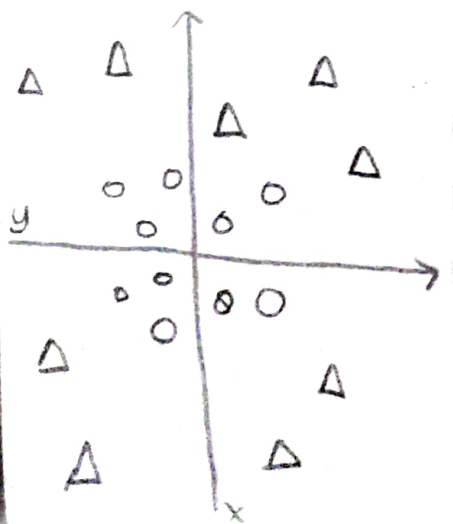
Advantages

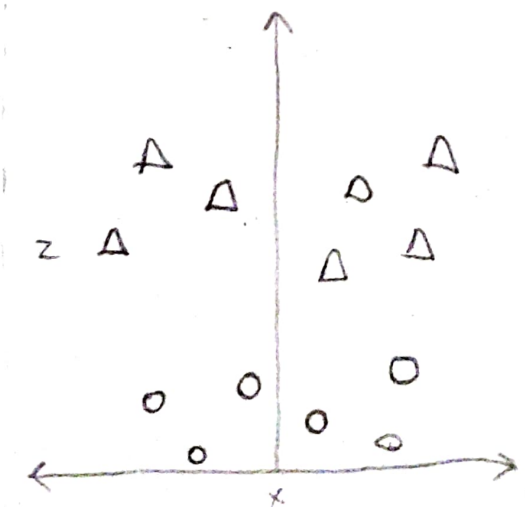
* It is effective in high dimensional spaces
* It is memory efficient
* It can be used to both classify data and predict continious numerical values.

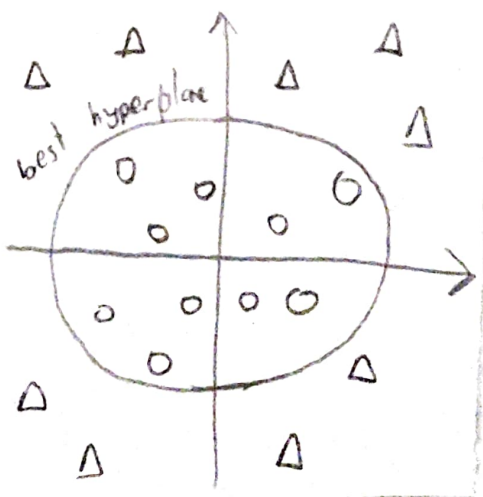
Disadvantages

* Not suitable for large datasets
* It has a risk for overfitting
* It can be costly computationally to train them (especially non-linear)

* For nonlinear dataset, we need to add third dimension to drive our hyperplane







Question 4

* Fasttext classification model is a model used for efficient learning of word representations and sentence (text) classification.
* Fasttext operates at a character level.
* It is written in C++
* Fasttext uses a hashtable for either word or character n-grams,
* Fasttext treats each word as composed of n-grams
* Because of that the vector for a word is made of the sum of this character n-grams.

for → ngram[min-n] is 3 and ngram[max n] is 6, let's create vector for "Kalem"

"<Ka", "Kal", "Kale", "Kalem>", "ale", "alem>", "lem", "lem>", "em>"

Advantages

* Easy to train your own models
* Previously trained model can be used to compute word vectors. ↗ This is especially for out of vocabulary words
* You can use it to generate vectors for paragraphs or sentences.
* It has very high accuracy rate and it is very fast
* Great choice for the language identification task.

Disadvantages

* Doesn't come with a language classification out of the box.
* It requires high memory usage. This can be lowered by determining min and max n-grams.

Question 5

1. Under-Sampling
   a. Random Undersampling
   b. Informative Undersampling
      I. EasyEnsemble
      II. BalanceCascade
2. Over-Sampling
   a. Random Over-Sampling
   b. SMOTE
   c. Ensembling Balance Bagging Classifier
3. Advanced Sampling

1.a/
* Randomly deletes examples in the majority class
* Since under-sampling reduces the data size, less time is needed for learning

1.b/
* Informative Undersampling is done by following a selection criterion to remove the examples from the majority class.

I/ EasyEnsemble extracts several subsets of independent samples from the majority class Then, it develops new classifiers using the combination of each subset with minority class

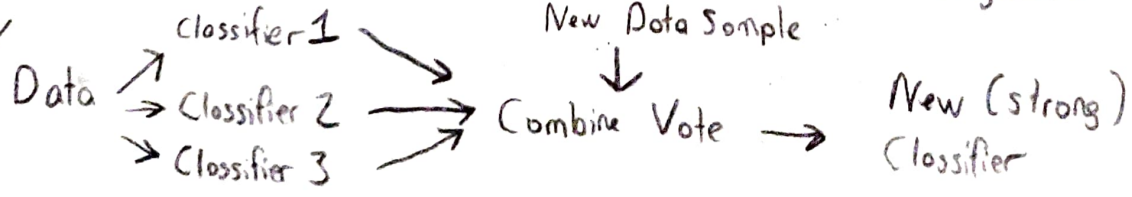
II/ BalanceCascade uses a supervised learning approach to select which majority class to ensemble

2/ Over-sampling replicates the samples from minority classes to balance the data

2.a/ Random over-sampling, randomly duplicates samples in the minority class, It may discard useful data and cofalso cause overfitting.

2.b/ SMOTE, one of the most popular technique used in class imbalance problem. It tries to balance class distribution by randomly increasing minority class samples by replicating them. Main idea in SMOTE is the generation of synthetic data between each sample of the minority class and it's nearest (k) neighbors.

2.c/



3/ Boosting can be an example for advanced sampling. Boosting places different weights on training distributions for each iteration. It efficiently alters the distributions of training data