EE3-23 MACHINE LEARNING COURSEWORK 2019

Provide answers to all 5 problems given on the next page. Maximum number of possible mark points are given below each problem. Each question adds up to **[100 points]**, which will be averaged over all 5 questions, multiplied by 0.2 and added to the marks obtained from the exam [80 points].

Much of reporting in the domain of Machine Learning is done using Latex, which is a document preparation system for high-quality typesetting available for Linux, Windows and Mac. Present your solution within this Latex file i.e. edit the source file and compile. The solutions must be clearly presented with equations or figures if appropriate but also succinctly. Extensive copying of lecture material indicates weak understanding and poor effort, so points will be deducted for including irrelevant information. Use mathematical formulations as much as possible, rather than verbal descriptions.

Do not: 1) remove the questions, 3) change the document style.

Submission deadline: 21 November 2019, PDF file only via Blackboard
Submission format: PDF file only via Blackboard, filename: username.pdf (e.g. ws217.pdf)

Declaration:
I, <YOUR NAME>, pledge that this assignment is completely my own work, and that I did not make use of this coursework from any other person, and that I did not allow any other person to make use of portions of my coursework. I understand that if I violate this honesty pledge, I am subject to disciplinary action pursuant to the appropriate sections of Imperial College London.

1. Practical ML scenario.

a) A supermarket chain would like to commission a face recognition system to automatically apply discount to customers with clubcard memberships. They can provide face images (100x100 resolution) of 1M customers with membership and 1M without (customer consent according to GDPR). They estimate the long term cost of loosing an unhappy customer due to incorrect recognition to be 5k and a cost of applying discount to a non member as 100 over a year. ML solution will be useful if it can guarantee that the test error will not differ from the training error by more than 10% with 90% certainty.

i) Identify relevant ML components and formulate it as an ML problem. Discuss the error in detail. [30 marks]

ii) What is required to guarantee that the predictor meets the criteria? [10 marks]

iii) Assume you choose a polynomial of degree $k = 10$ and a linear classifier as the predictor. What is the approximate number of data points that are needed if you want to guarantee that the test error of the predictor is within 0.05 from the training error with probability 0.99? Show your calculations. [10 marks]

b) Understanding of the relation between the training error, the validation error, the model complexity, and the amount of training data. Draw figures :

i) Errors vs complexity. [10 marks]

ii) Errors vs λ regularisation. [10 marks]

iii) Errors vs the number of data points. [10 marks]

Discuss 5 typical cases that may be observed during training, and point them on the curves :

A) Complexity of the hypothesis class is too high (overfitting). [4 marks]

B) λ gives optimal regularisation for the given learning setup. [4 marks]

C) λ too high, regularization too strong (underfitting). [4 marks]

D) Complexity of the hypothesis class is too low (underfitting). [4 marks]

E) The number of data points is to low, hence overfitting. [4 marks]

2. On Jun 9, 2016, ORB published a public poll result about Brexit: they asked 2,052 people, 55% of them supported leaving the EU while 45% wanted to stay. Assume that the people's decisions are independent and identically distributed.

a) With what confidence level can you reject the hypothesis that the probability of choosing to stay in the EU is at least 50% (i.e., the majority of people would like to stay in the EU). [40 marks]

b) How would the confidence in predicting the referendum outcome change if the poll was repeated on 2,052 people with the result of 52% remain, 48% leave? [30 marks]

c) How many people would have to be asked to have 99.9999% confidence that the final election would give the same result in case b)? [30 marks]

3. Consider neural network architecture given in Figure **??**

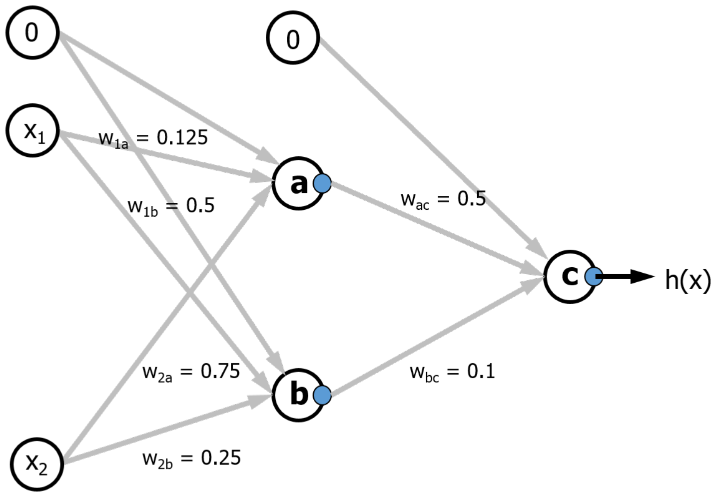


Figure 3.1 Neural network architecture

a) Forward propagation

i) Write down the formula for Maxout non-linearity activation function, draw this function and its gradient in a figure. [10 marks]

ii) What is the advantage and disadvantage of using Maxout over ReLU activation? [10 marks]

iii) Given input vector $\mathbf{x} = (8, -8)$ use the ReLU non-linearity to calculate outputs of all neurons and $h(x)$. [30 marks]

b) Backpropagation

i) Apply the backpropagation algorithm to calculate the update of weight $w_{2,a}$. Use L_2 loss without regularization, ReLU activation, learning rate of 0.2 and training example of $\mathbf{x} = (8, -8)$ with its label $y = 1.4$. [35 marks]

ii) What techniques can be used when training a neural network to avoid overfitting? [15 marks]

4. a) Support Vector Machines

You are given the following training data examples that belong to two classes:

Positive examples: $\begin{bmatrix} -3 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} -2 \\ -2 \end{bmatrix}, \begin{bmatrix} 2 \\ 2 \end{bmatrix}, \begin{bmatrix} 4 \\ 1 \end{bmatrix}$

