Machine Learning Introduction Tutorial
COMS3007

Meir Rosendorff (1490527), Kimessha Paupamah (1038238),
Devon Jarvis (1365149)

February 18, 2018

1 Question 1

a) Supervised learning (Classification)- Participants are classified into two discrete groups, either wearing glasses or not wearing glasses.

b) Reinforcement learning - Iterative and easy to create a reward function and simulate a game.

c) Supervised learning (Classification) or Unsupervised learning- If a large labelled dataset is given then use supervised learning, if not use unsupervised learning so that the algorithm can determine categories to classify the vehicles.

d) Supervised learning (Classification)- Case outcome is either guilty or not guilty, two discrete groups.

e) Supervised learning (Classification or Regression)- Performance has a continuous solution space if predicting exact marks. If predicting pass or fail then it would be a classification problem. Likewise if predicting a category (eg: 50% - 60%) then this too would be a classification problem. We determined this could not be an unsupervised learning problem as due to the fact that the solution space is continuous, unsupervised learning methods may struggle to find ways of splitting the output from the network to form discrete clusters or groups.

f) Supervised learning (Classification)- Input of symptoms, treatments and results of past medical records and can map to procedures of future treatment.
Unsupervised learning- Can cluster symptoms and have the algorithm categorise illnesses based on similar treatment needed.

g) Reinforcement learning- Reward is easily quantified by fulfillment of achievements or goals or by measuring score and the agent can easily interact with the game environment.

h) Supervised learning (Classification or Regression)- Prices are continous and thus predicting exact prices would be a regression problem, however a neural network can also determine if there will be a drop or rise in prices. This would be a classification problem as these would form two discrete groups.

2 Question 2

The prediction of cancer susceptibility (risk assessment), prediction of cancer recurrence and the prediction of cancer survival are all uses of supervised learning in industry. The above includes objectives such as detection of tumors as well as the prediction/prognosis of a cancer type. The integration of features into the model such as family history, age, diet, weight, high-risk habits and exposure to environmental carcinogens play a critical role in predicting the development of cancer. However, these type of features alone do not provide a robust decision from the model and other features such as molecular biomarkers, cellular parameters as well as the expression of certain genes have been proven as very informative indicators for cancer prediction and need to be included in the model. This is an example of a classification supervised learning problem as patients are grouped into classes based on the type of the cancer, high or low probability of the cancer reoccuring and the severity of the cancer (possibly on a discrete scale, like 1-10). A limitation to this method is that generally implementations of the model do not generalise to different types of cancer when implemented to perform tasks such as determining severity of the cancer or cancer susceptibility. This is viewed as quite a fundamental limitation. [1]

3 Question 3

Currently one of the most prevelent uses of unsupervised learning in industry is the use of clustering for anomaly detection to detect intrusions for network security. This technique establishes normal usage patterns and detect intrusions by observing deviations from these patterns. The self-organising map (SOM) is a common algorithm used for clustering and has been applied to this domain successfully in the past. It does suffer from the fact that while increasing the number of neurons in it's architecture will improve the resolution of the map, the computation time will also significantly increase as a result. The SOM is traditionally trained by using normal network traffic. The trained SOM reflects the distribution of the normal network connections. A connection is represented by 41 features, which include the basic features of the individual TCP connections, the content features within a connection, and the traffic features computed by using a two-second time window. After training each cluster is labelled according to the majority type of data in that cluster. [2]

4 Question 4

a) Soundwaves from the noises the animal makes, design of the animals claws, nose wetness (healthy dogs have wet noses and healthy cats have dry noses)

b) Nose-shape, face-shape, eye shape.

c) Weight of the bird, wing-span, general size, bone-structure and density of the species of bird.

d) Previous marks vs current marks, sleep hours per night, course averages.

e) Loudness, tempo, pitch. (all obtained from sampeling soundwaves from the music)

f) History of games played, fitness of team members, whether or not the team is playing at home.

g) BMI of target market, area of supermarket, history of chocolates bought, time of the year.

5 Question 5

a) Soundwaves - record the noises made by dogs (e.g barking) and cats (e.g purring); claws - measure claws; noses - dogs have wet noses, cats have dry noses.

b) Online databases have already labelled images, if these databases are not easily accessible photos may need to be taken and labelled manually.

c) Weight - weigh the bird; wing-span - measure the wing-span; size - measure height and width of bird; bone structure - study bone fragments from birds of same species.

d) Marks - kept track of in student record or database; Sleep schedule - ask students to record their sleeping patterns; Averages - kept track of in course record or database.

e) Obtained from measuring different measurements of wavelengths and sound waves from the music.

f) History - kept track of in records, fitness - measure fitness, home ground - this is easily identified from the league schedule or fixture list.

g) BMI - can be calculated from body measurements; history - kept track of in customer database; time of year - date.

6 References

[1] Kourou, K., Exarchos, T.P., Exarchos, K.P., Karamouzis, M.V. and Fotiadis, D.I., 2015. Machine learning applications in cancer prognosis and prediction. Computational and structural biotechnology journal, 13, pp.8-17.

[2] Lei, J.Z. and Ghorbani, A., 2004, May. Network intrusion detection using an improved competitive learning neural network. In Communication Networks and Services Research, 2004. Proceedings. Second Annual Conference on (pp. 190-197). IEEE