Exercise 2

Elements of Machine Learning, 2020, by Jens Petersen

The assignments in EML must be completed individually and written individually in English. Group discussions are allowed and encouraged, but in such cases you should list the group members in your handin. Your handin should include your solution to the exercises in a single pdf (not zipped), including code in a separate zip-file. Code used to solving the Combining multiple learners part should be attached as an appendix.

Exercises

From the textbooks (40 points)

Solve the following exercises

- PRML[1] 8.16, 8.27, ESL[2] 8.4, 10.1

Graphical Models (30 points)

A standard example in graphical model literature is the student network (Figure 1.

1. What is $p(\text{Intelligence} = 1 | \text{Letter} = 1, \text{SAT} = 1)$?
2. Convert the directed graph in Figure 1 to a moral graph
3. With the Letter node as root and the factor graph shown in Figure 2, Appendix A contains Sum-Product messages in the forward and backward passes and Appendix B contains Max-Sum messages in the forward pass. Give the missing messages (indicated by a question mark). Note you do not need to complete the messages relating to the Max-Sum backward pass. What are the marginal probabilities and most likely configuration of each node?

Combining multiple learners (30 points)

The goal of this exercise is to predict whether an email is spam or not using various ways of combining multiple learners. For this you will need the spam dataset described in ESL, which can be downloaded from `https://archive.ics.uci.edu/ml/datasets/spambase`. We will use classification trees as the base learner.

[1] Pattern Recognition and Machine Learning, Christopher M. Bishop

[2] Elements of Statistical Learning, Trevor Hastie, Robert Tibshirani, and Jerome Friedman

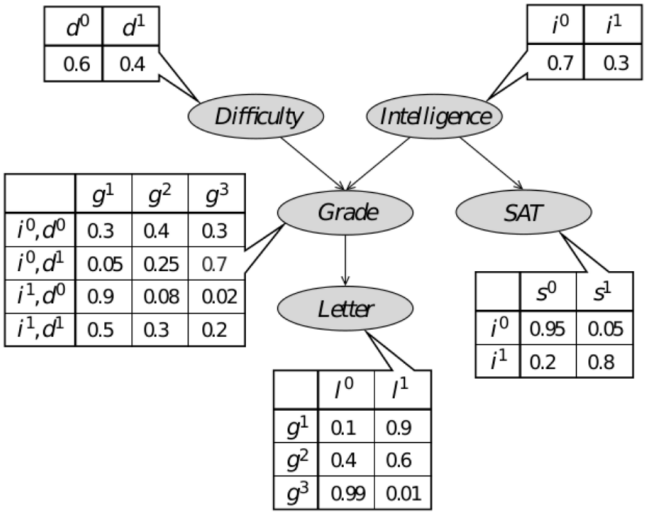

Figure 1: The student network. The model describes the relationship between class difficulty (0 = low, 1 = high), student intelligence (0 = not intelligent, 1 = intelligent), the grade the student gets in class (1 = good, 2 = average, 3 = bad), the student's score in the SAT exam (0 = low, 1 = high), and the recommendation letter the student gets from the professor (0 = not a good letter, 1 = good letter).

1. Split the dataset into non-overlapping training, validation, and test sets. Motivate your choice of splitting ratios, given the exercises below (6 points).

2. Implement, train and compare results of classifying spam (24 points)

   (a) using a basic classification tree method. You may use the decision tree implementation in scikit-learn for this `https://scikit-learn.org/stable/modules/tree.html`.

   (b) using bagging.

   (c) using a boosting algorithm, such as AdaBoost.M1.

   You are expected to describe, motivate, and compare possible methodological choices, including how you arrive at values for hyper parameters such as the number of iterations of AdaBoost, the number of base learners and the tree depth for the base learners. The bagging and boosting part of the exercise should be implemented by yourself, that is, it should not rely on existing implementations such as scikit-learn's ensemble methods `https://scikit-learn.org/stable/modules/ensemble.html`, however, you may use the base learner from 2a. Write up a comparison of the three approaches, what are your conclusions?

