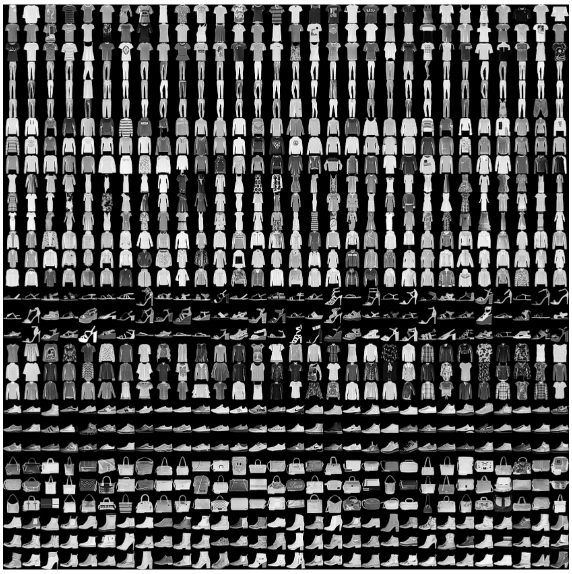
Machine Learning (Homework 3)

Due date : 2022/12/23 23:59:59

1 Support Vector Machine (SVM) (40%)

Support vector machine (SVM) is known as a popular method for pattern classification. In this exercise, you will implement SVM for classification. Here, the Fashion MNIST dataset is given in x_train.csv and t_train.csv. The input data contain three classes of apparels: T-shirt/top, Trouser and Sandal. Each example is a 28x28 gray-scaled image, associated with a class label.



Data Description

- **x_train** is a 300 × 784 matrix, where each row is all pixels of a training image.
- **t_train** is a 300 × 1 matrix, which records the classes of the training images. 0, 1, 2 represent the apparels: T-shirt/top, Trouser and Sandal, respectively.

In the training procedure of SVM, we need to optimize with respect to the Lagrange multipliers $a = \{a_n\}$. Here, we use the sequential minimal optimization to solve the problem. For details, you can refer to the paper [Platt, John. "Sequential minimal optimization: A fast algorithm for training support vector machines", 1998]. The classifier is written by

$$y(\mathbf{x}) = \sum_{n=1}^{N} a_n t_n k(\mathbf{x}, \mathbf{x}_n) = \mathbf{w}^{\top}\mathbf{x} + b$$

$$\mathbf{w} = \sum_{n=1}^{N} \alpha_n t_n \phi(\mathbf{x}_n)$$

$$b = \frac{1}{N_{\mathcal{M}}} \sum_{n \in \mathcal{M}} \left(t_n - \sum_{m \in \mathcal{S}} a_m t_m k\left(\mathbf{x}_n, \mathbf{x}_m\right) \right)$$

where $\mathcal{M}$ denotes the set of indices of data points having $0 < a_n < C$.

Scikit-learn is a free software machine learning library that provides sklearn.svm. You are allowed to use the library to calculate the multipliers (coefficients) rather than using the **prediction function** directly. In this exercise, you will implement SVM with linear kernel and polynomial kernel (bonus).

- **Linear kernel:**

$$k(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^\top \phi(\mathbf{x}_j) = \mathbf{x}_i^\top \mathbf{x}_j$$

- **Polynomial (homogeneous) kernel of degree 2:**

$$k(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^\top \phi(\mathbf{x}_j) = (\mathbf{x}_i^\top \mathbf{x}_j)^2$$

$$\phi(\mathbf{x}) = [x_1^2, \sqrt{2}x_1 x_2, x_2^2]$$

$$\mathbf{x} = [x_1, x_2]$$

SVM is a binary classifier, but the problem here has three classes. To solve this problem, there are two decision approaches, one is the 'one-versus-the-rest', and the other is the 'one-versus-one'.

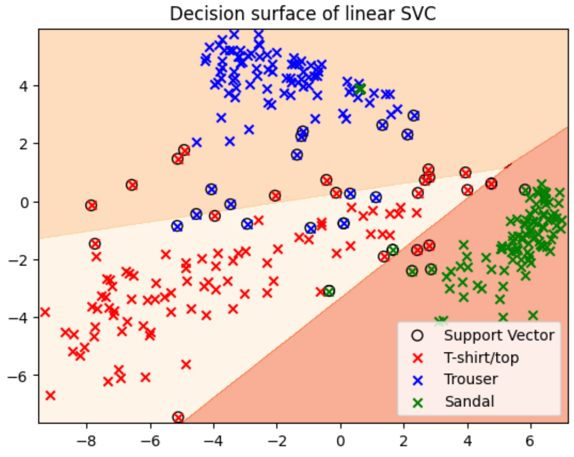
1. It is popular to use principal component analysis (PCA) to reduce the dimension of images to $d = 2$. Please implement it by yourself instead of using the method from sklearn. (10%)
2. Describe the difference between two decision approaches (one-versus-the-rest and one-versus-one). Decide which one you want to choose and explain why you choose this approach. (5%)
3. Use the principle values projected to top two eigenvectors obtained from PCA, and build a SVM with linear kernel to do multi-class classification. You can decide the upper bound C of a_n by yourself or just use the default value provided by sklearn. Then, plot the corresponding decision boundary and show the support vectors. The sample figures are provided below. (25%)

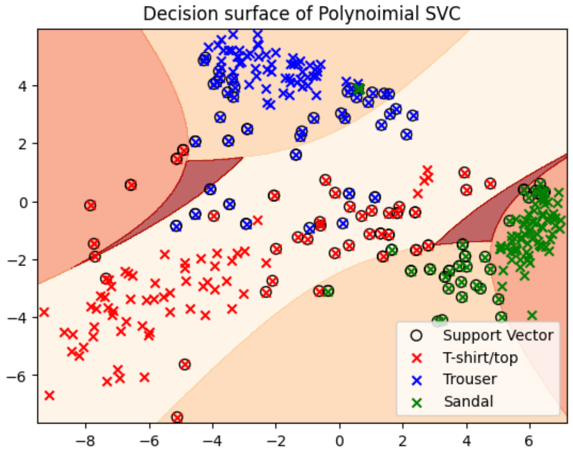
Bonus (10%)

- Repeat 3 with polynomial kernel (degree = 2).

