Machine Learning (Homework #3)

Due date: 1/6

I. Gaussian Process

a. **Gaussian Process for regression**: In this exercise, you are given a file gp_data.csv which are sampled from a Gaussian process. Please implement your Gaussian process for regression, with precision of the noise $\beta = 0.01$ and the kernel function of the form

$$k(\boldsymbol{x}, \boldsymbol{x}') = \theta_0 \exp\{-\frac{1}{2}\sum_{i=1}^{2}\eta_i(x_i - x_i')^2\} + \theta_1$$

Please plot the figure that compares the prediction values and target values using Test set (for example, Fig. 1), and show the corresponding root-mean-square error in the report

$$E_{\mathrm{RMS}} = \sqrt{2(y(\boldsymbol{x}_n) - t_n)^2/N}.$$

Data description: There are **input data X** and **target date T** in the file. The number of data points is 1600. Please use the first 1200 samples as train set and the last 400 samples as test set.
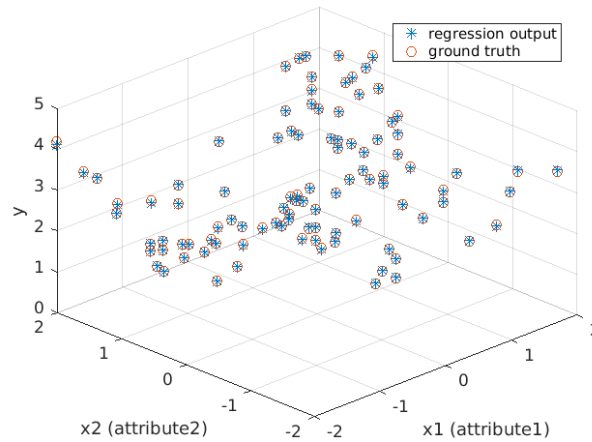

Fig.1

b. **Gaussian Process (GP) for binary classifier**: In this exercise, you will implement Gaussian Process for binary classifier with the kernel function of the form

$$k(\boldsymbol{x}, \boldsymbol{x}') = \theta_0 \exp\{-\frac{1}{2}\sum_{i=1}^{2}\eta_i(x_i - x_i')^2\} + \theta_1$$

You are given a file which is extract from HIV-1 protease cleavage data set. Please show the classification accuracy in the report.

1. **Data description**: There are 6590 octamers (8 amino acids) and a flag (-1 or 1) in the original data set. There are twenty kind of amino acid. Each kind of amino acid is expressed by a number between 0 to 1 and target label is denoted as 0 or 1 in the given data set:

G(Glycine, Gly)	0
P(Proline, Pro)	0.0526
A(Alanine, Ala)	0.1053
V(Valine, Val)	0.1579
L(Leucine, Leu)	0.2105
I(Isoleucine, Ile)	0.2632
M(Methionine, Met)	0.3158
C(Cysteine, Cys)	0.3684
F(Phenylalanine, Phe)	0.4211
Y(Tyrosine, Tyr)	0.4737
W(Tryptophan, Trp)	0.5263
H(Histidine, His)	0.5789
K(Lysine, Lys)	0.6316
R(Arginine, Arg)	0.6842
Q(Glutamine, Gln)	0.7368
N(Asparagine, Asn)	0.7895
E(Glutamic Acid, Glu)	0.8421
D(Aspartic Acid, Asp)	0.8947
S(Serine, Ser)	0.9474
T(Threonine, Thr)	1

2. The GP algorithm for classification is given below

Algorithm 1 Algorithm

1: Training
2: initialize $\mathbf{C}_N$, $\mathbf{t}_N$, $\mathbf{a}_N$;
3: $\mathbf{a}_N = 0$
4: **for** each epoch **do**
5:     calculate $\mathbf{W}_N$ which is a diagonal matrix with elements $\sigma(a_n)(1 - \sigma(a_n))$
6:     $\mathbf{a}_N^{new} = \mathbf{C}_N(\mathbf{I} + \mathbf{W}_N\mathbf{C}_N)^{-1}\{\mathbf{t}_N - \boldsymbol{\sigma}_N + \mathbf{W}_N\mathbf{a}_N\}$ equation (6.83)
7: **end for**
8: Testing
9: Given $\mathbf{a}_N$,$\mathbf{C}_N$, $\mathbf{t}_N$, $\boldsymbol{\sigma}_N$;
10: **for** each test sample **do**
11:     Calculate $\mathbb{E}[a_{N+1}|\mathbf{t}_N] = \mathbf{m}_{N+1} = \mathbf{k}^\top(\mathbf{t}_N - \boldsymbol{\sigma}_N)$
12:     Calculate $\text{var}[a_{N+1}|\boldsymbol{\sigma}_N] = \mathbf{C}_{N+1} = c - \mathbf{k}^\top(\mathbf{W}_N^{-1} + \mathbf{C}_N)^{-1}\mathbf{k}$
13:     Prediction value $\widehat{y}(i) = \text{sigmoid}(\mathbb{E}[a_{N+1}|\mathbf{t}_N])$
14: **end for**

- Note: you can decide the values of kernel parameters $(\boldsymbol{\theta}, \eta)$ by your own. Show your decision procedure.
- You have to implement all details in the Gaussian process model and do not use toolbox.

II. Support Vector Machine (SVM)

Support vector machines (SVM) is known as a popular method for pattern classification. In this exercise, you will implement the SVM for classification. You are given the Iris dataset (iris.mat) which contains the measurements of 150 specimens from each of three different species of iris—setosa (1), versicolor (2), and virginica (3) —with the following four features: sepal length, sepal width, petal length and petal width. For ease of visualization, here we only use the first two features: sepal length and sepal width in the iris.mat.

