DDA3020: Homework 3

November 14, 2022

Homework due: **11:59pm, November 28, 2022**.

1 Written Problems (8 points)

1. Given the following loss function, please plot the computational graph, and derive the update procedure of parameters using back-propagation algorithm,

$$\mathcal{L}(\Theta) = \mathrm{CE}\Big(\mathbf{y}, \mathrm{softmax}\Big(\mathbf{W}_2 \tanh(\mathbf{W}_1 \boldsymbol{x} + \mathbf{b}_1) + \mathbf{b}_2 + \mathbf{W}_4 \max(0, \mathbf{W}_3 \boldsymbol{x} + \mathbf{b}_3) + \mathbf{b}_4\Big)\Big) + \lambda \sum_{i=1}^{4} \|\mathbf{W}_i\|_F^2, \tag{1}$$

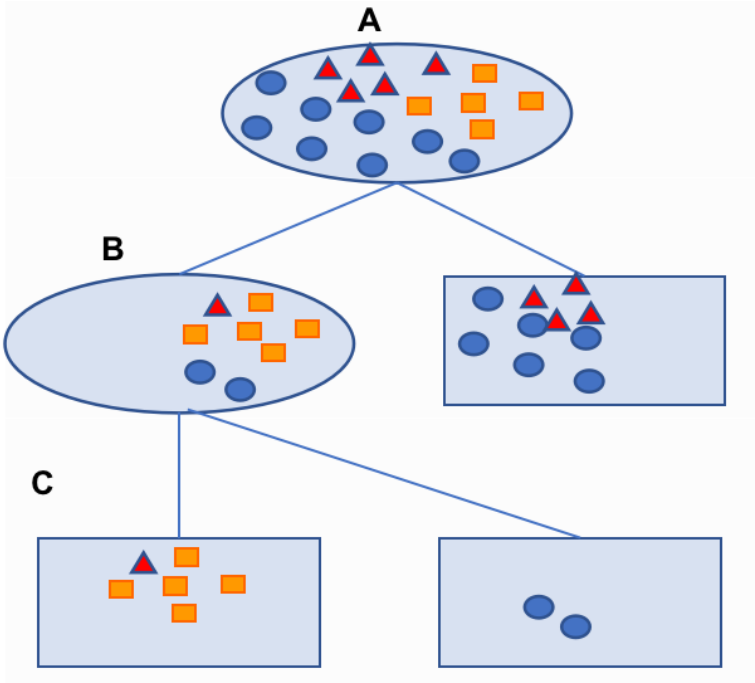
   where $\Theta = \{\mathbf{W_1}, \mathbf{W_2}, \mathbf{W_3}, \mathbf{W_4}, \mathbf{b_1}, \mathbf{b_2}, \mathbf{b_3}, \mathbf{b_4}\}$ denote the parameters and $\mathbf{W}_i \in \mathbb{R}^{d \times d}$, $\mathbf{b}_i \in \mathbb{R}^{d \times 1}$, $i = 1, \ldots, 4$. $\boldsymbol{x} \in \mathbb{R}^d$ indicates the input features; $\boldsymbol{y} \in \mathbb{R}^d$ is the ground-truth label; CE denotes cross-entropy (2 points).

2. (1 point) The input shape is $100 \times 100 \times 3$. Consider the following 4 layers, $Conv_1 + Maxpool_1 + Conv_2 + Maxpool_2$.

   - $Conv_1$: 8 $5 \times 5 \times 3$ filters, stride=3, padding=p_1
   - $Maxpool_1$: 2×2 filter, stride=2, padding=p_2
   - $Conv_2$: 16 $3 \times 3 \times 8$ filters, stride=2, padding=p_3
   - $Maxpool_2$: 3×3 filter, stride=2, padding=p_4

   (1) Find the minimum p_1, p_2, p_3, p_4 that make the filters fit the input of each layer. (2) Based on the computed p_1, p_2, p_3, p_4, please compute the shape of activation map of each layer and the total number of parameters of these layers.

