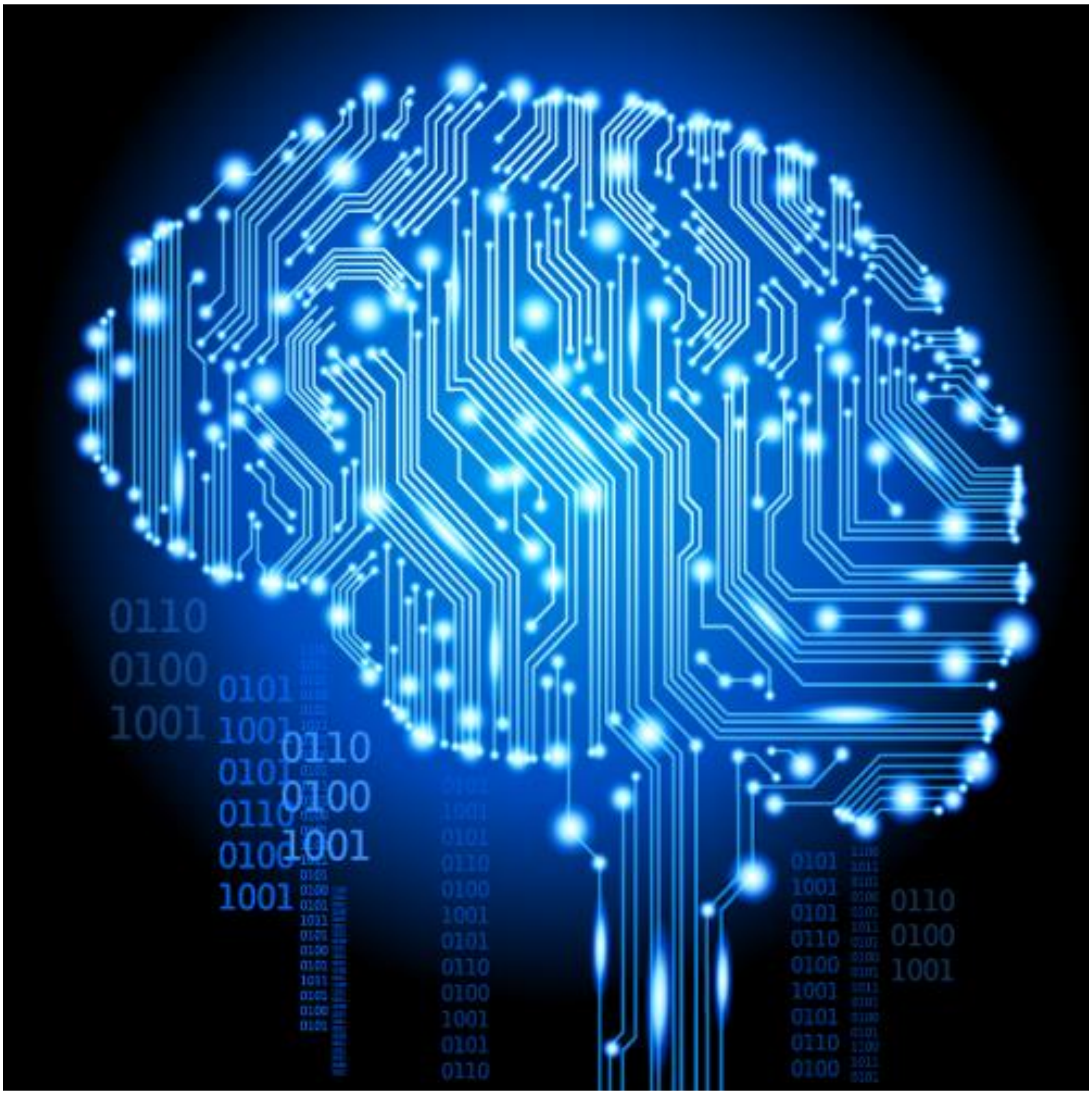
The University of Nottingham

UNITED KINGDOM · CHINA · MALAYSIA

COMP3009 Machine Learning

Computer Based Coursework Manual – Autumn 2021

Assessed coursework 1: Support Vector Machines



The University of Nottingham
UNITED KINGDOM · CHINA · MALAYSIA

Assignment 1: Support Vector Machines

This is the first set of assessed labs. In this assignment, you will train Support Vector Machines (SVMs) using different kernels, set hyper-parameters using inner-cross-fold validation. You will implement these using Matlab. From here on, you will **have to** work as a group, handing in a single report and code by **11th Nov, 2021 at 4 pm** UK time on Moodle. You can split and distribute the work to individual members, but each individual is expected to understand every aspect of the work.