Negative examples: $\begin{bmatrix} -2 \\ -4 \end{bmatrix}, \begin{bmatrix} 0 \\ -5 \end{bmatrix}, \begin{bmatrix} 1 \\ -3 \end{bmatrix}, \begin{bmatrix} 2 \\ -2 \end{bmatrix}, \begin{bmatrix} 4 \\ -2 \end{bmatrix}$

i) Plot these points, and draw the maximum margin SVM classifier. Which of the training examples constitute the support vectors? [16 marks]

ii) The hyperplane $H(w,b)$ for this SVM classifier is defined as $w^T x + b = 0$, where $||w||^2 = 1$. Characterize the w and b parameters for the above training data. [12 marks]

iii) Suppose you are given the following additional training examples:

Positive example: $\begin{bmatrix} 0 \\ 2 \end{bmatrix}$ Negative examples: $\begin{bmatrix} 7 \\ 1 \end{bmatrix}, \begin{bmatrix} -3 \\ -5 \end{bmatrix}$

What are the w and b parameters of the SVM classifier now? [8 marks]

b) Unsupervised learning

i) What are the differences between supervised and unsupervised learning? Give an example for each of them. [12 marks]

ii) Given a training dataset $x^{(1)}, \dots, x^{(n)}$, where $x^{(i)} \in \mathbb{R}^d$, the k-means clustering algorithm is given as follows:

1. Randomly choose k cluster centroids: $\mu_1, \dots, \mu_k \in \mathbb{R}^d$

2. Repeat until convergence

   For $i = 1, \dots, n$, set

$$c^{(i)} = \operatorname*{argmin}_j ||x^{(i)} - \mu_j||^2$$

   For $j = 1, \dots, k$, set

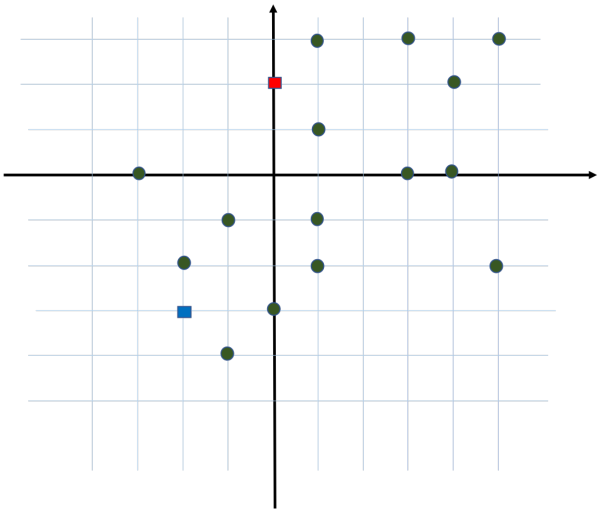
$$\mu_j = \frac{\sum_{i=1}^n \mathbb{1}\{c^{(i)} = j\} x^{(i)}}{\sum_{i=1}^n \mathbb{1}\{c^{(i)} = j\}}$$

   where $\mathbb{1}\{x\} = 1$ if x is true, and 0 otherwise.

   Explain each of the above steps. [20 marks]

iii) Consider the 14 data points $x^{(1)}, \dots, x^{(14)} \in \mathbb{R}^2$, represented by green dots in the figure below. The blue and red points indicate the initial centroids.

Run one iteration of the 2-means clustering algorithm: indicate which cluster each data point belongs to, and update the centroids.



[20 marks]

iv) What are the final centroids and the clusters if you iterate the 2-means algorithm until convergence? [12 marks]

5. Consider the Markov decision process (MDP) with three states: s_1, s_2 and s_3. There are two actions that can be taken at any state, denoted by a_1 and a_2. State transition probabilities when action a_1 is taken is given below:

a_1	s_1	s_2	s_3
s_1	0.8	0.1	0.1
s_2	0.6	0.3	0.1
s_3	0.1	0.5	0.4

For example, at state s_1, if we take action a_1, the next state is s_1 with probability 0.8, s_2 with probability 0.1, and s_3 with probability 0.1. State transition probabilities when action a_2 is taken is given as follows:

a_1	s_1	s_2	s_3
s_1	0.4	0.4	0.2
s_2	0.7	0.2	0.1
s_3	0.1	0.1	0.8

The reward gained at each state is independent of the action taken, and is given by $R(s_1) = 1$, $R(s_2) = 2$ and $R(s_3) = -1$. (Feel free to use a computer to evaluate the answers to some of the questions below. You don't need to report your code; just explain the steps taken.)

a) Write down the state-value function $v_\pi(s)$ for a discounted MDP. [10 marks]

b) Write down the Bellman equation for the state-value function. [10 marks]

c) Assume that action a_1 is taken at each step. Find the values of all the states for a discount factor of $\gamma = 0.9$. [10 marks]

d) Find the state values of the policy in part c) with $\gamma = 0.1$. [10 marks]

We have $v(s_1) = 1.10, v(s_2) = 2.12, v(s_3) = -0.92$.

e) Starting with an initial value function of $v_1(s_1) = v_1(s_2) = v_1(s_3) = 0$, and $\lambda = 0.9$, apply one synchronous iteration of the value iteration algorithm to compute a new value function $v_2(s)$. [20 marks]

f) Apply the value iteration 10 times to compute the values $v_{11}(s)$ for both $\lambda = 0.1$ and $\lambda = 0.9$. Compare these values to those obtained in parts b) and c). Which one do you expect to be greater? Why? [20 marks]

g) What are the optimal values of the states, and optimal actions to take for each state for $\gamma = 0.9$? How do these values compare to those found in part b)? [20 marks]