Exam Advanced Machine Learning
24 October 2019, 15.15–18.00

This exam consists of 6 problems, each consisting of several questions. All answers should be motivated, including calculations, formulas used, etc. The use of a calculator is not allowed.

Question 1: Logistic regression
Consider the problem of the binary classification task depicted in Figure 1 (left). We attempt to solve this with the simple linear logistic regression model:

$$\mathbb{P}(y = 1 \,|\, \mathbf{x}, \mathbf{w}) = g(w_1 x_1 + w_2 x_2) = \frac{1}{1 + e^{-w_1 x_1 - w_2 x_2}}.$$

For simplicity, we do not use the bias parameter w_0. The training data can be separated with zero training error: see line L_1 in Figure 1 (right), for instance.

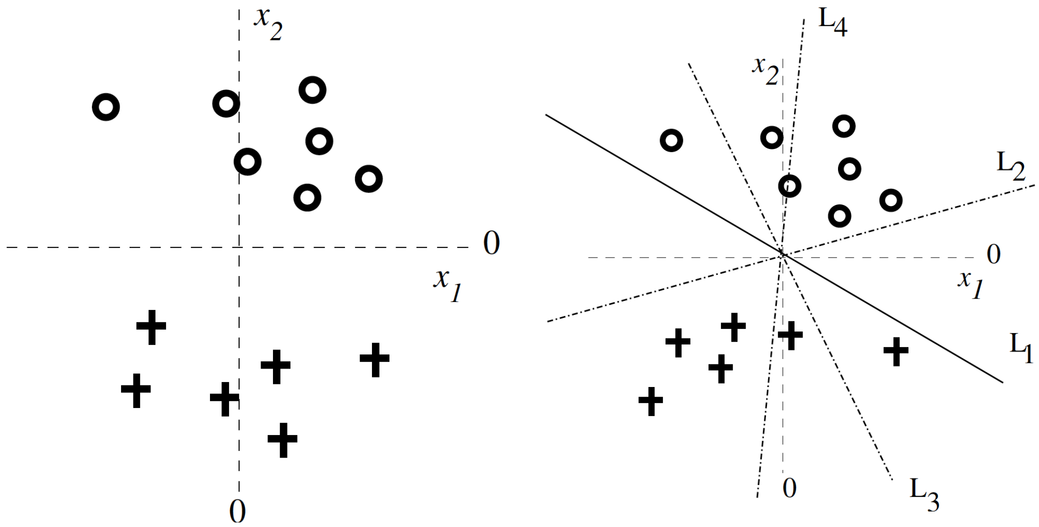
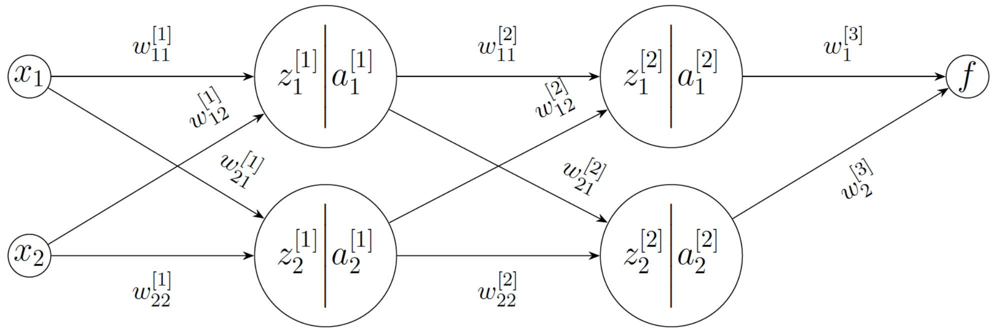


Figure 1: Classification problem

(a) Consider the regularization approach where we use regularization on *only* w_2. Thus, the regularization penalty is λw_2^2. We would like to know which of the four lines in Figure 1 (right) could arise from L_1 as a result of such regularization. For each potential line L_2, L_3, and L_4 determine whether it can result from regularization of w_2. Explain your answer.

