STAT3007/7007 Deep Learning, Tutorial 4

2022 Semester 2

1. (Perceptron)

(a) Consider training a perceptron on the training set

$$\begin{aligned}&(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), (\mathbf{x}_3, y_3), (\mathbf{x}_4, y_4), (\mathbf{x}_5, y_5), (\mathbf{x}_6, y_6)\\ =\ &((-3, 1), -1), ((-2, 1), -1), ((-1, 1), -1), ((0.1, 1), 1), ((0.2, 1), 1), ((0.3, 1), 1).\end{aligned}$$

i. Let us choose $\mathbf{w}_0 = 0$. Thus initially the perceptron classifies all examples as negative. Now use the perceptron algorithm with $\eta = 1$ to update the weights. How many updates do you need and what is your final weight vector? Show your working.

ii. Consider the perceptron convergence theorem. If we choose $\mathbf{w}^* = (1, 0)$. What is the margin γ? What is the radius R? What is the upper bound on the number of updates required for the perceptron algorithm to converge?

(b) Consider a dataset $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n) \in \mathbf{R}^{d+1} \times \{-1, +1\}$. In lecture, we defined the margin of a separating hyperplane as the minimum distance of an example to it. For a separating hyperplane $\mathbf{w}^\top \mathbf{x} = 0$, its margin is $\gamma = \min_i \frac{y_i \mathbf{x}_i^\top \mathbf{w}}{\|\mathbf{w}\|}$.
Note that each $\mathbf{x}_i$ is the original d-dimensional feature vector augmented with a dummy feature of value 1. We can also put the examples and the hyperplane $\mathbf{w}^\top \mathbf{x} = 0$ in the original d-dimensional feature space, and then compute the margin of the hyperplane in this d-dimensional feature space. Is the margin computed in the original feature space the same as the one computed in the augmented feature space?

(c) Consider a dataset $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n) \in \mathbf{R}^{d+1} \times \{-1, +1\}$.

i. Assume that $\mathbf{w}$ and $\mathbf{w}'$ are the normal vectors of two separating hyperplanes for the dataset. Show that for any $\lambda \in [0, 1]$, $\lambda\mathbf{w} + (1 - \lambda)\mathbf{w}'$ is also the normal vector of a separating hyperplane.

ii. Assume that $\mathbf{w}$ and $\mathbf{w}'$ are the normal vectors of two separating hyperplanes for the dataset. Is it true that for any $\lambda \in \mathbf{R}$, $\lambda\mathbf{w} + (1-\lambda)\mathbf{w}'$ is also the normal vector of a separating hyperplane? If not, find all λ values such that $\lambda\mathbf{w} + (1 - \lambda)\mathbf{w}'$ is also the normal vector of a separating hyperplane?

2. (Adaline) We explained in lecture why Adaline can help to reduce the loss by using the first-order Taylor series expansion of the loss. Give a similar justification for the logistic approximation discussed in the lecture.

3. (Hopfield nets) Consider a Hopfield net for storing two patterns $\mathbf{a}_1 = (-1, -1, 1, 1)$ and $\mathbf{a}_2 = (1, 1, -1, -1)$.

(a) How many neurons are there in the Hopfield net? Draw the structure of the network.

(b) What is the weight matrix for the network?

(c) Assume that we are given a noisy input pattern $(-1, 1, 1, 1)$. To retrieve the stored pattern, we need to first initialise the activation states of the neurons with their corresponding values in the given pattern. Perform one iteration of synchronous updates on all the neurons. What is the updated pattern?

(d) For the input pattern $(-1, 1, 1, 1)$, how many iterations of synchronous updates do you need until the network converges? What is the final pattern?

(e) Now consider the input pattern $(1, 1, 1, 1)$. Update the neurons in the order 1, 2, 3, 4, 1, 2, 3, 4,... until convergence. What is the final pattern?

(f) Consider the input pattern $(1, 1, 1, 1)$. Update the neurons in the order 4, 3, 2, 1, 4, 3, 2, 1,... until convergence. What is the final pattern? Do you get the same pattern as (f)?