Assignment

STAT30270 – Statistical Machine Learning

Deadline - Friday 11th April 2025 at 17:00

Exercise 1

Consider the following training data:

$$\mathbf{X} = \begin{bmatrix} -0.96 & 3.36 \\ 0.89 & 2.11 \\ -1.44 & 3.18 \\ -3.02 & 3.77 \\ -0.17 & 1.43 \end{bmatrix}, \qquad \mathbf{y} = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \\ 1 \end{bmatrix},$$

where $\mathbf{X}$ are the inputs, and $\mathbf{y}$ is the binary vector corresponding to the target variable. A logistic regression model is estimated on these data using gradient descent with step size parameter of 0.3. At iteration t, the algorithm returns the following estimate of the weights:

$$\hat{\mathbf{w}}^{(t)} = (0.5,\ 0.8,\ -0.2).$$

1. Compute the estimate $\hat{\mathbf{w}}^{(t+1)}$ at iteration $t+1$ of the gradient descent algorithm. Show clearly all computations.

   (5 marks)

2. Show numerically that the loss function decreases when updating the parameters from $\hat{\mathbf{w}}^{(t)}$ to $\hat{\mathbf{w}}^{(t+1)}$.

   (5 marks)

Exercise 2

Consider the quantities: accuracy (a), sensitivity (s), specificity (e) and prevalence (p).

1. Show that:

   $$a = ps + (1-p)e$$

   (5 marks)

2. Given a prevalence of $p = 0.5$, show that a random classifier that assigns the positive class with probability θ has an expected accuracy of 0.5 for any value of $\theta \in [0, 1]$.

   (5 marks)

3. Given a prevalence $p = 0.1$, show that a random classifier that assigns the positive class with probability $0.1 \leq \theta \leq 0.5$ has, at best, an expected accuracy of 0.82.

   (5 marks)

Exercise 3

Consider some two-dimensional input observations $\mathbf{x}_i = (x_{i1}, x_{i2})$ and the associated target variable $y_i \in \{-1, +1\}$. A soft-margin support vector classifier has been trained on these data. The associated separating hyperplane is given below:

$$\{x_1, x_2 \,:\, 2 - 3x_1 + x_2 = 0\}$$

Consider the following subset of data points $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, 5$:

$\mathbf{x}_i$	x_{i1}	x_{i2}	y_i
$\mathbf{x}_1$	1	2	+1
$\mathbf{x}_2$	0.5	0	+1
$\mathbf{x}_3$	1.5	1.5	-1
$\mathbf{x}_4$	3	4	+1
$\mathbf{x}_5$	2	2	-1
$\mathbf{x}_6$	1	-0.5	-1
$\mathbf{x}_7$	-1	-3.5	-1
$\mathbf{x}_8$	1	0.5	-1

1. Which data points **are guaranteed** to correspond to support vectors? Justify your answer.

   (5 marks)

2. Recall the constraint for a soft margin support vector classifier:

   $$y_i(w_0 + \mathbf{x}_i^\top \mathbf{w}) \geq 1 - \xi_i, \qquad \xi_i \geq 0, \qquad i = 1, \ldots, 8.$$

   For each data point $\mathbf{x}_i$ compute the **minimal feasible** ξ_i given the constraint and the values of w_0 and $\mathbf{w}$. What do you notice? (**Hint:** Try to answer the question "What is the smallest value of ξ_i that ensures the constraint is fulfilled?")

   (5 marks)

3. Compute the accuracy of the estimated support vector classifier on the above data.

   (5 marks)

Exercise 4 – Data analysis

Processminer data set

The data are a subset of a larger dataset provided by *Processminer*, originally collected in a pulp-and-paper manufacturing facility. In this setting, a single machine takes in various raw materials and outputs reels of paper.

Multiple sensors are installed throughout the machine, measuring both the characteristics of the input materials (e.g., pulp fiber content, chemicals, etc.) and relevant process variables (e.g., blade type, couch vacuum, rotor speed, etc.).

Paper manufacturing is essentially a continuous rolling process. However, on occasion, the paper breaks. When a break happens and it is detected, production is stopped, the reel is removed, any identified issues are fixed, and the process is restarted. This results in a short downtime and minor production delays.

If a break goes unnoticed, however, it can cause significant disruptions: damage to machinery, contamination of subsequent production runs, and defective output that cannot be used. This leads in increased material waste, higher operational costs, and prolonged downtime as additional corrective measures are taken.

The goal of this project is to build a predictive model that accurately identifies breaks in the paper during process, using the data gathered by the sensors. This will allow for timely intervention, enabling production to be paused before further damage occurs.

The dataset is available in the object `data_processminer`, provided in the file `data_assignment_processminer.RData`. The dataset includes the following variables:

- `y`: Target binary variable indicating a break (`break`) or not (`no_break`) during the process.
- `x1` - `x59`: Continuous input variables representing sensor readings related to raw materials and process variables. Descriptions are omitted for anonymity.

Predicting paper breaks during the manufacturing process

The goal is to develop a supervised learning model to predict whether a paper break will occur, based on certain conditions as provided by the sensor measurements.

1. Implement logistic regression and support vector classifiers to predict paper breaks in the process given the numerical sensor reading features. Use an appropriate framework to compare and tune the different models, evaluating and discussing their relative merits. Select the best model for predicting potential paper breaks in the process.

   (50 marks)

2. Use appropriately some test data to evaluate the generalized predictive performance of the best selected classifier. Provide a discussion about the reliability of the selected model at predicting paper breaks in a real world scenario.

   (10 marks)

Guidelines:

- **Clearly justify and explain all decisions made during the analysis.**.
- If you wish, you can use only a subset of the features of the data in the model building stage. However, for full marks you **must clearly motivate your choice** and explain why some features are discarded.
- You will not be evaluated on the basis the predictive performance of your classifiers, but you would need to show that attempts have been considered to build a classifier with reasonable performance.

Submission rules and instructions

- Write a short and tidy report and submit it as a single pdf file (approximately max 10-12 pages, code excluded).
- Include the R code used for the data analysis in the report. The report can be produced using R Markdown (or similar tool), with the code included in the main text or as an appendix. **The code must be working and the analysis must be reproducible in all parts**.
- In general, for full marks you **must explain** concisely and clearly **all reasoning**, as well as **show all steps** and **computations** in your answers. Correct answers alone will not achieve full marks.
- You can approach/solve the exercise questions (1, 2, 3) using R. However, **bear in mind**: 1. During the exam you will be required to solve exercises using pen and paper; 2. You will need to show all steps and computations for full marks, merely displaying results from running code will not give full marks.
- For the data analysis task, the use of the `caret` package (or packages with similar automatic tuning and model comparison functionalities) is not allowed.
- For the data analysis task, submitting only code without any output, commentary, or discussion will not receive any marks.
- Multiple submissions before deadline are allowed and only the latest one will be considered for marking.
- **Submission after deadline will incur in penalization as UCD rules**. See "Module details" document under the "General information" tab on Brightspace.
- **Plagiarism is strictly prohibited and will result in severe penalties**. See "Module details" document and "SMS academic integrity protocol" under the "General information" tab on Brightspace to review which actions constitute plagiarism and further information.
- **Any instance of plagiarism** will result in a **zero grade** in this assessment component **for all students involved**.
- By submitting your solutions document to this assignment, you confirm that you have read and understood the regulations and policies regarding assessment and plagiarism outlined here, in the "Module Details" document, the "SMS Academic Integrity Protocol", and related documents. You also agree to abide by all the regulations stated in these documents.