Solutions to the exam in Neural Networks and Learning Systems - TBMI26 Exam 2019-03-18

Part 1

1. The following methods are supervised learning methods:

   - Back-propagation
   - k-NN
   - LDA (Linear discriminant Analysis)
   - SVN (Support Vector Machines)

2. Slack-variables are used to allow som samples to be miss-classified, in order not to overfit to noisy data and outliers.

3. Which of these functions can be used in the hidden layers of a back-prop network?

   - $y = s$ - No, the activation function in a hidden layer needs to be non-linear to be useful
   - $y = \tanh(s)$ - Yes
   - $y = \frac{s}{\|s\|}$ - No, this function is not differentiable
   - $y = e^{(-s^2)}$ - Yes

4. The first eigenvalue of the data covariance matrix describes the maximum variance of the data.

5. The multi-armed bandit illustrates the exploration-exploitation problem.

6. The general definition of a kernel function is $k(\mathbf{x}, \mathbf{y}) = \phi(\mathbf{x})^T \phi(\mathbf{y})$, where $\phi((\mathbf{x})$ is a non-linear function.

7. "k-means clustering" and "mixture of gaussians" are two examples of the EM-algorithm.

8. A random forrest.

9. The "empirical risk" is the number of wrong classifications.

10. The purpose of the hidden layers in a multi-layer perceptron classifier is transform the data to a space where the problem i linearly separable.

Part 2

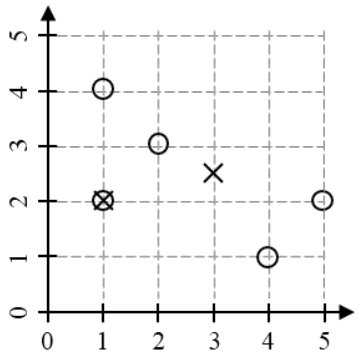
11. The prototype vectors are updated to the means of the closest data points, i.e.,

$$\mathbf{p}_1 = \frac{1}{4}\left[\begin{pmatrix}1\\4\end{pmatrix} + \begin{pmatrix}2\\3\end{pmatrix} + \begin{pmatrix}4\\1\end{pmatrix} + \begin{pmatrix}5\\2\end{pmatrix}\right] = \begin{pmatrix}3\\2.5\end{pmatrix}$$

and

$$\mathbf{p}_2 = \begin{pmatrix}1\\2\end{pmatrix}$$

as shown in the figure.



The algorithm must perform 2 more iteration before no more points change clusters.

12. Besides the training data, we use *validation data* for monitoring the generalization error during training and, finally, *test data* for testing the final performance after training.

13. Ordinary PCA can be fomulated as the following eigenvalue problem:

$$\mathbf{X}\mathbf{X}^T\mathbf{e} = \lambda\mathbf{e}$$

Multiplying from the left with the transpose of the data matrix $\mathbf{X}$ gives

$$\mathbf{X}^T\mathbf{X}\mathbf{X}^T\mathbf{e} = \lambda\mathbf{X}^T\mathbf{e}$$

and, if we let $\mathbf{f} = \mathbf{X}^T\mathbf{e}$, this can be written as

$$\mathbf{X}^T\mathbf{X}\mathbf{f} = \lambda\mathbf{f}$$

which is a new eigenvalue problem of the inner product matrix $\mathbf{X}^T\mathbf{X}$, i.e. a kernel matrix.

14. The ReLU function is linear for positive values and zero for negative values. The advantage is that the gradient is constant and does not decrease when propagated back thru the network. This means that it avoids the so-called "vanishing gradient problem".

15. One possible implementation of the decision tree:

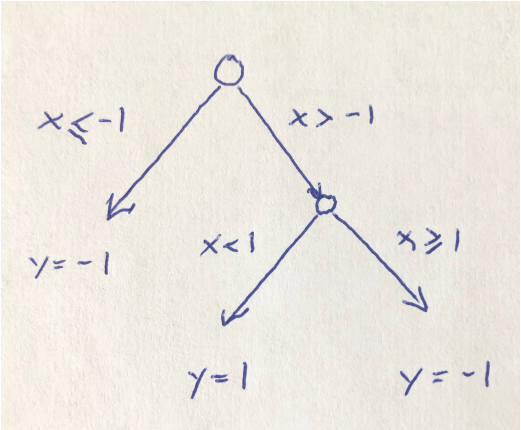


Figure 1: Decision tree

Part 3

16. a) The drawn net should have bias weights and have all parameters defined.