In the training procedure of SVM, we need to optimize with respect to the Lagrange multiplier $\boldsymbol{\alpha} = \{\alpha_n\}$. Here, we use the Sequential Minimal

Optimization (smo.m) to solve the problem. The idea of SMO algorithm is to update only two α_n and fix the others at each iteration. Then, we can find α_n by quadratic optimization and update another pair of α_n until all $\boldsymbol{\alpha}$ converge. Finally, we use the result to calculate bias b. For details, one can refer to the paper [Platt, John. "Sequential minimal optimization: A fast algorithm for training support vector machines." (1998)].

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

- One-versus-the-rest approach constructs a classifier distinguishing between one class C_k and the remaining. Predict the label by finding the corresponding classifier which reports the highest confidence score.
- One-versus-one approach constructs a classifier distinguishing between two classes. Predict the label by Voting. All classifiers vote for one class, finally one class which gets the highest number gets predicted

In this exercise, you will implement two kinds of kernel SVM

SVM:
$$y(\mathbf{x}) = \sum_{n=1}^{N} \alpha_n t_n k(\mathbf{x}, \mathbf{x_n}) = \mathbf{w}^\top + b$$

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

Linear kernel:

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

Polynomial (homogeneous) kernel of degree 2:

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

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

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

a. Choose one decision approach. Build two kinds of kernel SVM to do multiclass classification. Please show the accuracy on training data and test data. Mention the decision approach you use in this task.

b. Please plot the corresponding decision boundary and support vector. The following Figures show the examples.
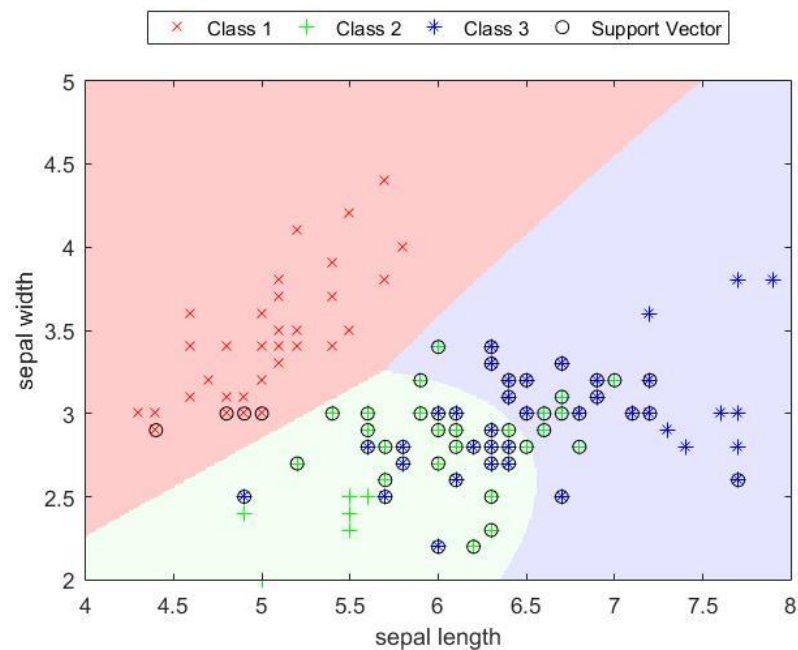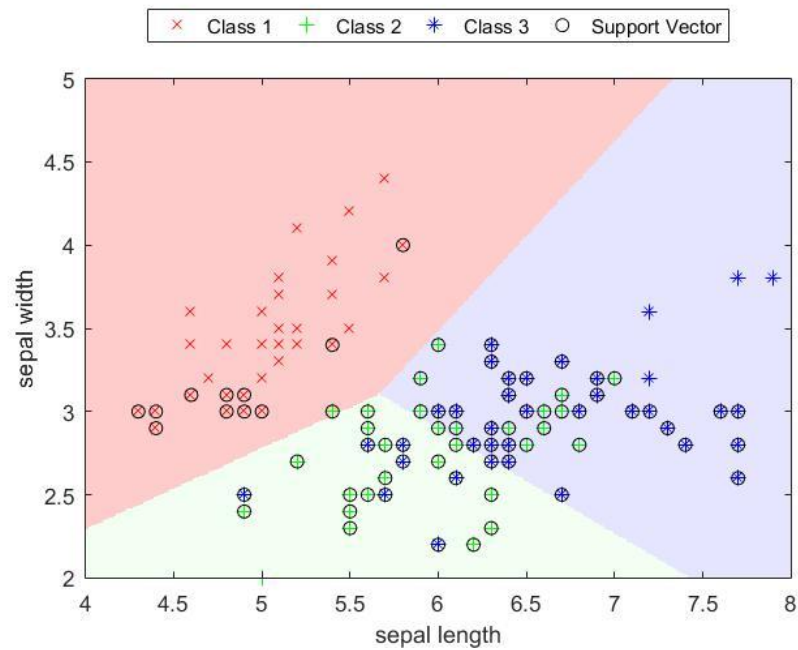
c. Do some discussion about different kernels in SVM.

Note:

If you use other languages, you can use toolbox only for computation of α_n, but the kernel function should be built by yourself.

Appendix

$[\text{alpha}, \text{bias}] = \text{sm}(\mathbf{K}, \mathbf{y}, c, \text{tol})$

$\mathbf{K}$: $n \times n$ `kernel matrix,` $k_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$;

$\mathbf{y}$: $1 \times n$ `vector of labels, -1 or 1`

c: `a regularization parameter such that` $0 \le \alpha_i \le \frac{c}{n}$

tol: `tolerance for terminating criterion`