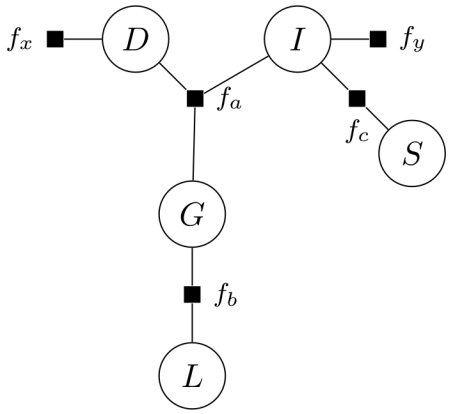

Figure 2: Factor graph corresponding to the student network with variables abbreviated by their first letters, where $f_x(D) = p(D)$, $f_y = p(I)$, $f_a(D, I, G) = p(G|D, I)$, $f_b(G, L) = p(L|G)$, and $f_c(I, S) = p(S|I)$

Appendix A - sum product messages

Using the factor graph, we can compute the forward pass of the sum-product algorithm using the node L as root:

1. $\mu_{f_x \to D}(D) = f_x(D)$
2. $\mu_{f_y \to I}(I) = f_y(I)$
3. $\mu_{S \to f_c}(S) = 1$
4. $\mu_{f_c \to I}(I) = \sum_S f_c(S, I)$
5. $\mu_{D \to f_a}(D) = \mu_{f_x \to D}(D)$
6. $\mu_{I \to f_a}(I) = \mathbf{?}$
7. $\mu_{f_a \to G}(G) = \mathbf{?}$
8. $\mu_{G \to f_b}(G) = \mu_{f_a \to G}(G)$
9. $\mu_{f_b \to L}(L) = \sum_G f_b(L, G)\mu_{G \to f_b}(G)$

The backward step, starts from the root and propagates back to the leaves:

1. $\mu_{L \to f_b}(L) = 1$
2. $\mu_{f_b \to f_G}(G) = \sum_L f_b(L, G)\mu_{L \to f_b}(L)$
3. $\mu_{G \to f_a}(G) = \mu_{f_b \to G}(G)$
4. $\mu_{f_a \to D}(D) = \mathbf{?}$

5. $\mu_{D \to f_x}(D) = \mathbf{?}$
6. $\mu_{f_a \to I}(I) = \sum_{G,D} f_a(I, G, D) \mu_{D \to f_a}(D) \mu_{G \to f_a}(G)$
7. $\mu_{I \to f_y}(I) = \mu_{f_a \to I}(I)$
8. $\mu_{I \to f_c}(I) = \mu_{f_a \to I}(I)$
9. $\mu_{f_c \to S}(S) = \sum_I f_c(I, S) \mu_{I \to f_c}(I)$

Sum product forward pass messages in detail

$$\mu_{f_x \to D}(D) = \begin{array}{c|c} D & p(D) \\ \hline 0 & 0.6 \\ 1 & 0.4 \end{array} \tag{1}$$

$$\mu_{f_y \to I}(I) = \begin{array}{c|c} I & p(I) \\ \hline 0 & 0.7 \\ 1 & 0.3 \end{array} \tag{2}$$

$$\mu_{S \to f_c}(S) = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \tag{3}$$

$$\mu_{f_c \to I}(I) = \sum_S f_c(S, I) = \sum_S \begin{array}{cc|c} I & S & f_c(S,I) = p(S|I) \\ \hline 0 & 0 & 0.95 \\ 0 & 1 & 0.05 \\ 1 & 0 & 0.2 \\ 1 & 1 & 0.8 \end{array} = \begin{array}{c|c} I & \mu_{f_c \to I}(I) \\ \hline 0 & 1 \\ 1 & 1 \end{array} \tag{4}$$

$$\mu_{D \to f_a}(D) = \mu_{f_x \to D}(D) = \begin{array}{c|c} D & p(D) \\ \hline 0 & 0.6 \\ 1 & 0.4 \end{array} \tag{5}$$

