PHYS379: Group Project - Machine Learning Topic

Edward McCann

Department of Physics, Lancaster University, Lancaster, LA1 4YB, UK

Keywords: machine learning; supervised learning; linear regression; logistic regression

Introduction

In PHYS379, you should undertake an open-ended group project and submit a group report describing the project. The group project is worth twenty credits. Please see the document "PHYS379-group-project-notes.pdf" for details about organisation and assessment.

Open-ended project

Your task in the project is to model the properties of physical systems using machine learning. Precisely what you do is your choice - the project is open-ended. You will need to formulate a research question (or questions) that is more specific than simply "model the properties of physical systems using machine learning". What are you trying to discover? You should do some research in the library or on the internet to find out what the interesting questions are. It may be that you will not finalise the research question until the latter stages of the project, once you have found what is (or is not) interesting. Remember that this is group work - it will make sense to plan for parallel tasks in order to use your personnel effectively. For example, you might want parallel numerical and analytic investigations, investigations looking at different aspects of the same problem, or you might deliberately duplicate effort in order to check important results (in the "real" world, results should always be checked before publication, getting different people to work independently is the best way to do this). There are some ideas below, you can follow some or all of them, or you can follow your own ideas.

Getting started

Machine learning

Machine Learning is the act of a computer doing some task, without being explicitly programmed to do that task [1-4]. One example is supervised learning, in which the program takes a set of input variables x_{ij} and their corresponding outputs y_i (the "training set"). It then finds a fit between these variables in order to predict the output on new, unseen data (the "test set"). There are m data points with index $i = 1, \cdots m$, and each input variable has n features/characteristics with index $j = 1, \cdots n$. There is also unsupervised learning, where the program finds some pattern in a set of inputs, such as finding clusters in the data, or detecting anomalies.

Linear Regression

The simplest form of supervised learning is "linear regression" [2,5]. The program takes the training set, and finds the line of best fit through this data. This line can then be used to predict the results from the previously unseen test set, and the accuracy of the predictions quantify the quality of the model. Regression algorithms attempt to minimise the "cost function" (or "loss function"), which, for linear regression, is defined as

$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}(h_{\theta i} - y_i)^2 \qquad (1)$$

$$h_{\theta i} = \theta_0 + \sum_{j=1}^{n} x_{ij}\theta_j \qquad (2)$$

where $J(\theta)$ is the cost function, $h_{\theta i}$ is the "hypothesis function" for the regression, x_{ij} and y_i are the input and output variables for some datapoint, and θ is the set of parameters the algorithm changes to minimise $J(\theta)$. Note that, for linear regression, the hypothesis function is simply a straight line through the data.

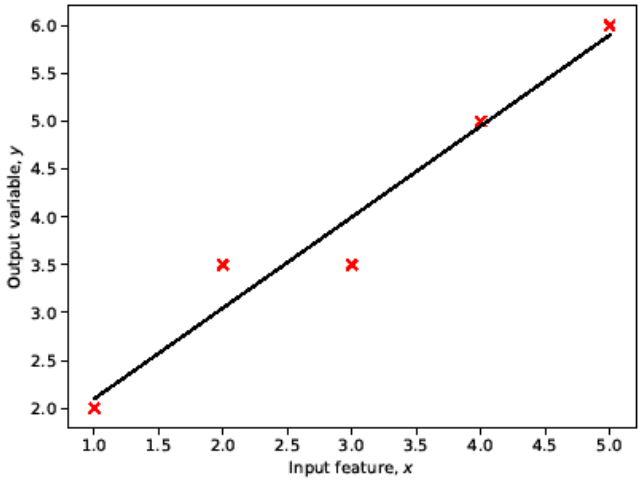

Figure 1: Example data set for linear regression with a single feature ($n = 1$). The dataset used is shown in Table 1, and has a line of best fit $y = 1.15 + 0.95x$.

x	y	Contribution to cost function $J(\theta_0 = 0, \theta_1 = 0)$	Contribution to cost function $J(\theta_0 = 1, \theta_1 = 1)$	Contribution to cost function $J(\theta_0 = 1.15, \theta_1 = 0.95)$
1	2	4	0	0.01
2	3.5	12.25	0.25	0.2025
3	3.5	12.25	0.25	0.25
4	5	25	0	0.0025
5	6	36	0	0.01
Total cost function $J(\theta)$		8.95	0.05	0.0475

Table 1: An example of how the datapoints from Fig.1 contribute to the cost functions for different values of θ, using $h_{\theta i} = \theta_0 + \theta_1 x_i$, and calculating each contribution as $(h_{\theta i} - y_i)^2$. The total cost function is determined as $J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}(h_{\theta i} - y_i)^2$ and the final column shows the minimum value of $J(\theta)$ used to create the straight line fit in Fig.1. This minimum value is determined through the iterative process outlined in Eq.(3).

The simplest way to minimise the cost function is using gradient descent. This algorithm calculates the gradient of the cost function for each θ, then moves downhill, to approach the minima of the function. An implementation of gradient descent is given by

$$\theta_i = \theta_i - \alpha \frac{d}{d\theta_i} J(\theta_0, \theta_1, \cdots, \theta_n) \qquad (3)$$

where α is the "learning rate" of the system. Repeating Eq.(3) will result in convergence to a minimum for sufficiently small α. An example of gradient descent is shown in Fig.1 and Table 1 for a simple example with only one feature ($n = 1$). In this case, we can simply write $h_{\theta i} = \theta_0 + \theta_1 x_i$. We begin with an initial guess of $\theta_0 = 0, \theta_1 = 0$, and the gradient descent algorithm finds the minimum of the cost function to be 0.0475, with parameters $\theta_0 = 1.15, \theta_1 = 0.95$.