(b) If we change the regularization to the absolute value and also regularize w_1, we get a regularization penalty having the form $\lambda\,(|w_1| + |w_2|)$. As we increase the regularization parameter λ, describe what will happen to the weights w_1 and w_2, and in which order.

Question 2: Neural networks
Consider this three layer network:



Let σ be the sigmoid function for the activations. Then, we have

$$Z^{[1]} = \begin{bmatrix} z_1^{[1]} \\ z_2^{[1]} \end{bmatrix} = \begin{bmatrix} w_{11}^{[1]} & w_{12}^{[1]} \\ w_{21}^{[1]} & w_{22}^{[1]} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad , \quad A^{[1]} = \begin{bmatrix} a_1^{[1]} \\ a_2^{[1]} \end{bmatrix} = \begin{bmatrix} \sigma(z_1^{[1]}) \\ \sigma(z_2^{[1]}) \end{bmatrix}$$

$$Z^{[2]} = \begin{bmatrix} z_1^{[2]} \\ z_2^{[2]} \end{bmatrix} = \begin{bmatrix} w_{11}^{[2]} & w_{12}^{[2]} \\ w_{21}^{[2]} & w_{22}^{[2]} \end{bmatrix} \begin{bmatrix} a_1^{[1]} \\ a_2^{[1]} \end{bmatrix} \quad , \quad A^{[2]} = \begin{bmatrix} a_1^{[2]} \\ a_2^{[2]} \end{bmatrix} = \begin{bmatrix} \sigma(z_1^{[2]}) \\ \sigma(z_2^{[2]}) \end{bmatrix}$$

(a) Given that $f = w_1^{[3]} a_1^{[2]} + w_2^{[3]} a_2^{[2]}$, compute the following four derivatives:

$$\delta_1 = \frac{\partial f(x)}{\partial z_1^{[2]}}, \qquad \delta_2 = \frac{\partial f(x)}{\partial Z^{[2]}}, \qquad \delta_3 = \frac{\partial f(x)}{\partial Z^{[1]}}, \qquad \delta_4 = \frac{\partial f(x)}{\partial w_{11}^{[1]}}.$$

(b) Explain why dropout in a neural network acts as a regularizer.

(c) Briefly explain one method for dealing with the problem of exploding gradients and one method for dealing with the problem of vanishing gradients in deep neural networks.

Question 3: Graphical models
We wish to develop a graphical model for the following transportation problem.

A transport company is trying to choose between two alternative routes for commuting between Rotterdam and Amsterdam. In an experiment, two identical trucks leave Rotterdam at the same but otherwise random time T. The trucks take different routes, arriving at their (common) destination in Amsterdam at times A_1 and A_2.

The transit time for each route depends on the congestion along the route, and the two congestions are unrelated. Let us represent the random delays introduced along the routes by variables C_1 and C_2. Finally, let F represent the identity of the truck that reaches Amsterdam first. We view F as a random variable that takes values 1 or 2.

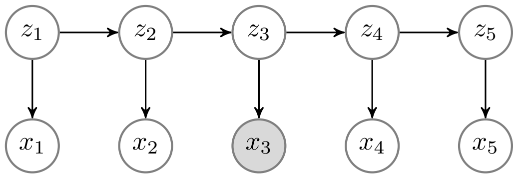
(a) Draw a graphical model with edges so that it captures the relationships between the variables in this transportation problem.

(b) Write the expression for the joint probability $\mathbb{P}(T, A_1, A_2, C_1, C_2, F)$ of the network in its *reduced* factored form.

(c) Check if T is conditionally independent of C_1 given F in the network. Provide a good argument for your answer.

Question 4: Hidden Markov Models (HMMs)
The following questions pertain to hidden Markov models. The subparts in this question are not related to each other.