$$\mu_{I \to f_a}(I) = \textbf{?} \tag{6}$$

$$\mu_{f_a \to G}(G) = \textbf{?} \tag{7}$$

$$\mu_{G \to f_b}(G) = \mu_{f_a \to G}(G) = \begin{array}{c|c} G & \mu_{f_a \to G}(G) \\ \hline 1 & 0.362 \\ 2 & 0.2884 \\ 3 & 0.3496 \end{array} \tag{8}$$

$$\mu_{f_b \to L}(L) = \sum_G f_b(L, G)\mu_{G \to f_b}(G) = \sum_G \begin{array}{cc|c} G & L & f_b(L,G) \\ \hline 1 & 0 & 0.1 \\ 2 & 0 & 0.4 \\ 3 & 0 & 0.99 \\ 1 & 1 & 0.9 \\ 2 & 1 & 0.6 \\ 3 & 1 & 0.01 \end{array} * \begin{array}{c|c} G & \mu_{f_a \to G}(G) \\ \hline 1 & 0.362 \\ 2 & 0.2884 \\ 3 & 0.3496 \end{array}$$

$$= \begin{array}{cc|c} G & L & f_b(L,G) * \mu_{f_a \to G}(G) \\ \hline 1 & 0 & 0.1 * 0.362 = 0.0362 \\ 2 & 0 & 0.4 * 0.2884 = 0.11536 \\ 3 & 0 & 0.99 * 0.3496 = 0.346104 \\ 1 & 1 & 0.9 * 0.362 = 0.3258 \\ 2 & 1 & 0.6 * 0.2884 = 0.17304 \\ 3 & 1 & 0.01 * 0.3496 = 0.003496 \end{array} = \begin{array}{c|c} L & \mu_{f_b \to L}(L) \\ \hline 0 & 0.497664 \\ 1 & 0.502336 \end{array} \tag{9}$$

Sum product backward pass messages in detail

$$\mu_{L \to f_b}(L) = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \tag{1}$$

$$\mu_{f_b \to f_G}(G) = \sum_L f_b(L, G) \mu_{L \to f_b}(L) = \sum_L$$

G	L	$f_b(L,G)$
1	0	0.1
2	0	0.4
3	0	0.99
1	1	0.9
2	1	0.6
3	1	0.01

$=$

G	$\mu_{f_b \to G}(G)$
1	1
2	1
3	1

(2)

$$\mu_{G \to f_a}(G) = \mu_{f_b \to G}(G) =$$

G	$\mu_{f_b \to G}(G)$
1	1
2	1
3	1

(3)

$$\mu_{f_a \to D}(D) = \textbf{?} \tag{4}$$

$$\mu_{D \to f_x}(D) = \textbf{?} \tag{5}$$

$$\mu_{f_a \to I}(I) = \sum_{G,D} f_a(I,G,D)\mu_{D \to f_a}(D)\mu_{G \to f_a}(G)$$

$$= \sum_{G,D}
\begin{array}{cccc}
I & D & G & f_a(G,I,D) \\ \hline
0 & 0 & 1 & 0.3 \\
0 & 0 & 2 & 0.4 \\
0 & 0 & 3 & 0.3 \\
0 & 1 & 1 & 0.05 \\
0 & 1 & 2 & 0.25 \\
0 & 1 & 3 & 0.7 \\
1 & 0 & 1 & 0.9 \\
1 & 0 & 2 & 0.08 \\
1 & 0 & 3 & 0.02 \\
1 & 1 & 1 & 0.5 \\
1 & 1 & 2 & 0.3 \\
1 & 1 & 3 & 0.2
\end{array}
*
\begin{array}{ccc}
G & D & \mu_{D \to f_a}(D) * \mu_{G \to f_a}(G) \\ \hline
1 & 0 & 0.6 \\
2 & 0 & 0.6 \\
3 & 0 & 0.6 \\
1 & 1 & 0.4 \\
2 & 1 & 0.4 \\
3 & 1 & 0.4
\end{array}$$

