Assignment VI (2015.2 - T01)

Submission Deadline: December, 18th, 2015.

Instructions:

- Submission deadline: December, 18th, 2015.
- This assignment must be delivered as a report (with Introduction, Methodology, Results, Conclusion and References).
- Source codes must be delivered as attachments.
- When answering the questions, books and papers citations are allowed as long as they are listed in "references" section of the report.
- When submitting this assignment, please inform if you are registered or not in SIGAA.
- If you are enrolled in SIGAA, provide your university ID number when submitting the assignment.

Exercises

Exercise 01

Using Dermatology Data Set [1] available at:
`http://archive.ics.uci.edu/ml/datasets/Dermatology`,
compare *Support Vector Machines classifier* (SVM) with *Fisher's Linear Discriminant* and *The Perceptron Algorithm* in terms of classification accuracy. You can implement SVM or use a third-party toolbox or API.

The Dermatology data set is divided in six classes. Each class has a number which identifies a skin disease. Class 1, which is the class for *Psoriasis*, has 112 samples; Class 2, which is the class code for *Seboreic Dermatitis*, has 61 samples; Class 3, which is the class code for *Lichen Planus*, has 72 samples; Class 4, which is the class code for *Pityriasis Rosea*, has 49 samples; Class 5, which is the class code for *Cronic Dermatitis*, has 52 samples and Class 6, which is the class code for *Pityriasis Rubra Pilaris*, has 20 samples. There are 8 samples in the data set **that is missing an attribute value (Age)**. They are distinguished with "?". **Remove those samples from the dataset**.

The data set, called `dermatology.data`, is a text file which the rows (358) are the number of samples. So each row represents a sample. In each sample, the first 34 columns are the attributes and in the last column (35th column) is the sample class-id.

Knowing that the Dermatology data set must be tested using the *Multiple Classes Approach*, consider a K-class discriminant comprising K linear functions of the form

$$y_k(\mathbf{x}) \quad = \quad \mathbf{w}_k^T \phi(\mathbf{x})$$

and then assign a sample $\mathbf{x}$ to a class $\mathcal{C}_k$ if $y_k(\mathbf{x}) > y_j(\mathbf{x})$ for all $j \neq k$.

To use the Dermatology data set for evaluating SVM and the other two linear models for clasification follow the following steps:

1. Load `dermatology.data` file;

2. Put all the rows along with the first 34 columns in a matrix $\mathbf{X}$;

3. Put the labels in a vector $\mathbf{l}$;

4. Create a matrix $\mathbf{T}$ where each row corresponds to the rows of $\mathbf{l}$ and each of these rows has 6 columns filled with -1s, but the column that it correspond tho the label of the sample which is 1. For example, the first sample of the data set is labeled as "2" (*seboreic dermatitis*), so the class in the first row of $\mathbf{T}$ it shall have:

$$-1, 1, \ -1, \ -1, \ -1, \ -1$$

5. Transform $\mathbf{X}$ to zero mean and unit variance;

6. Randomize the rows of $\mathbf{X}$ and do not forget to put $\mathbf{T}$ in the same row order of $\mathbf{X}$;

7. Divide $\mathbf{X}$ into $\mathbf{X}_l$ (learning data) and $\mathbf{X}_t$ (testing data). In a way that $\mathbf{X}_l$ must have 80% of the data of $\mathbf{X}$ and the rest must be in $\mathbf{X}_t$. Do the same with $\mathbf{T}$: 80% of the rows of $\mathbf{T}$ for $\mathbf{T}_l$ (learning) and resting 20% for $\mathbf{T}_t$ (testing);

8. Use $\mathbf{X}_l$ and $\mathbf{T}_l$ in training phase;

9. Evaluate results with $\mathbf{X}_l$ and $\mathbf{T}_l$ collecting the *classification accuracy* and the *mean squared error*;

10. Repeat steps 6 to 9 fifty times. Then inform the maximum classification rate, the minimum classification rate, mean classification rate and the standard deviation of the classification rates. Do it for all classifiers and then build a table with the results.

Exercise 02

Using the Data Set available in this course's website, compare *Support Vector Regression* (SVR) with *Gaussian Processes for Regression* (GPR) and *Bayesian linear regression* (BLR) in terms of *mean squared error*

$$\text{MSE} \;=\; \frac{1}{N}\sum_{n=1}^{N}(y_n - t_n)^2 \,.$$

Just like the previous exercise, you can use a third-party software/API/toolbox to make use of SVR. The data was generated using the following model:

$$t_n \;= y(\mathbf{x}_n) + \epsilon_n$$

where ϵ_n is a gaussian noise in the n-th sample. In this problem, the gaussian noise has a standard deviation of $\sigma = 2$.

The data available in the website of this course are divided in 4 subsets. Follow the instructions below:

- Use `data-input-learning.txt` and `data-output-learning.txt` in the training phase. For training phase, there are 3000 samples.
- Use `data-input-testing.txt` and `data-output-testing.txt` in the testing phase. For testing, there are 1000 samples. Use these samples to evaluate the *mean squared error.*
- Each sample has 5 variables.
- Plot the measured outputs t_n and the predicted outputs $y(\mathbf{x}_n)$ of each regression algorithm (SVR, BLR and GPR).
- Build a table containing the *mean squared error* of each regression algorithm (SVR, BLR and GPR).

Where to find SVM/SVR Software

A list of available SVM-related softwares/APIs can be found at
`http://www.kernel-machines.org/software/`.

References

[1] M. Lichman. UCI machine learning repository, 2013.