b) In order to calculate the error gradient, let us first line up the whole chain of functions involved:

$$\varepsilon = \frac{1}{N}\sum_{i=1}^{2}\sum_{\mu=1}^{N}\varepsilon_i^\mu \tag{1}$$

$$= \frac{1}{N}\sum_{i=1}^{2}\sum_{\mu=1}^{N}(d_i^\mu - y_i^\mu)^2 \tag{2}$$

$$y_i^\mu = \sum_{j=0}^{2} V_{ij}u_j^\mu \tag{3}$$

$$u_j^\mu = \sigma(s_j^\mu) = \frac{1}{1+e^{-s}} \tag{4}$$

$$s_j^\mu = \sum_{k=0}^{2} W_{jk}x_k^\mu \tag{5}$$

Now, we want to calculate each of the partial derivatives $\frac{\partial\varepsilon}{\partial V_{ij}}$ and $\frac{\partial\varepsilon}{\partial W_{jk}}$.
The chain rule gives us for V (the weights in the output layer):

$$\frac{\partial\varepsilon}{\partial V_{ij}} = \frac{1}{N}\sum_{\mu=1}^{N}\sum_{n=1}^{2}\frac{\partial\varepsilon_n^\mu}{\partial y_n^\mu}\frac{\partial y_n^\mu}{\partial V_{nj}} = \frac{2}{N}\sum_{\mu=1}^{N}(d_i^\mu - y_i^\mu)(-1)u_j^\mu = \tag{6}$$

$$= \frac{2}{N}\sum_{\mu=1}^{N}(y_i^\mu - d_i^\mu)\,u_j^\mu \tag{7}$$

Note that the sum over the number of output neurons has changed index name from i to n, so that we don't confuse it with the index in the differentiation variable V_{ij}. However, at the second equal sign this sum vanishes because the only term contributing is when $n = i$.
The chain rule gives us for W (the weights in the hidden layer).

$$\frac{\partial\varepsilon}{\partial W_{jk}} = \frac{1}{N}\sum_{\mu=1}^{N}\sum_{i=1}^{2}\frac{\partial\varepsilon_i^\mu}{\partial y_i^\mu}\frac{\partial y_i^\mu}{\partial u_j^\mu}\frac{\partial u_j^\mu}{\partial s_j^\mu}\frac{\partial s_j^\mu}{\partial W_{jk}^\mu} = \tag{8}$$

$$= \frac{1}{N}\sum_{\mu=1}^{N}\sum_{i=1}^{2}2(d_i^\mu - y_i^\mu)(-1)\left(V_{ij}\sigma'(s_j^\mu)x_k^\mu\right) = \tag{9}$$

$$= \frac{2}{N}\sum_{\mu=1}^{N}x_k^\mu\sum_{i=1}^{2}(y_i^\mu - d_i^\mu)V_{ij}\frac{e^{-s_j^\mu}}{(1+e^{-s_j^\mu})} \tag{10}$$

Note that we don't need sums for k and j since W_{jk} j^{th} and k^{th} term of x_k, V_{ij}, and s_j

17. a)

g_{00}	=	**2**	2	3	2	(0.5p)
g_{01}	=	**0**	1	0	1	(0.25p)
g_{02}	=	**0**	2	0	2	(0.25p)
g_{10}	=	**4**	4	2	1	(0.5p)
g_{11}	=	**4**	0	2	0	(0.25p)
g_{12}	=	**2**	0	1	0	(0.25p)

b)

$$d_0 = g_{00} + g_{10} = f_0 * h_{00} + f_1 * h_{10} \overset{h_{00}=h_{h10}}{=} (f_0 + f_1) * h_{00}$$

$$\begin{pmatrix} d_1(x) \\ d_2(x) \end{pmatrix} = \begin{pmatrix} g_{01}(x) + g_{11}(x) \\ g_{02}(x) + g_{12}(x) \end{pmatrix} = \begin{pmatrix} f_0(x) \cdot 1 + f_1(x) \cdot 2 \\ f_0(x) \cdot 2 + f_1(x) \cdot 1 \end{pmatrix} = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix} \begin{pmatrix} f_0(x) \\ f_1(x) \end{pmatrix}$$

c) $I \cdot J \cdot K$ parameters have to be learned.