3. Compute the Gini index, the entropy and the classification error for each node of the tree in the figure below, where different color (also shape) represents different class. (1 point)



4. (2 points) Suppose we randomly sample a training set D from some unknown distribution. For each training set D we sample, we train a regression model h_D to predict y from x (one dimensional). We repeat this process 10 times resulting in 10 trained models.
Recall that $y = t(x) + \epsilon$, where $\epsilon \in \mathcal{N}(0, \sigma^2)$. Here, we specify $\sigma^2 = 1$.
For a new test sample $(x, y) = (3, 7)$ sampled from the same distribution that generated the training sets, we suppose $t(x = 3) = 7.2$, and ϵ is instantiated as -0.2.
Suppose the predictions of the new test sample based on the 10 trained models are 6, 8, 9, 5, 9, 5, 4, 8, 9, 4.
(a) Based on this 10 trials, please compute the empirical mean squared error (MSE), Bias^2 and Variance on this test sample.(1 point)
(Hint:

   - For a new test sample (x, y), we define its mean squared error (MSE) by different models as
   $$MSE(x, y) = E_D[(h_{D_i}(x) - y)^2].$$
   - It can observed that $E_{(x,y),D}[(h_{D_i}(x) - y)^2] = E_{(x,y)}[MSE(x, y)]$.
   - The empirical estimation of MSE based on above 10 trained models is
   $$\widehat{MSE}(x, y) = \frac{1}{10}\sum_{i=1}^{10}(h_D(x) - y)^2.$$

(b) Explain why $\widehat{MSE}(x, y) \neq \text{Bias}^2 + \text{Variance} + \sigma^2$ (1 point)

5. (2 points) Neural networks with different activation functions.
Consider a two-layer network function of the form

$$y_k(\mathbf{x}, \mathbf{w}) = \sigma\left(\sum_{j=1}^{M} w_{kj}^{(2)} h\left(\sum_{i=1}^{D} w_{ji}^{(1)} x_i + w_{j0}^{(1)}\right) + w_{k0}^{(2)}\right),$$

in which the hidden unit nonlinear activation function $h(\cdot)$ is given by logistic sigmoid function of the form

$$\sigma(a) = \frac{1}{1+\exp(-a)}.$$

Show that there exists an equivalent network, which computes exactly the same function, but with the hidden unit activation function given by tanh(a) where the tanh function is defined by

$$\tanh(a) = \frac{e^a - e^{-a}}{e^a + e^{-a}}.$$

That is, if there's another two-layer network function with tanh(a) as hidden unit activation function:

$$\hat{y}_k(\mathbf{x}, \hat{\mathbf{w}}) = \sigma\left(\sum_{j=1}^{M} \hat{w}_{kj}^{(2)} \tanh\left(\sum_{i=1}^{D} \hat{w}_{ji}^{(1)} x_i + \hat{w}_{j0}^{(1)}\right) + \hat{w}_{k0}^{(2)}\right),$$

then there exists linear transformation between these w and $\hat{w}$, that enable $y_k(x, w) = \hat{y}_k(x, \hat{w})$ for all x.
Hint: first find the relation between $\sigma(a)$ and $\tanh(a)$, and then show that the parameters of the two networks differ by linear transformations.

2 Programming (8 points)

1. Decision Tree (required) (4 points)

   Task description Fit (*i.e.*, regression) the real variable *Sales* in the **Carseats** dataset, using decision tree, bagging, and random forests. All these algorithms can be implemented by calling **sklearn** in Python. The loss is set as sum of squared error (SSE).