(6)

$$= \sum_{G,D}
\begin{array}{ccc}
I & D & G & f_a(G,I,D) * (\mu_{D \to f_a}(D) * \mu_{G \to f_a}(G)) \\ \hline
0 & 0 & 1 & 0.3 * 0.6 = 0.18 \\
0 & 0 & 2 & 0.4 * 0.6 = 0.24 \\
0 & 0 & 3 & 0.3 * 0.6 = 0.18 \\
0 & 1 & 1 & 0.05 * 0.4 = 0.02 \\
0 & 1 & 2 & 0.25 * 0.4 = 0.1 \\
0 & 1 & 3 & 0.7 * 0.4 = 0.28 \\
1 & 0 & 1 & 0.9 * 0.6 = 0.54 \\
1 & 0 & 2 & 0.08 * 0.6 = 0.048 \\
1 & 0 & 3 & 0.02 * 0.6 = 0.012 \\
1 & 1 & 1 & 0.5 * 0.4 = 0.2 \\
1 & 1 & 2 & 0.3 * 0.4 = 0.12 \\
1 & 1 & 3 & 0.2 * 0.4 = 0.08
\end{array}
=
\begin{array}{cc}
I & \mu_{f_a \to I}(I) \\ \hline
0 & 1 \\
1 & 1
\end{array}$$

$$\mu_{I \to f_c}(I) = \mu_{f_y \to I}(I)\mu_{f_a \to I}(I) =
\begin{array}{cc}
I & \mu_{f_a \to I}(I) \\ \hline
0 & 1 \\
1 & 1
\end{array}
*
\begin{array}{cc}
I & \mu_{f_y \to I}(I) \\ \hline
0 & 0.7 \\
1 & 0.3
\end{array}
=
\begin{array}{cc}
I & \mu_{I \to f_c}(I) \\ \hline
0 & 0.7 \\
1 & 0.3
\end{array}
\tag{7}$$

$$\mu_{f_c \to S}(S) = \sum_I f_c(I,S)\mu_{I \to f_c}(I) = \sum_I f_c(I,S)
\begin{array}{cc}
I & \mu_{I \to f_c}(I) \\ \hline
0 & 0.7 \\
1 & 0.3
\end{array}$$

$$= \sum_I
\begin{array}{ccc}
I & S & \mu_{I \to f_c}(I) \\ \hline
0 & 0 & 0.95 \\
1 & 0 & 0.2 \\
0 & 1 & 0.05 \\
1 & 1 & 0.8
\end{array}
*
\begin{array}{cc}
I & \mu_{I \to f_c}(I) \\ \hline
0 & 0.7 \\
1 & 0.3
\end{array}
=
\begin{array}{cc}
S & \mu_{f_a \to S}(S) \\ \hline
0 & 0.725 \\
1 & 0.275
\end{array}
\tag{8}$$

Appendix B - max sum forward messages

The forward and backward messages for the max-sum algorithm are similar to the sum-product one (Appendix A), but instead of marginalizing we take the maximum, and instead of multiplying we take the sum:

1. $\mu_{f_x \to D}(D) = \log(f_x(D))$
2. $\mu_{f_y \to I}(I) = \log(f_y(I))$
3. $\mu_{S \to f_c}(S) = 0$
4. $\mu_{f_c \to I}(I) = \max_S \log(f_c(S, I))$
5. $\mu_{D \to f_a}(D) = \mu_{f_x \to D}(D)$
6. $\mu_{I \to f_a}(I) = \mathbf{?}$
7. $\mu_{f_a \to G}(G) = \mathbf{?}$
8. $\mu_{G \to f_b}(G) = \mu_{f_a \to G}(G)$
9. $\mu_{f_b \to L}(L) = \max_G \log(f_b(L, G)) + \mu_{G \to f_b}(G)$

