

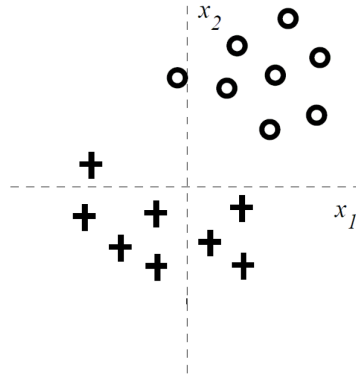
## Exam Advanced Machine Learning

24 October 2018, 12.00–14.45

This exam consists of 5 problems, each consisting of several questions. All answers should be motivated, including calculations, formulas used, etc.

### Question 1: Logistic regression

Consider the following training data for solving the classification problem illustrated in the following figure. Notice that the training data can be separated with *zero* training error with a linear separator.



We attempt to solve the binary classification task with a simple linear logistic regression model

$$\mathbb{P}(y = 1 \mid \mathbf{x}, \mathbf{w}) = g(w_0 + w_1 x_1 + w_2 x_2) = \frac{1}{1 + e^{-w_0 - w_1 x_1 - w_2 x_2}}.$$

Consider training the *regularized* linear logistic regression model, where only *one* of the parameters is regularized. In other words, the regularization penalty is equal to  $\lambda w_j^2$  for only one  $j \in \{0, 1, 2\}$  in contrast to the normal penalty  $\lambda \sum_{j=0}^2 w_j^2$ . How does the training error change when the regularization strength  $\lambda$  increases. State whether the training error increases, stays the same, or decreases as  $\lambda$  grows larger. Provide a justification for your answer for:

- (a) Regularization of  $w_2$
- (b) Regularization of  $w_1$
- (c) Regularization of  $w_0$

### Question 2: Decision trees

Master Yoda is concerned about the number of Jedi apprentices that have turned to the Dark Side, so he has decided to train a decision tree on some historical data to help identify problem cases in the future. The following table summarizes whether or not each of the 12 initiates turned to the Dark Side based on their age when their Jedi training began, whether or not they completed their training, their general disposition, and their species.

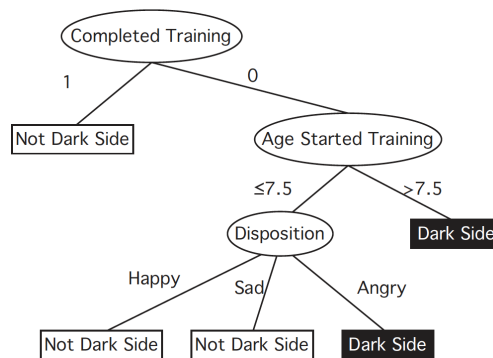
| Dark Side | Age Started Training | Completed Training | Disposition | Species      |
|-----------|----------------------|--------------------|-------------|--------------|
| 0         | 5                    | 1                  | Happy       | Human        |
| 0         | 9                    | 1                  | Happy       | Gungan       |
| 0         | 6                    | 0                  | Happy       | Wookiee      |
| 0         | 6                    | 1                  | Sad         | Mon Calamari |
| 0         | 7                    | 0                  | Sad         | Human        |
| 0         | 8                    | 1                  | Angry       | Human        |
| 0         | 5                    | 1                  | Angry       | Ewok         |
| 1         | 9                    | 0                  | Happy       | Ewok         |
| 1         | 8                    | 0                  | Sad         | Human        |
| 1         | 8                    | 0                  | Sad         | Human        |
| 1         | 6                    | 0                  | Angry       | Wookiee      |
| 1         | 7                    | 0                  | Angry       | Mon Calamari |

Recall that the impurity measured by the classification deviance is given by

$$D_{\text{node}} = -2 \sum_{k=1}^m n_k \log\left(\frac{n_k}{n}\right) = -2 \left[ \sum_{k=1}^m n_k \log(n_k) - n \log(n) \right],$$

with the number of occurrences  $n_1, \dots, n_m$  for the  $m$  possible outcomes ( $n = n_1 + \dots + n_m$ ).

