CSDS 440: Assignment 6

Shaochen (Henry) ZHONG, `sxz517`

Mingyang TIE, `mxt497`

Due and submitted on 10/16/2020

Fall 2020, Dr. Ray

Problem 23

(i, ii, iii)

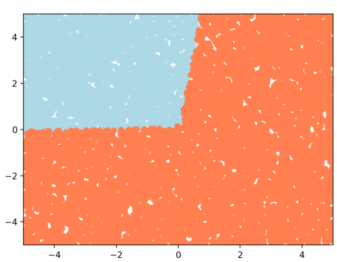

Figure 1: $w \in [-10, 10]$

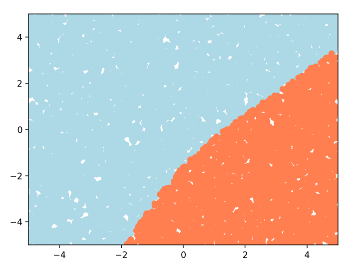

Figure 2: $w \in [-3, 3]$

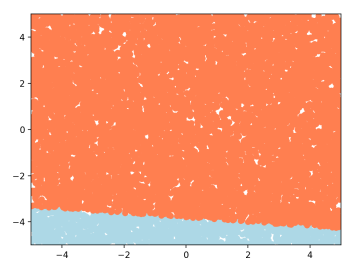

Figure 3: $w \in [-0.1, 0.1]$

By observation we may tell with the weights w decaying, the output graph becomes more "linear" which is better againt overfitting. For color reference, we have Output = 1 and Output = 0.

Problem 24

For traditional backpropagation of n_k in k-th layer, we have $\frac{\partial L}{\partial n_k} = \sum_{k+1 \in \text{Downstream}(n_k)} \frac{\partial L}{\partial n_{k+1}} \cdot \frac{\partial n_{k+1}}{\partial n_k}$ assuming there is an edge from n_k to n_{k+1}. For the non-feedforward structure the question suggests, n_k will still calculate the above loss, but it will also consider other nodes in the k-th layer which it connects to n_k. We denotes these nodes as $n_{k_i} = \{n_{k_1}, n_{k_2}, n_{k_3}, ...\}$, where edges like $n_k \to n_{k_i}$ exist.

Since there is no cycle, these n_{k_i} nodes will not connect to any Downstream(n_k) nodes. So we can just do backpropagation layer by layer and once the above $\sum\limits_{k+1 \in \text{Downstream}(n_k)} \frac{\partial L}{\partial n_{k+1}} \cdot \frac{\partial n_{k+1}}{\partial n_k}$ is calculated, we will calculate the loss of n_{k_i} nodes with respect to n_k and add to the loss of n_k. Which will give us:

$$
\begin{aligned}
\frac{\partial L}{\partial n_k} &= \sum_{k+1 \in \text{Downstream}(n_k)} \frac{\partial L}{\partial n_{k+1}} \cdot \frac{\partial n_{k+1}}{\partial n_k} + \sum_{n_{k_i}} \frac{\partial L}{\partial n_{k_i}} \cdot \frac{\partial n_{k_i}}{\partial n_k} \\
\implies \frac{\partial L}{\partial w_{(k-1)k}} &= \frac{\partial L}{\partial n_k} \frac{\partial n_j}{w_{(k-1)k}} \\
&= \frac{\partial L}{\partial n_k} \cdot x_{(k-1)k}
\end{aligned}
$$

Assume there is a $n_{k-1} \to n_k$ edge.

Problem 25

$$
\begin{aligned}
P(T = i \mid c_1, \ldots, c_N) &= \frac{P(c_1, \ldots, c_N \mid T = i) \cdot P(T = i)}{\sum\limits_{i \in \{1,2,3\}} P(c_1, \ldots, c_N \mid T = i) \cdot P(T = i)} \\
&= \frac{\prod\limits_{j=1}^{N} P(c_j \mid T = i) \cdot P(T = i)}{\sum\limits_{i \in \{1,2,3\}} \big(\prod\limits_{j=1}^{N} P(c_j \mid T = i) \cdot P(T = i) \big)}
\end{aligned}
$$

Since no prior information is exposed, we know $P(T = i) = \frac{1}{3}$ for $i \in \{1, 2, 3\}$. Thus we have:

$$
P(T = i \mid c_1, \ldots c_N) = \frac{\prod\limits_{j=1}^{N} P(c_j \mid T = i) \cdot \frac{1}{3}}{\sum\limits_{i \in \{1,2,3\}} \big(\prod\limits_{j=1}^{N} P(c_j \mid T = i) \cdot \frac{1}{3} \big)}
$$

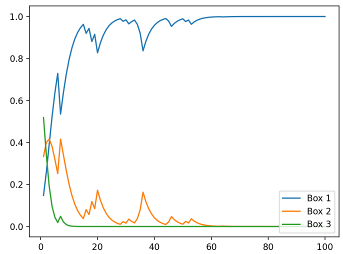

Figure 4: Box 1

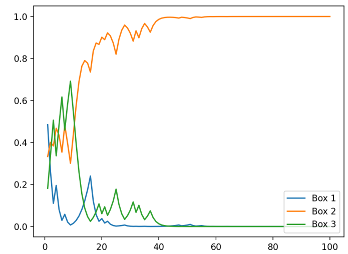

Figure 5: Box 2

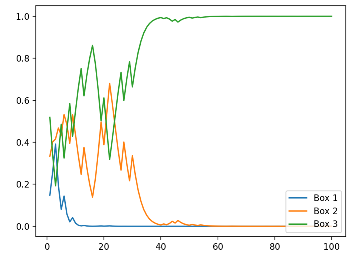

Figure 6: Box 3

In comparision to *Problem 27*. In *Problem 25* we have all boxes having probablity of $\frac{1}{3}$ instead of $\{0.1, 0.1, 0.8\}$ in *Problem 27*. Reflecting the setting of the questions.

