SVM Project

In this project you are asked to run experiments on the Wine dataset from UCI database. There are 178 examples, each labeled as 0, 1 or 2. For convenient, wine.csv file(with column name) is provided. You are asked to train SVM classifiers for this problem using **scikit-learn**. The challenge is to select the free parameters to maximize the accuracy. You are asked to produce a total of 4 programs and their corresponding best classifier models:

1. A classifier trained with 50% of the data using a polynomial kernel. Program should be named as **SVM-p50.py** and model should be named as **SVM-p50_model.sav**.
2. A classifier trained with 50% of the data using an exponential (rbf) kernel. Program should be named as **SVM-e50.py** and model should be named as **SVM-e50_model.sav**.
3. A classifier trained with 75% of the data using a polynomial kernel. Program should be named as **SVM-p75.py** and model should be named as **SVM-p75_model.sav**.
4. A classifier trained with 75% of the data using an exponential (rbf) kernel. Program should be named as **SVM-e75.py** and model should be named as **SVM-e75_model.sav**.

- Your programs must set the random seed of python to 1 to make sure that your results are reproducible. In other words, we should be able to reproduce your model using the program.
- The "p" programs must use a polynomial kernel, and the "e" programs must use an exponential kernel.
- Your model will be tested on the entire dataset.
- The training and testing of each program should not take more than 3 minutes.

Installing scikit-learn

```
pip install -U scikit-learn
```

Provided files and programs

1. **wine.csv** dataset file.
2. An example program **SVM-50.py**.
3. **fraction_wine.py** program to produce random training dataset.
4. An example jupyter notebook for Wisconsin breast cancer dataset from UCI (which only has 2 categories) can be found in lecture notes.

What you need to do

Determine the parameters for the SVM to maximize the accuracy (F1 score).

Grading

1. We will load and run the model you produced on entire dataset.
2. We will also generate random subsets of training examples by running the program **fraction_wine.py** with a seed that is kept secret. If, for example, the seed is 7, generating a fraction of 50% can be done as follows:

```
python3 fraction_wine.py  --dataset wine.csv --frac 0.5 --seed 7
This creates the file wine_7_50.csv.
```

Your grade will be based on the accuracy of your models trained with the generated examples and tested on the entire training data.

What you need to submit

Your submission should be a single zip archive named **netid.zip**, where **netid** is your net id. The zip archive should contain the following:

1. Source code of the python scripts. They should be named as follows:

```
SVM-p50.py, SVM-e50.py, SVM-p75.py, SVM-e75.py,
```

2. The saved best models. They should be named as follows:

```
SVM-p50_model.sav, SVM-e50_model.sav, SVM-p75_model.sav, SVM-e75_model.sav.
```

3. Documentation describing the results of experiments/accuracy that your programs achieve on the provided data.

SCIKIT-LEARN

Scikit-learn is a popular free software machine learning library for the Python programming language. Their description of SVM can be found in the following link:
https:// scikit −learn.org/stable/modules/svm.html
The method that corresponds to what was covered in class is **SVC** (Support Vector Classification). The description of its parameters can be found in:
https:// scikit −learn.org/stable/modules/generated/sklearn.svm.SVC.html

Running SVC with a polynomial kernel (and soft margins) requires the following parameters to be set:

```
C = positive float value. This is the soft margins parameter.
Kernel = 'poly'
degree = nonnegative integer value
gamma = positive float value. You cannot use 'scale' or 'auto'. (1.0 in class.)
coef0 = float value. (1.0 in class.)
```

Running SVC with a exponential kernel (and soft margins) requires the following parameters to be set:

```
C = positive float value
Kernel = 'rbf'
gamma = positive float value. You cannot use 'scale' or 'auto'.
```

The kernels:

name	formula	formula given in class	
linear:	$K(x,y) = x'y$	$x'y$	same
polynomial:	$K(x,y) = (\gamma\ x'y + r)^d$	$(x'y+1)^d$	same with $r=1, \gamma=1$
rbf:	$K(x,y) = e^{(-\gamma\|x-y\|^2)^d}$	$e^{-\|x-y\|^2/(2\sigma^2)}$	same with $d=1, \gamma=\frac{1}{2\sigma^2}$