   Dataset **Carseats** contains 400 data points (saved in 400 rows). For each data, the first column is the value of target variable *Sales* that we want to fit; the remaining 9 columns indicate 9 features (or attributes), as shown in Fig. 1. There is no fixed train/test splitting. In this project, you have two options: simply set the first 300 rows as the training set, and the remaining 100 rows as the testing set; randomly split the whole dataset to 300 train + 100 test, and try multiple times. The dataset can be obtained from `https://github.com/selva86/datasets/blob/master/Carseats.csv`

   What you should do

   - **Data statistics**: analyze the statistics of the target variable and each feature, and try to visualize the statistics (*e.g.*, histogram) (0.5 point)
   - **Decision tree**: solve the above problem using decision tree method; report the train/test errors with respect to different maximum depths, different least node sizes; plot the learned tree (2 points)
   - **Bagging of trees**: solve the above problem using the bagging method, with decision tree as the base learner; report the train/test errors with respect to different depths, different number of trees (2 points)
   - **Random forests**: solve the above problem using the random forest method, with decision tree as the base learner; report the train/test errors with respect to different number of trees, different values of m (the number of candidate attributes to split in every step, see Slides 'W7-Decision Tree', Page 68) (2 points)
   - Plot the curve of bias^2 with respect to different number of trees in random forests, *e.g.*, $\#tree = 10, 20, \ldots, 100$. Then, describe the relationship between bias^2 and different number of trees; repeat the procedure for variance. (1.5 points)

Carseats

Sales	CompPrice	Income	Advertising	Population	Price	ShelveLoc	Age	Education	Urban	US
9.5	138	73	11	276	120	Bad	42	17	Yes	Yes
11.22	111	48	16	260	83	Good	65	10	Yes	Yes
10.06	113	35	10	269	80	Medium	59	12	Yes	Yes
7.4	117	100	4	466	97	Medium	55	14	Yes	Yes
4.15	141	64	3	340	128	Bad	38	13	Yes	No
10.81	124	113	13	501	72	Bad	78	16	No	Yes
6.63	115	105	0	45	108	Medium	71	15	Yes	No
11.85	136	81	15	425	120	Good	67	10	Yes	Yes
6.54	132	110	0	108	124	Medium	76	10	No	No
4.69	132	113	0	131	124	Medium	76	17	No	Yes
9.01	121	78	9	150	100	Bad	26	10	No	Yes
11.96	117	94	4	503	94	Good	50	13	Yes	Yes
3.98	122	35	2	393	136	Medium	62	18	Yes	No
10.96	115	28	11	29	86	Good	53	18	Yes	Yes
11.17	107	117	11	148	118	Good	52	18	Yes	Yes
8.71	149	95	5	400	144	Medium	76	18	No	No
7.58	118	32	0	284	110	Good	63	13	Yes	No
12.29	147	74	13	251	131	Good	52	10	Yes	Yes
13.91	110	110	0	408	68	Good	46	17	No	Yes
8.73	129	76	16	58	121	Medium	69	12	Yes	Yes
6.41	125	90	2	367	131	Medium	35	18	Yes	Yes
12.13	134	29	12	239	109	Good	62	18	No	Yes
5.08	128	46	6	497	138	Medium	42	13	Yes	No
5.87	121	31	0	292	109	Medium	79	10	Yes	No
10.14	145	119	16	294	113	Bad	42	12	Yes	Yes

Figure 1: Some examples of **Carseats**.

2. Handwritten Digit Recognition using sk-learn (4 points)

   Task description Use the function "MLPClassifier" of sk-learn to construct a fully connected network to classify the MNIST data. The dataset can be downloaded from `http://yann.lecun.com/exdb/mnist/`. The training-testing spilliting has been given. Show the performance of your neural network with different structures:

   - number of hidden layers chosen from $\{1, 2, 3\}$
   - number of hidden nodes chosen from $\{50, 200, 784\}$ (let all hidden layers have the same number of nodes)

Note that you should submit A3_StudentID.pdf (report, together with the written answers), and A3_StudentID.ipynb (code). Please zip them into "A3_StudentID.zip". The reference report is in Assignment 1. You can check it on BlackBoard.
(You can submit several files in one submission. Don't submit them in different submissions.) Your report for the programming questions should include necessary formulas, charts, and explanations. The number of pages should be 4-5.