THE HONG KONG UNIVERSITY OF SCIENCE & TECHNOLOGY
Machine Learning
Homework 2

Due Date: See course website

Your answers should be typed, not handwritten. You can submit a Word file or a pdf file. Submissions are to be made via Canvas. Note that penalty applies if your similarity score exceeds 40. To minimize your similarity score, don't copy the questions.

Question 1 Consider the following dataset:

Instance	y	x_1	x_2
1	1	0	0
2	1	0	0
3	1	0	1
4	1	0	1
5	0	1	0
6	0	1	0
7	1	1	1
8	0	1	1

(a) Give the Naïve Bayes model for the data. There is no need to use Laplace smoothing, and there is no need to show the process of calculation.

(b) Calculate the posterior probabilities of the Instances 1 and 7 belonging to the two classes according to the model of the previous sub-question. Show the process of calculation.

Question 2 [Optional]
Suppose there are K i.i.d training sets $S_k = \{\mathbf{x}_{ki}, y_{ki}\}_{i=1}^{m}$ $(k = 1, \ldots, K)$ for a regression problem with a hypothesis class $\mathcal{H}$. For each k, let

$$h_k = \arg\min_{h \in \mathcal{H}} \hat{\epsilon}(h), \text{ where } \hat{\epsilon}(h) = \frac{1}{m}\sum_{i=1}^{m}(y_{ki} - h(\mathbf{x}_{ki}))^2$$

The variance component of the expected error of h_k is:

$$\text{Var}(h_k) = E_{\mathbf{x}} E_{S_k}[E_{S_k}(h_k(\mathbf{x})) - h_k(\mathbf{x})]^2.$$

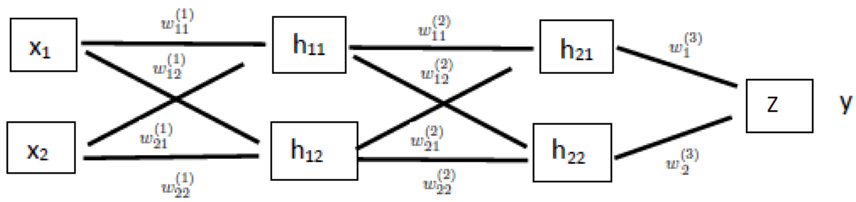
Because the training sets are i.i.d, $\text{Var}(h_k)$ is the same for different k. Let $\text{Var}(h_k) = \sigma^2$.

(a) Let $\bar{h} = \frac{1}{K}\sum_{k=1}^{K} h_k$. Show that the variance component of the expected error of $\bar{h}$ is:

$$\text{Var}(\bar{h}) = \frac{1}{K}\sigma^2.$$

(b) Based on part (a), a variance reduction technique called **bagging** is proposed. Find out how bagging works, and explain why it reduces variance.

Question 3 Consider the following feedforward neural network with one input layer, two hidden layers, and one output layer. The hidden neurons are **tanh** units, while the output neuron is a sigmoid unit.



The weights of the network and their initial values are as follows:

Between input and first hidden: $\begin{bmatrix} w_{11}^{(1)} & w_{12}^{(1)} \\ w_{21}^{(1)} & w_{22}^{(1)} \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$

Between two hidden layers: $\begin{bmatrix} w_{11}^{(2)} & w_{12}^{(2)} \\ w_{21}^{(2)} & w_{22}^{(2)} \end{bmatrix} = \begin{bmatrix} -1 & -1 \\ 1 & 1 \end{bmatrix}$

Between second hidden and output: $\begin{bmatrix} w_1^{(3)} \\ w_1^{(3)} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

For simplicity, assume the units do not have bias parameters. Let there be only one training example $(x_1, x_2, y) = (1, 2, 0)$.

(a) Consider feeding $(x_1, x_2) = (1, 2)$ to the network. What are the outputs of the hidden units? What is the logit $z = u_{21}w_1^{(3)} + u_{22}w_2^{(3)}$ calculated at the output unit? The output of the output unit is a probability distribution $p(y|x_1 = 1, x_2 = 2, \theta)$. What is the distribution?

(b) Next consider backpropagation. The loss function for the training example is $L = -\log p(y = 0|x_1 = 1, x_2 = 2, \theta)$. What is the error $\frac{\partial L}{\partial z}$ for the output unit? What are the errors for the hidden units? What are $\frac{\partial L}{\partial w_{22}^{(2)}}$ and $\frac{\partial L}{\partial w_{22}^{(1)}}$? If we want to reduce the loss on the example, should we increase or decrease the two parameters?

Question 4: Why is the sigmoid activation function not recommended for hidden units, but it is fine for an output unit.

Question 5: What is dropout used for in deep learning? Why does it work? Answer briefly.

Question 6: What are the key ideas behind the Adam algorithm for training deep neural networks? Answer briefly.