(a) The following network depicts a sequence of 5 observations from a hidden Markov model, where z_1, z_2, z_3, z_4, and z_5 is the hidden state sequence. Are x_1 and x_5 independent given x_3? Please motivate your answer.



(b) Assume that the following sequences are very long and the pattern highlighted with spaces is repeated:

Sequence 1: $\quad$ 100 100 100 100 $\cdots$ 100

Sequence 2: $\quad$ 1 100 100 100 100 $\cdots$ 100

If we model each sequence with a different first-order hidden Markov model, what is the number of hidden states for the latent variable that a reasonable model selection method would report? Explain your answer and give the triplet (π, A, φ) for each sequence explicitly. Remember that π is the initial distribution for the latent variables, A the transition matrix for the latent variables, and φ the emission probabilities.

(c) Is the following statement true or false? The sequence of observed states sampled from a hidden Markov model satisfies the first-order Markov property. Please provide an argument for your answer.

Question 5: Reinforcement learning
We are using Q-learning to learn a policy in a system with two states s_1 and s_2, and two actions a and b. Assume that the discount factor $\gamma = 0.8$, and that the learning rate $\alpha = 0.2$. The current values of Q are

$Q(s_1, a)$	2.0
$Q(s_1, b)$	2.0
$Q(s_2, a)$	4.0
$Q(s_2, b)$	2.0

(a) Suppose that when we were in state s_1, we took action b, received reward 1.0, and moved to state s_2. Which item of the Q-table will change, and what is the new value?

(b) Do you agree or disagree with the following statement: Q-learning can only be used when the learner has prior knowledge of how its actions affect its environment. Please provide an argument to your answer.

(c) For Q-learning to converge, we need to correctly manage the exploration versus exploitation trade-off. What property needs to hold for the exploration strategy?

Question 6: Bayesian linear regression

Suppose that you are doing machine learning with linear regression models using basis functions $\varphi(\cdot)$. Thus, the prediction is given by $y(\mathbf{x}, \mathbf{w}) = \mathbf{w}^\top \varphi(\mathbf{x})$. We assume that the data points are drawn independently from the distribution $p(t \mid \mathbf{x}, \mathbf{w}, \sigma^2) = \mathcal{N}(t \mid y(\mathbf{x}, \mathbf{w}), \sigma^2)$, where σ^2 is the variance. We are applying a Bayesian framework to learn the parameters $\mathbf{w}$ by setting the prior distribution to $p(\mathbf{w}) = \mathcal{N}(\mathbf{w} \mid \mathbf{m}_0, \mathbf{S}_0)$ having mean $\mathbf{m}_0$ and covariance $\mathbf{S}_0$.

(a) Show by "completing the squares" that the posterior distribution $p(\mathbf{w} \mid \mathbf{t}) = \mathcal{N}(\mathbf{w} \mid \mathbf{m}_N, \mathbf{S}_N)$ where

$$\mathbf{m}_N = \mathbf{S}_N(\mathbf{S}_0^{-1}\mathbf{m}_0 + \mathbf{\Phi}^\top \mathbf{t}/\sigma^2),$$

$$\mathbf{S}_N^{-1} = \mathbf{S}_0^{-1} + \mathbf{\Phi}^\top \mathbf{\Phi}/\sigma^2.$$

(b) Suppose that we consider the prior $\mathbf{S}_0 = \alpha^{-1}\mathbf{I}$ for the covariance. Show that $\mathbf{m}_N$ converges to $\mathbf{m}_N = (\mathbf{\Phi}^\top \mathbf{\Phi})^{-1}\mathbf{\Phi}^\top \mathbf{t}$ as $\alpha \to 0$.

(c) Given an interpretation of the result in part b.

partial grade	1	2	3	4	5	6
(a)	3	4	3	2	3	3
(b)	2	1	1	3	1	2
(c)		2	2	1	1	2

Final grade is: (sum of partial grades) / 4.0 + 1.0