Machine Learning Homework 5

Gaussian Process & SVM

Due Date: 2025/05/07 23:59

I. Gaussian Process

In this section, you are going to implement the Gaussian Process and visualize the result.

- Training data
  - **input.data** is a 34x2 matrix. Every row corresponds to a 2D data point (X_i,Y_i).
  - $Y_i = f(X_i) + \epsilon_i$ is a noisy observation, where $\epsilon_i \sim N(\bullet|0, \beta^{-1})$. You can use $\beta = 5$ in this implementation.
- What you are going to do
  - **Task1:**

    Apply Gaussian Process Regression to predict the distribution of *f* and visualize the result. Please use a rational quadratic kernel to compute similarities between different points.

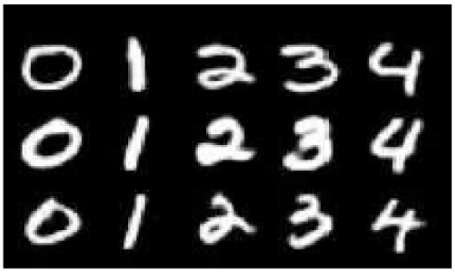
    Details of the visualization:
    - Show all training data points.
    - Draw a line to represent the mean of *f* in range [-60,60].
    - Mark the 95% confidence interval of *f*.

    (You can use matplotlib.pyplot to visualize the result, e.g. use matplotlib.pyplot.fill_between to mark the 95% confidence interval, or you can use any other package you like.)
  - **Task2:**

    Optimize the kernel parameters by minimizing negative marginal log-likelihood, and visualize the result again. (You can use scipy.optimize.minimize to optimize the parameters.)

II. SVM on MNIST dataset

Use SVM models to tackle classification on images of hand-written digits (digit class only ranges from 0 to 4, as the figure shown below).



- Training data
  - **X_train.csv** is a 5000x784 matrix. Every row corresponds to a 28x28 gray-scale image.
  - **Y_train.csv** is a 5000x1 matrix, which records the class of the training samples.
- Testing data
  - **X_test.csv** is a 2500x784 matrix. Every row corresponds to a 28x28 gray-scale image.
  - **Y_test.csv** is a 2500x1 matrix, which records the class of the test samples.
- What you are going to do
  - **Task1**
    Use different kernel functions (linear, polynomial, and RBF kernels) and compare their performance.
  - **Task2**
    Please use C-SVC (you can choose by setting parameters in the function input, C-SVC is soft-margin SVM). Since there are some parameters you need to tune for, please do the grid search for finding parameters of best performing model. For instance, in C-SVC you have a parameter C, and if you use RBF kernel you have another parameter γ, you can search for a set of (C, γ) which gives you best performance in cross-validation. (Thereare lots of sources on the internet, just google for it.)
  - **Task3**
    Use linear kernel + RBF kernel together (therefore a new kernel function) and compare its performance with respect to others. You would need to find out how to use a user-defined kernel in libsvm.

III. Report

- Submit a report in PDF format. The report should be written in English.
- Please strictly follow the report format. We will deduct some points according to the situation if you don't follow it.
- Please don't explain the code line by line. You need to explain it clearly and well structurally. For example, explain what the function is used for and explain what formula you have used in the function.
- Since this homework is mainly graded by report, please spend more time on it. (e.g. well organized) We won't give you any points if you just finish the code.

A. Report format:

I. Gaussian Process

1. **Code (20%)**
    - Expected Content
        - Paste your code snippets in screenshot or in formatted code block with comments and explain your code.
        - For example, show the formula of rational quadratic kernel and the process you optimize the kernel parameters.
        - !! Note if you don't explain your code, you can't get any points in section 2 and 3 either!
    - Code for Task 1 (10%)
    - Code for Task 2 (10%)
2. **Experiments (20%)**
    - Expected Content
        - Show experiment settings and results, including the figures and the hyperparameters we asked you to show.
        - Note that if you don't explain your code in the above section, you cannot get any points in this section either.
    - Experiment for Task 1 (10%)
    - Experiment for Task 2 (10%)
3. **Observation and Discussion (10%)**
    - Anything you want to discuss, such as comparing the performance when using different hyperparameters.
    - If you need to refer to images or code snippets in previous sections, you can either
        i. Add a duplicate one in this section.
        ii. Add title or numbering system to all images and code blocks and refer to them by their corresponding identifier.

iii. Put some parts of your discussions in the middle of the previous sections, but you must make it super obvious for us. (e.g. Add title or headings to tell us that the paragraph is part of “Observations and Discussion”)

II. SVM on MNIST

1. **Code (20%)**
   - Expected Content
     - Paste your code snippets in screenshot or in formatted code block with comments and explain your code.
     - !! Note if you don’t explain your code, you can’t get any points in section 2 and 3 either!
   - Code for Task 1 (5%)
   - Code for Task 2 (10%)
   - Code for Task 3 (5%)
2. **Experiments (20%)**
   - Expected Content
     - Show experiment settings and results, including the figures and the hyperparameters we asked you to show.
   - Experiment for Task 1 (6%)
   - Experiment for Task 2 (8%)
   - Experiment for Task 3 (6%)
3. **Observations and Discussion (10%)**
   - Anything you want to discuss, such as comparing the performance when using different hyperparameters.
   - If you need to refer to images or code snippets in previous sections, you can either
     i. Add a duplicate one in this section.
     ii. Add title or numbering system to all images and code blocks and refer to them by their corresponding identifier.
     iii. Put some parts of your discussions in the middle of the previous sections, but you must make it super obvious for us. (e.g. Add title or headings to tell us that the paragraph is part of “Observations and Discussion”)

B. Turn in

1. Report (.pdf)
2. Source code

You should compress your source code and the report into a zip file and name it like ML_HW5_yourstudentID_name.zip, e.g. ML_HW5_0856XXX_王小明.zip.

- If the zip file name has format error (.rar is not allowed) or the report is not in pdf format, there will be a penalty (-10).
- Submit your homework in time.
    - After the deadline, you can still submit in the following 7 days, you will get only 70% of the original score.
    - Starting from the seventh day after the deadline, you cannot submit your homework and you will **get 0 score**.
    - Whenever you submit your homework, the latest submission will be used for grading. (so don't accidentally submit something after the deadline, you will get 30% discount no matter what)

♦ Packages allowed in this assignment:

You are only allowed to use the LIBSVM library, numpy, scipy.optimize, scipy.spatial.distance, and package for visualizing results. Official introductions can be found online.
Important: scikit-learn is not allowed.