Max sum forward pass messages in detail (please excuse the use of equalities with rounded off values)

$$\mu_{f_x \to D}(D) = \log(f_x(D)) = \begin{array}{c|c} D & \mu_{f_x \to D}(D) \\ \hline 0 & \log(0.6) \\ 1 & \log(0.4) \end{array} \tag{1}$$

$$\mu_{f_y \to I}(I) = \log(f_y(I)) = \begin{array}{c|c} I & \mu_{f_y \to I}(I) \\ \hline 0 & \log(0.7) \\ 1 & \log(0.3) \end{array} \tag{2}$$

$$\mu_{S \to f_c}(S) = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \tag{3}$$

$$\mu_{f_c \to I}(I) = \max_S \log(f_c(S, I)) = \max_S \begin{array}{cc|c} I & S & f_c(S,I) = p(S|I) \\ \hline 0 & 0 & 0.95 \\ 0 & 1 & 0.05 \\ 1 & 0 & 0.2 \\ 1 & 1 & 0.8 \end{array} = \begin{array}{c|c|c} I & \mu_{f_c \to I}(I) & \arg\max_S \\ \hline 0 & \log(0.95) & 0 \\ 1 & \log(0.8) & 1 \end{array}$$

$$= \begin{array}{c|c|c} I & \mu_{f_c \to I}(I) & \arg\max_S \\ \hline 0 & -0.051 & 0 \\ 1 & -0.223 & 1 \end{array} \tag{4}$$

$$\mu_{D \to f_a}(D) = \mu_{f_x \to D}(D) = \begin{array}{c|c} D & \mu_{f_x \to D}(D) \\ \hline 0 & \log(0.6) \\ 1 & \log(0.4) \end{array} \tag{5}$$

$$\mu_{I \to f_a}(I) = \textbf{?} \tag{6}$$

$$\mu_{f_a \to G}(G) = \textbf{?} \tag{7}$$

$$\mu_{G \to f_b}(G) = \mu_{f_a \to G}(G) = \begin{array}{c|c} G & \mu_{f_a \to G}(G) \\ \hline 1 & \log(0.1296) \\ 2 & \log(0.1596) \\ 3 & \log(0.1862) \end{array} = \begin{array}{c|c} G & \mu_{f_a \to G}(G) \\ \hline 1 & -2.043 \\ 2 & -1.835 \\ 3 & -1.681 \end{array} \tag{8}$$

$$\mu_{f_b \to L}(L) = \max_G(\log(f_b(L,G)) + \mu_{G \to f_b}(G))$$

$$= \max_G \left(\begin{array}{cc|c} G & L & \log((f_b(L,G)) \\ \hline 1 & 0 & \log(0.1) \\ 2 & 0 & \log(0.4) \\ 3 & 0 & \log(0.99) \\ 1 & 1 & \log(0.9) \\ 2 & 1 & \log(0.6) \\ 3 & 1 & \log(0.01) \end{array} + \begin{array}{c|c} G & \mu_{G \to f_b}(G) \\ \hline 1 & \log(0.1296) \\ 2 & \log(0.1596) \\ 3 & \log(0.1862) \end{array} \right)$$

$$= \max_G \begin{array}{cc|c} G & L & \log((f_b(L,G)) + \mu_{G \to f_b}(G) \\ \hline 1 & 0 & \log(0.1 * 0.1296) = \log(0.013) \\ 2 & 0 & \log(0.4 * 0.1596) = \log(0.0638) \\ 3 & 0 & \log(0.99 * 0.1862) = \log(0.184338) \\ 1 & 1 & \log(0.9 * 0.1296) = \log(0.11664) \\ 2 & 1 & \log(0.6 * 0.1596) = \log(0.09576) \\ 3 & 1 & \log(0.01 * 0.1862) = \log(0.001862) \end{array} \tag{9}$$

$$= \begin{array}{c|c|c} L & \mu_{f_b \to L}(L) & \arg\max_G \\ \hline 0 & \log(0.184338) & 3 \\ 1 & \log(0.11664) & 1 \end{array} = \begin{array}{c|c|c} L & \mu_{f_b \to L}(L) & \arg\max_G \\ \hline 0 & -1,691 & 3 \\ 1 & -2,148 & 1 \end{array}$$