18. The direction that optimality separates the two classes is given by

$$\mathbf{W} = \mathbf{C}_{tot}^{-1} (\mathbf{m}_1 - \mathbf{m}_2)$$

where $\mathbf{C}_{tot}$ is the sum of the individual covariance matrices for the two respective classes, and $\mathbf{m}_1$ and $\mathbf{m}_2$ are the centers of the two classes. We begin with the "crosses" class:

$$\mathbf{X}_1 = \begin{pmatrix} -1 & 1 & 2 & 3 & 5 \\ 2 & 2 & 4 & 3 & 4 \end{pmatrix}$$

$$\mathbf{C}_1 = \frac{1}{N-1} (\mathbf{X}_1 - \mathbf{m}_1) (\mathbf{X}_1 - \mathbf{m}_1)^T = \left[\mathbf{m}_1 = \begin{pmatrix} 2 \\ 3 \end{pmatrix} \right] = \frac{1}{4} \begin{pmatrix} 20 & 7 \\ 7 & 4 \end{pmatrix}$$

Now for the "circles" class:

$$\mathbf{X}_2 = \begin{pmatrix} 0 & 2 & 3 & 4 & 6 \\ -1 & 0 & 0 & 1 & 0 \end{pmatrix}$$

$$\mathbf{C}_2 = \frac{1}{N-1} (\mathbf{X}_2 - \mathbf{m}_2) (\mathbf{X}_2 - \mathbf{m}_2)^T = \left[\mathbf{m}_1 = \begin{pmatrix} 3 \\ 0 \end{pmatrix} \right] = \frac{1}{4} \begin{pmatrix} 20 & 4 \\ 4 & 2 \end{pmatrix}$$

We now get:

$$\mathbf{C}_{tot} = \mathbf{C}_1 + \mathbf{C}_2 = \frac{1}{4} \begin{pmatrix} 40 & 11 \\ 11 & 6 \end{pmatrix}$$

$$\mathbf{C}_{tot}^{-1} = \frac{4 * 4}{40 * 6 - 11 * 11} * \frac{1}{4} \begin{pmatrix} 6 & -11 \\ -11 & 40 \end{pmatrix} \approx \begin{pmatrix} 0.20 & -0.37 \\ -0.37 & 1.34 \end{pmatrix}$$

$$\mathbf{W} = \mathbf{C}_{tot}^{-1} (\mathbf{m}_1 - \mathbf{m}_2) \approx \begin{pmatrix} -1.31 \\ 4.40 \end{pmatrix}$$

Normalize $\mathbf{W}$:

$$\hat{\mathbf{W}} = \frac{\mathbf{W}}{||\mathbf{W}||_2} \approx \frac{1}{4.59} \begin{pmatrix} -1.31 \\ 4.40 \end{pmatrix} \approx \begin{pmatrix} -0.29 \\ 0.96 \end{pmatrix}$$

Project the data on $\hat{\mathbf{W}}$ which gives the following:

$$\mathbf{Y}_1 = \hat{\mathbf{W}}^T \mathbf{X}_1 \approx \begin{pmatrix} 2.20 & 1.63 & 3.26 & 2.02 & 2.41 \end{pmatrix}$$

$$\mathbf{Y}_2 = \hat{\mathbf{W}}^T \mathbf{X}_2 \approx \begin{pmatrix} -0.96 & -0.57 & -0.86 & -0.18 & -1.71 \end{pmatrix}$$

The projected data is shown in the figure below.

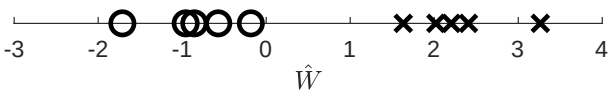


Figure 2: Data projected on $\hat{\mathbf{W}}$

19. A definition of an optimal (deterministic) Q-function:

$$\begin{aligned} Q^*(x,a) &= r(x,a) + \gamma V^*(f(x,a)) \\ &= r(x,a) + \gamma Q^*(f(x,a), \mu^*(f(x,a))) \\ &= r(x,a) + \gamma \max_b Q^*(f(x,a),b) \end{aligned}$$

and for the V-function:

$$V^*(x) = \max_b Q^*(x,b)$$

We can find a solution by going through the state models recursively.

