CSEN 240 Machine Learning
Homework #2

[Level 1: high-level understanding]

1. (4 points) Which of the following can describe SVM? (select all that apply)
    a. It is a supervised learning algorithm.
    b. It is an unsupervised learning algorithm.
    c. It can be used to solve a classification problem.
    d. The algorithm is based on the concept of finding a hyperplane that separates the data with a maximal margin, in order to classify future data points with more confidence.
2. (4 points) Which of the following problems SVM is best suited for? (select one)
    a. Predict housing prices based on location, size, and other factors.
    b. Classify images, such as identifying the contents of a photograph
    c. Group customers according to their purchasing behavior
    d. Extract important features from a dataset, allowing for more accurate analysis
3. (3 points) What is the purpose of the training set, validation set, and testing set in machine learning? (One each)
    a. To evaluate the performance of the model on unseen data
    b. To fine-tune the model's parameters and learn the underlying patterns in the data
    c. To evaluate the performance of the model during the training process
4. (4 points) What are the ML problems due to data?
    a. If we are trying to predict the price of a house, but the training data includes irrelevant information such as the color of the house or the name of the previous owner
    b. If we are trying to predict the success of a new product based on customer data, but we only have a small amount of historical customer data.
    c. If we are trying to predict customer behavior but the data contains a lot of missing values or errors
    d. If we are trying to predict customer behavior, but the training data only includes data from a single demographic group.

    The choices are (one each)
        i. insufficient quantity of training data
        ii. nonrepresentative training data
        iii. poor-quality data
        iv. irrelevant features
5. (10 points) Please watch the guest lecture by Mr. FanFan Jiang on his AI-based system for generating English-to-ASL (American Sign Language) datasets. Here is a quick review:
    - The initial objective of this project is to create an automated sign language production system capable of generating high-quality sign language captions. However, sign language production have been hindered by the limited availability of publicly accessible sign language datasets. To address this challenge, FanFan

devised a system tailored to generate continuous American Sign Language (ASL) datasets.

- FanFan made use of Google's key-point tracking tool to analyze publicly accessible news clips that contained ASL captions. Following this, he devised post-processing and data-filtering methods to surmount constraints associated with current human detection techniques, thereby guaranteeing the development of high-quality datasets. Subsequently, FanFan harnessed a Transformer neural network to translate English sentences into key-point movements. The sequences of key-point movements were subsequently used to produce lifelike 3D-rendered ASL videos.
- Please describe whether FanFan has addressed the following issues and if so, what has he done:
    (A) insufficient quantity of training data
    (B) nonrepresentative training data
    (C) poor-quality data
    (D) irrelevant features

[Level 2: manual exercise]

6. (20 points) Given the following dataset of labeled points in two dimensions, use a support vector machine to find the best separating hyperplane. (Note: please use hard margin. Using high-school geometry should be sufficient; no need to solve the NP optimization problem.)
    - Positive samples: {(3,3), (4,4)}
    - Negative samples: {(2,1), (1,2)}
    - Please report the equation for the hyperplane.
    - Use the constructed hyperplane to predict the class of a new data point at (2,2.4)

[Level 3: extension of the basic algorithm]

7. (15 points) There may be outliers or noises in the data from real-world applications. To address this issue, a soft margin can be used in a modified optimization problem, known as a soft-margin SVM:
    - Objective: $\min \frac{1}{2}\|\vec{w}\|^2 + C\sum_i \xi_i$
    - Constraint: $y_i(\vec{w}\cdot\overrightarrow{x_i} + b) = 1 - \xi_i$ and $\xi_i \geq 0$
    - ξ_i is the slack, which allows $\overrightarrow{x_i}$ to be inside the margin
    - SVM without the slacks is known as hard-margin SVM.

    a. Where is $\overrightarrow{x_i}$ relative to where the margin is when its ξ_i value is 0?
    b. Where is $\overrightarrow{x_i}$ relative to where the margin is when $0 \leq \xi_i \leq 1$?
    c. Where is $\overrightarrow{x_i}$ relative to where the margin is when $\xi_i > 1$?

[Level 4: computer-based exercise]

8. (20 points) (Using HW2-7.ipynb as the template)
    a. Using popular scikit-learn package for SVM

b. Assuming (0, 1, 1) is the ground truth of the decision boundary, create 40 unique samples (20 are positive and 20 are negative).
c. First, evenly split the 40 samples into two sets: one is called training samples, and the other is called testing samples.
d. Second, train a hard-margin SVM using 100% of the training samples, and test the accuracy of the unseen testing samples. (Repeat 100 times for the mean and standard deviation of accuracies)
e. Third, train a hard-margin SVM using 50% of the training samples (e.g., 5 positive and 5 negative samples), and test the accuracy of the unseen testing samples. (Repeat 100 times for the mean and standard deviation of accuracies)
f. What do you observe? Can you explain?
g. If necessary, please use the following website to test the statistically significant difference https://www.graphpad.com/quickcalcs/ttest1/?format=SD

9. (20 points) (Using HW2-8.ipynb as the template.)
    a. Use LIBSVM (https://www.csie.ntu.edu.tw/~cjlin/libsvm/)
        i. Need to install libsvm-official (in colab, to install new python package, use "!" in front of command line)
        ii. Follow "A Practical Guide to Support Vector Classification" https://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf
        iii. Use A.1 Astroparticle Physics dataset: svmguide1 and svmguide1.t
    b. Use the radial basis function (RBF) (i.e., t =2)
    c. First, train the model without scaling the dataset with γ = 2, C = 32, then report your prediction accuracy on the testing data
    d. Second, scale datasets with default parameters, train the model, and then report your prediction accuracy on the testing data. What do you observe? Why?
    e. Third, change C = 2, 8, 32, 128, 512, repeat the model training (using scaled datasets), and report the prediction accuracy of the training data and that of the testing data. What do you observe across different C's? Why?