Data

As a group, you will have to select *two* datasets: one for regression, and one for classification. You have to choose your dataset from the UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/index.php). After selection of the datasets, please have your group's assigned lab assistant approve your choice of datasets. Approval depends on: (a) uniqueness of your combination of datasets (every group must have a unique combination); (b) suitability in terms of difficulty (very large datasets make evaluation time-consuming). Please use the MS Teams channel, in-person lab support or the lab assistant's university email to send your selection of datasets.

Task 1: Train SVMs with the linear kernel

Use as the classifier a linear SVM, with the box-constraint C (which determines the importance of the slack-variables) set to be always one (i.e. no hyperparameters to tune). You have to build models for binary classification and regression. For binary classification, use the function:

```
Mdl = fitcsvm(X,Y, Name, Value);
```

to train a model, where X are the features and Y the labels of a two-class problem, and 'Name', 'Value' are variable name and value pairs to set. To set the kernel to linear, use the pair

```
'KernelFunction','linear'
```

and to set the box-constraint parameter to 1, use:

```
'BoxConstraint',1
```

So the full command would be:

```
Mdl = fitcsvm(X,Y,  'KernelFunction','linear', 'BoxConstraint',1);
```

For regression, use

```
Mdl = fitrsvm(X,Y, Name, Value)
```

to train a model. Note that for regression, you have to set the value of the parameter 'Epsilon' to something sensible. Note epsilon is half of the error-insensitive tube diameter. Try a few different values of 'Epsilon'. The model's performance is not important at this point. Report your findings.

Task 2: Train SVMs with Gaussian RBF and Polynomial kernels

Use Matlab's built-in fitcsvm/fitrsvm to train models, now using RBF and Polynomial kernels. Write an inner-fold cross-validation routine for finding the optimal hyperparameters. For classification, these are C and the kernel parameters: 'KernelScale', sigma for RBF and 'PolynomialOrder', q for the polynomial kernel. For regression, you have to set 'Epsilon' in addition. Do not use the automatic parameter optimisation methods. You must write your own, using only basic functions (although you can compare against that if you wish to check if you're doing the right thing). You may want to do these tests on a subset of the data if training is too slow.

Report for each model you trained how many support vectors were selected, both in absolute terms and in terms of a % of the training data available.

Task 3: Method Evaluation

Perform 10-fold cross-validation on the binary classification and regression tasks. Report results for linear, Gaussian, and polynomial kernels. Optimise hyper-parameters with the inner-cross-validation procedure developed in Task 2.

Report on the classification rate for the classification problems and RMSE for the regression problems. You may once again opt to use less data (if your selected data is too large) for your evaluation.

Task 4: Additional Questions

Answer the following three questions in your report:

Question 1: What does the kernel parameter of the Gaussian RBF kernel signify (sigma)? What happens when you increase its value?
Question 2: Explain what happens when a hard-margin SVM is fit to a dataset of two classes with overlapping features. What value do you need to set C (the slack-variable hyper-parameter) to attain a hard-margin SVM?
Question 3: Explain why do you need to use both inner cross-validation and an outer k-fold cross-validation (known as nested cross-validation) in the machine learning process?

Deliverables

For the completion of Assignment 1, the following have to be submitted on Moodle:

1. The code for training and evaluating SVMs, including the loading and transforming of the data, and code for performance evaluation.

2. A report of up to 1000 words (excluding result matrices and graphs) containing: a cover sheet, a brief introduction, all the elements asked for in each task, and a brief conclusion.

Marking Criteria

Clarity of report: 10%
Code quality: 20%
Implementation correctness: 10%
Parameter setting: 15%
Method evaluation: 15%
Questions: 30%