- What is the initial impurity of *Dark Side*? You do not have to evaluate the logarithms.
- What is the impurity when Yoda choses to split on *Complete Training*? Again, you do not have to evaluate the logarithms.
- Suppose that the decision tree after training has the following structure.



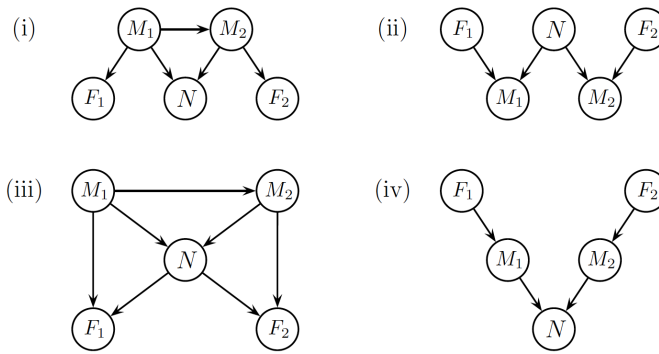
Consider the possibility that the input data above is noisy and not completely accurate, so that the decision tree you learned may not accurately reflect the function you want to learn. If you were to evaluate the three initiates represented by the data points below, on which one would you be most confident of your prediction, and why?

| Name   | Age Started Training | Completed Training | Disposition | Species      |
|--------|----------------------|--------------------|-------------|--------------|
| Ardath | 5                    | 0                  | Angry       | Human        |
| Barbar | 8                    | 0                  | Angry       | Gungan       |
| Caldar | 8                    | 0                  | Happy       | Mon Calamari |

### Question 3: Graphical models

Two astronomers in two different parts of the world make measurements  $M_1$  and  $M_2$  of the number of stars  $N$  in some small regions of the sky using their telescopes. Normally, there is a small possibility of error by up to one star in each direction. Each telescope can be, with a much smaller probability, badly out of focus (events  $F_1$  and  $F_2$ ). In such a case, the scientist will undercount by three or more stars or, if  $N$  is less than three fail to detect any stars at all.

- (a) Consider the four graphical models shown below. Which of them correctly, but not necessarily efficiently, represents the above information? Note that there may be multiple answers.



- (b) Write the expression for the joint probability  $\mathbb{P}(M_1, M_2, N, F_1, F_2)$  of *each* network in its *reduced* factored form.
- (c) Check if  $M_1$  is conditionally independent of  $M_2$  given  $N$  in each network.

### Question 4: Neural networks

Consider a single sigmoid threshold unit with three inputs  $x_1$ ,  $x_2$ , and  $x_3$ .

$$y = g(w_0 + w_1x_1 + w_2x_2 + w_3x_3) \quad \text{where} \quad g(z) = \frac{1}{1 + e^{-z}}.$$

We input values of either 0 or 1 for each of these inputs.

- (a) Assign values to the weights  $w_0$ ,  $w_1$ ,  $w_2$ , and  $w_3$  so that the output of the sigmoid unit is greater than 0.5 if and only if  $(x_1 \text{ AND } x_2) \text{ OR } x_3 = 1$ .
- (b) Describe how the face verification problem can be solved by using a convolutional neural network (CNN). Please mention in your answer, the required input, the architecture of the CNN, the cost function, and the output.
- (c) During the lecture, we have discussed one-to-many, many-to-one, and many-to-many recurrent neural networks (RNNs). Please give an application of each type of RNN.

**Question 5: Kaggle competition**

These questions pertain to the Kaggle competition on flight delays.

- (a) Which model(s) did you use in the competition? Elaborate on the choices on the model(s) and the corresponding (hyper)parameter(s).
- (b) Which features (default and/or created) did you use in your model and why?
- (c) Did overfitting play a role in the process of developing your solution? If so, how did it play a role and what did you do about it? If not, please explain why.
- (d) How good do you think the performance of your model is? Please explain.

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

Final grade is: (sum of partial grades) / 10 + 1.0