System A:

$$\begin{aligned} V^*(5) &= 0 \\ V^*(4) = Q^*(4 \to 5) = 4 + \gamma V^*(5) &= 4 \\ V^*(3) = Q^*(3 \to 5) = 1 + \gamma V^*(5) &= 1 \end{aligned}$$

For state 2 we have two options:

$$\begin{aligned} Q^*(2 \to 3) = 1 + \gamma V^*(3) &= 1 + \gamma \\ Q^*(2 \to 4) = 0 + \gamma V^*(4) &= 4\gamma \end{aligned}$$

Since the optimal path always will be followed:

$$\begin{aligned} V^*(2) = \max(1+\gamma, 4\gamma) &= \begin{cases} 1+\gamma & \text{if } \gamma < \frac{1}{3} \\ 4\gamma & \text{if } \gamma \geq \frac{1}{3} \end{cases} \\ V^*(1) = Q^*(1 \to 2) = 0 + \gamma V^*(2) &= \begin{cases} \gamma + \gamma^2 & \text{if } \gamma < \frac{1}{3} \\ 4\gamma^2 & \text{if } \gamma \geq \frac{1}{3} \end{cases} \end{aligned}$$

System B:

According to the state model we have :

$$\begin{aligned} V^*(2) = Q^*(2, \to 3) &= 0 + \gamma V^*(3) = \gamma V^*(3) \\ V^*(1) = Q^*(1, \to 2) &= 3 + \gamma V^*(2) = 3 + \gamma^2 V^*(3) \end{aligned}$$

In state 3 we have two options:

$$\begin{aligned} Q^*(3, \to 1) = 1 + \gamma V^*(1) &= 1 + 3\gamma + \gamma^3 V^*(3) \\ Q^*(3, \to 2) = 3 + \gamma V^*(2) &= 3 + \gamma^2 V^*(3) \end{aligned}$$

Assume that $Q^*(3 \to 1)$ is larger than $Q^*(3 \to 2)$. Then we get

$$V^*(3) = Q^*(3 \to 1) = 1 + 3\gamma + \gamma^3 V^*(3) \to V^*(3) = \frac{1+3\gamma}{1-\gamma^3}$$

If we instead assume that $Q^*(3 \to 2)$ is larger than $Q^*(3 \to 1)$ we get

$$V^*(3) = Q^*(3 \to 2) = 3 + \gamma^2 V^*(3) \to V^*(3) = \frac{3}{1-\gamma^2}$$

The first assumption is true when:

$$\begin{aligned}
\frac{1+3\gamma}{1-\gamma^3} &> \frac{3}{1-\gamma^2} \rightarrow \\
(1+3\gamma)(1-\gamma^2) &> 3(1-\gamma^3) \rightarrow \\
1+3\gamma-\gamma^2-3\gamma^3 &> 3-3\gamma^3 \rightarrow \\
\gamma^2-3\gamma+2 &< 0 \rightarrow \\
\left(\gamma-\frac{3}{2}\right)^2 &< \frac{1}{4} \rightarrow \\
\frac{-1}{2} < \gamma &- \frac{3}{2} < \frac{1}{2} \rightarrow \\
1 < &\gamma < 2
\end{aligned}$$

But $0 < \gamma < 1$, so this condition never holds. Therefore, the first assumption is never valid and $Q^*(3 \rightarrow 2)$ is always larger than $Q^*(3 \rightarrow 1)$ for our range of γ. The optimal path is therefore $3 \rightarrow 2$, and we get:

$$\begin{aligned}
Q^*(3 \rightarrow 1) &= 1 + \frac{3\gamma}{1-\gamma^2} \\
V^*(3) = Q^*(3 \rightarrow 2) &= \frac{3}{1-\gamma^2} \\
V^*(2) &= \frac{3\gamma}{1-\gamma^2} \\
V^*(1) &= \frac{3}{1-\gamma^2}
\end{aligned}$$