For a more complex system, with a larger number of features, the data must be normalised so the range of each feature is approximately the same. This allows you to compare the importance of each feature in the regression algorithm, with larger values of θ_i corresponding to a more important feature.

Logistic Regression

It is also possible to use machine learning in classifying data e.g. how likely is it a student will get into university with a particular set of grades? The minimisation process in logistic regression is similar to linear regression, however the hypothesis function and cost function are given by

$$h_\theta(x) = \frac{1}{1 + e^{-z}} \qquad (4)$$

$$J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \mathrm{Cost}(h_\theta(x), y) \qquad (5)$$

$$\mathrm{Cost}(h_\theta(x), y) = \begin{cases} -\log\big(h_\theta(x)\big) & if\ y = 1 \\ -\log\big(1 - h_\theta(x)\big) & if\ y = 0 \end{cases}$$

where $\boldsymbol{z}$ is the line of best fit $\boldsymbol{\theta_0} + \sum_{j=1}^{n} \boldsymbol{x_j \theta_j}$ [2,6]. This new hypothesis function gives the probability of the output being 1 i.e. if $\boldsymbol{h_\theta(x) = 0.7}$ this represents a 70% chance of the output being 1.

Once the cost function has been minimised, you can use your results to predict new results. How does varying $\boldsymbol{\alpha}$ affect the results? How close does the algorithm get to analytical solutions? Can you increase the accuracy of the model by including polynomial features ($\boldsymbol{y = x^2}$)?

Open-ended part

Please note:

1. **The aim is for your group to do its own, original calculations.** A literature review and, perhaps, reproduction of other people's results may be important steps along the way, but these are not the final aim.

2. The machine learning topic has run in PHYS379 for three years and it has generally been very successful. One difficulty is finding original data that hasn't been analysed before; a possible solution is to try to generate your own. Another common difficulty is under- or over-fitting (i.e. using a machine learning algorithm / neural network that is inappropriately complex for the given problem).
3. You will be able to find machine learning algorithms pre-written (e.g. for Python, Maple, MATLAB). Be very wary of using them, because (i) if someone else has written the code, how much credit will you get in PHYS379? (ii) possibly more importantly, if someone else has written the code, to what extent do you understand what it actually does?

Ideas include:

- Prediction of semiconductor or superconductor properties using linear regression.
- Surveying a group of people to see if there's a trend (e.g. do people with blue eyes prefer tea to coffee?). Note that you can't ask survey participants for "protected characteristics", as defined by Ref.[7].
- Prediction of song genre using the Million Songs Database, see [8] and the paper associated with it [9].
- Independent Component Analysis to unmerge images or waveforms using the algorithm outlined in [10].
- Cluster Analysis for segmenting students based on their interests [11].
- Logistic Regression for handwriting recognition [12].
- Logistic Regression for spam filtering using the bag-of-words model [13].
- Classifying social media posts based on their language.
- Make your own physics simulation, generate learning examples, then train machine learning to see how well it makes predictions subsequently. e.g. the path of a projectile travelling through a set of different masses with different locations.
- Game or sports outcome predictor. e.g. given the state of a chess match after 20 moves, is it possible to predict the final outcome?
- See [14] and Refs therein for an interesting (but brief) discussion of the interpretation of machine learning models.

Acknowledgement

Thanks to Alex Warwick who devised this topic during the summer of 2019.

Bibliography

[1] "Machine Learning", https://en.wikipedia.org/wiki/Machine_learning

[2] "The Elements of Statistical Learning" by T. Hastie, R. Tibshirani, and J.H. Friedman, (Springer, New York, 2001).

[3] "Pattern Recognition and Machine Learning" by C.M. Bishop (Springer, New York, 2006).

[4] "Machine Learning | Andew Ng'', https://www.youtube.com/watch?v=PPLop4L2eGk

[5] "Linear Regression", https://en.wikipedia.org/wiki/Linear\textunderscore regression

[6] "Logistic Regression", https://en.wikipedia.org/wiki/Logistic\textunderscore regression

[7] “Equality Act 2010”, https://www.legislation.gov.uk/ukpga/2010/15/section/4

[8] “The Million Songs Database”, http://millionsongdataset.com

[9] T. Bertin-Mahieux, D.P.W Ellis, B. Whitman and P. Lamere, in Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR 2011), 2011.

[10] A. Hyvärinen and E. Oja, Neural Networks. **13** (4-5), 407 (2000).

[11] “K-means Clustering”, https://en.wikipedia.org/wiki/K-means_clustering

[12] “Optical Character Recognition”, https://en.wikipedia.org/wiki/Optical_recognition

[13] “Bag-of-words Model”, https://en.wikipedia.org/wiki/Bag-of-words_model

[14] “Interpretability of machine-learning models in physical sciences” L. M. Ghiringhelli, arXiv:2104.10443, https://arxiv.org/abs/2104.10443