Hint

- You need to implement the whole algorithm except for multipliers (coefficients).





2 Gaussian Mixture Model (60%)

In this exercise, you will implement a Gaussian mixture model (GMM) and apply it in image segmentation. First, use a K-means algorithm to find K central pixels. Second, use the expectation maximization (EM) algorithm (please refer to textbook p.438-p.439) to optimize the parameters of the model. The input image is given by hw3.jpg. According to the maximum likelihood, you can decide the color μ_k, $k \in [1, \ldots, K]$ of each pixel x_n of output image

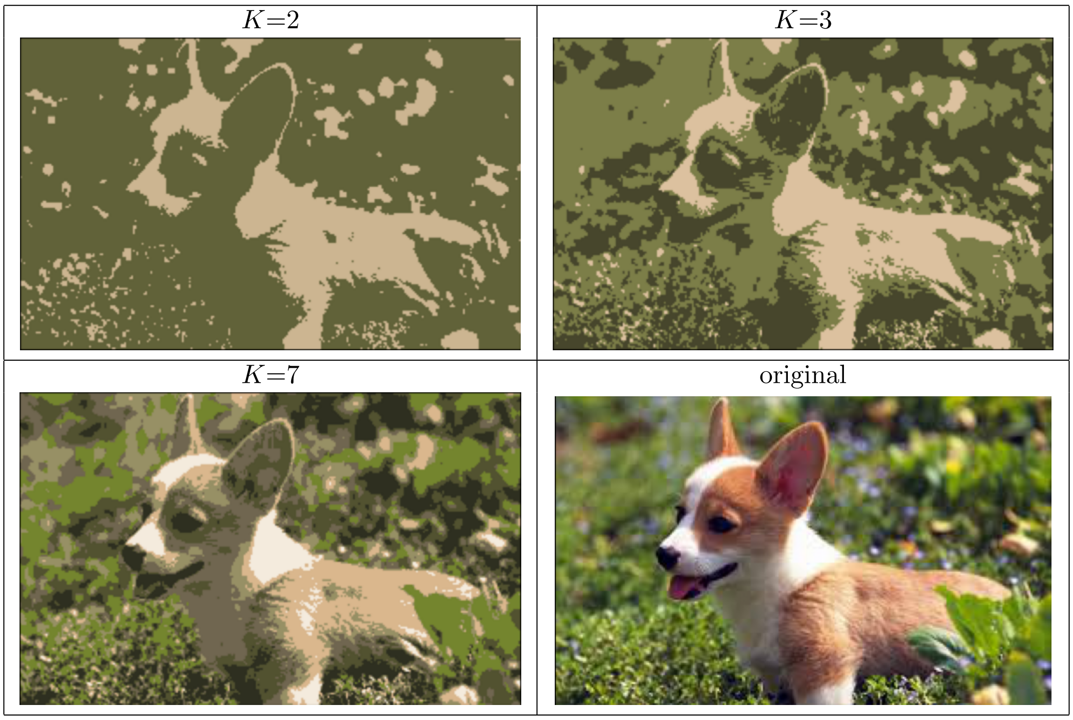
1. Please build a K-means model by minimizing

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma_{nk} \|x_n - \mu_k\|^2$$

   and show the table of the estimated $\{\mu_k\}_{k=1}^{K}$.

2. Use $\mu = \{\mu_k\}_{k=1}^{K}$ calculated by the K-means model as the means, and calculate the corresponding variances σ_k^2 and mixing coefficient π_k for the initialization of the GMM $p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \sigma_k^2)$. Optimize the model by maximizing the log likelihood function $\log p(x|\pi, \mu, \sigma^2)$ over N pixels through EM algorithm. Plot the learning curve for log likelihood of GMM. (Please terminate EM algorithm when the number of iterations arrives at 100.)

3. Repeat steps 1 and 2 for K = 2, 3, 7 and 20. Please show the resulting images of K-means model and GMM, respectively. Below are some examples.


K=2

K=3

K=7

original

4. Make some discussion about what is crucial factor to affect the output image between K-means and Gaussian mixture model (GMM), and explain the reason.

5. The input image shown below comes from the licence-free dataset for personal and commercial use. Image from: https://pickupimage.com/free-photos/Cat-in-the-forest/2333003

3 Rules

- Please name the assignment as hw3_StudentID.zip (e.g. hw3_0123456.zip).
- In your submission, two files are required.
  Note : Only the following two files are accepted, so the code of each exercise should be written in one .py file.
  - **hw3_StudentID.ipynb** file which contains all the results and codes for this homework. Also, it should contain the description or explanation for this homework. (Please write all codes in one file.)
  - **hw3_StudentID.py** file which is downloaded from the .ipynb file.
- Implementation will be graded by
  - Completeness
  - Algorithm Correctness
  - Discussion
- Only Python implementation is acceptable.
- Only the packages we provided is acceptable.
- DO NOT PLAGIARIZE. (We will check program similarity score.)