Also it is observable that as more candies were picked, which box we are picking from becomes clear; reflecting the change of decision boundary. Also, boxes with better similiarity (in terms of candy distribution) with the box we picking from will maintain a higher probablity for a longer period time comparing to boxes lack of similiarity to the box we picking from. e.g. When we are picking from Box 1, Box 2 has a higher probablity than Box 3 for a longer period of time before it goes to 0. This is because Box 1 and Box 2 have similar distribution, while Box 3 has the almost opposite distribution to Box 1.

Problem 26

Similar to pervious question, we have:

$$\begin{aligned}
P(c_{N+1} = C \mid c_1, \ldots, c_N) &= \frac{\sum_{i\in\{1,2,3\}} P(c_1, ..., c_N, c_{N+1} = C \mid T = i)}{P(c_1, \ldots, c_N)} \\
&= \frac{\sum_{i\in\{1,2,3\}} P(c_1, ..., c_N \mid T = i) P(c_{N+1} = C \mid T = i)}{P(c_1, \ldots, c_N)} \\
&= \sum_{i\in\{1,2,3\}} P(T = i \mid c_1, ..., c_N) P(c_{N+1} = C \mid T = i)
\end{aligned}$$

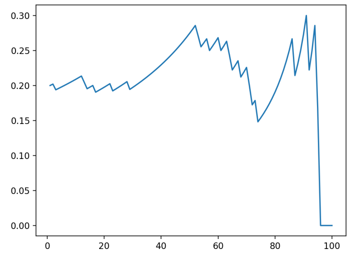

Figure 7: Box 1

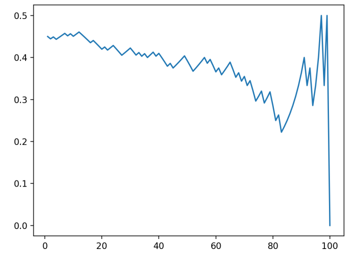

Figure 8: Box 2

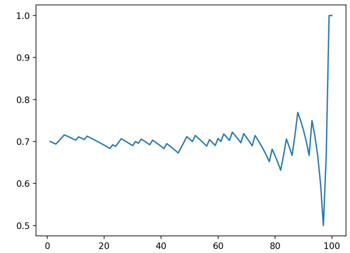

Figure 9: Box 3

First it is observable that boxes with more C inside has a higher probablity, and every graph starts with the probablity that representing their distribution, reflecting the setting of the problem.

As N is growing, the probablity changes more and more suddenly. This is because candies picked at a later time will have more impact on the probablity – e.g. the $N+1$ round will either make the probablity be 1 or 0 depending on if there is any C left after the N-th pick.

Problem 27

Base on *Problem 25*, we now have $P(T=1,2)=0.1$ and $P(T=3)=0.8$. Now substitute this finding in to the following equation.

$$P(T=i \mid c_1,\dots,c_N) = \frac{\prod_{j=1}^{N} P(c_j \mid T=i)\cdot P(T=i)}{\sum_{i\in\{1,2,3\}} \big(\prod_{j=1}^{N} P(c_j \mid T=i)\cdot P(T=i)\big)}$$

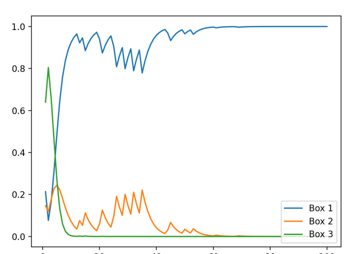

Figure 10: Box 1

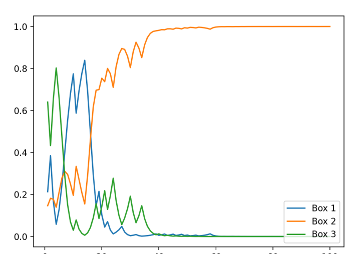

Figure 11: Box 2

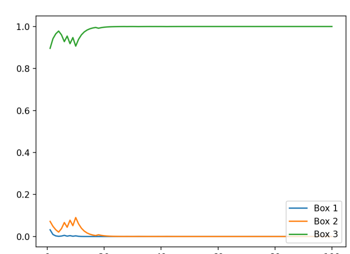

Figure 12: Box 3

In comparision to *Problem 25*. In *Problem 27* we have each box having an initial probablity of

$\{0.1, 0.1, 0.8\}$ instead of $\frac{1}{3}$ in *Problem 25*. Reflecting the setting of the questions.

Also it is observable that as more candies were picked, which box we are picking from becomes clear; reflecting the change of decision boundary. Also, boxes with better similiarity (in terms of candy distribution) with the box we picking from will maintain a higher probablity for a longer period time comparing to boxes lack of similiarity to the box we picking from. e.g. When we are picking from Box 1, Box 2 has a higher probablity than Box 3 for a longer period of time before it goes to 0. This is because Box 1 and Box 2 have similar distribution, while Box 3 has the almost opposite distribution to Box 1.