Part-1: Math

1. Classification and Cross-entropy loss

x_n is an input data sample, and it is a vector.
y_n is the ground-truth class label of x_n.
$\hat{y}_n$ is the output "soft-label" of a logistic regression classifier given the input x_n
n is from 1 to N
the number of classes is K

(1) write down the formula of binary cross-entropy
$\hat{y}_n$ is a scalar

$$L = -1/N \sum y_n \log(\hat{y}_n) + (1+y_n) \log(1-\hat{y}_n))$$

(2) write down the formula of cross-entropy
$\hat{y}_n$ is a vector of K elements
Assume y_n is in the format of one-hot-encoding

$$L = -1/N \sum \sum y(n,k) \log(\hat{y}_{nk})$$

(3) Assume there are two classes: class-0, class-1, and x_n is in class-1 (ground-truth $y_n = 1$)
Assume the output is $\hat{y}_n = 0.8$ from a binary logistic regression classifier (linear + sigmoid)
Compute the binary cross-entropy loss associated with the data sample x_n

$$L = -1 \cdot \log(0.8) + 0 \cdot \log(0.8) = 0.223$$

(4) Assume there are three classes: class-0, class-1 and class-2, and x_n is in class-2 (ground-truth $y_n = 2$).
Assume the output is $\hat{y}_n = [0.1, 0.2, 0.7]^T$ from a multi-class logistic regression classifier (linear + softmax)
Convert y_n into the format of one-hot-encoding
Compute the cross-entropy loss associated with the data sample x_n

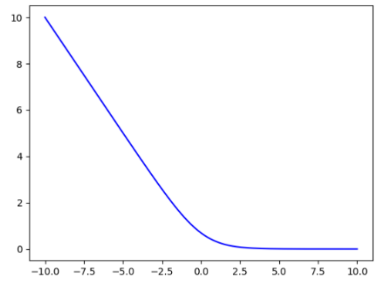
$$y_n \begin{bmatrix} y_0 \\ y_1 \\ y_2 \end{bmatrix}$$

$$\begin{array}{cccc} \text{label} & 0 & 1 & 2 \\ & \begin{vmatrix} 1 \\ 0 \\ 0 \end{vmatrix} & \begin{vmatrix} 0 \\ 1 \\ 0 \end{vmatrix} & \begin{vmatrix} 0 \\ 0 \\ 1 \end{vmatrix} \end{array}$$

$$L = -\frac{1}{n} \sum \sum y(n,k) \log(\hat{y}_{nk})$$

$$-\frac{1}{3} (1 \cdot \log(0.1) + 1 \cdot \log(0.2) + 1 \cdot \log(0.3)) = 1.423$$

(5) Show that the function $f(x) = -log\left(\frac{1}{1+e^{-x}}\right)$ is convex in x. log is natural log
Here is a plot of the function, and it seems that the function is convex.



$$\frac{d^2}{dx^2} = \frac{\partial^2}{\partial x^2} \left[-\ln \frac{1}{1+e^{-x}} \right]$$

$$\frac{df}{dx} = \left[-\frac{1}{e^x+1} \right.$$

$$\frac{e^x}{(e^x+1)^2} \geq 0 \quad \text{Convex}$$

Hint: show that $\frac{\partial^2 f}{\partial x^2} \geq 0$ then it is convex. This explains why cross entropy loss is convex.
The concept of cross entropy is from information theory

2. Regression

x_n is an input data sample, and it is a vector.
y_n is the ground-truth target value of x_n.
y_n could be a vector or a scalar
$\hat{y}_n$ is the predicted target value of a regressor (e.g. linear regressor) given the input x_n
n is from 1 to N

(1) write down the formula of MSE loss when y_n is a vector

MSE loss : $L = \frac{1}{N} \sum_{n=1}^{N} (\hat{y}_n - y_n)^2$

(2) write down the formula of MAE loss when y_n is a vector

MAE loss :

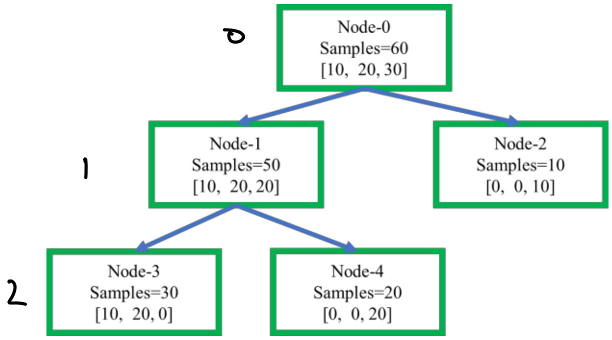
$$L = \frac{1}{N} \sum_{n=1}^{N} (\hat{y}_n - y_n)$$

3. Decision Tree

A decision tree is a partition of the input space.
Every leaf node of the tree corresponds to a region of the final partition of the input space.

(1) The output of a decision tree for regression looks like stairs. Why? — because the target value in the same region for the regression are similar
(2) Is it a good strategy to build a deep tree such that the number of samples in each leaf node is 1?
(3)~(6) are related to the tree below

↓ no because we will lose accuracy



60

(3) What is the total number of training samples according to the above tree ?
(4) What is the depth of the tree ? 2
(5) How many 'pure' nodes (entropy =0) are there in the above tree ? 2
(6) How many leaf/terminal nodes?

3

4. Bagging and Random forest

(1) Could it be useful to apply bagging with KNN classifiers using a fixed K? Why useful/useless?
no knn classifiers are to stable
(2) Random-forest uses a method to reduce the correlation between trees. What is the method?
random sampling of data and features

5. Boosting

What is the main difference between boosting (e,g. XGBoost) and bagging (e.g. Random-forest)?
Boosting = not Average   Bagging = RF average or majority

6. Stacking

model need to work together →(1) Could it be useful to stack many polynomial models of the same degree?
(2) Could it be useful to stack models of different types/structures?
(you may try this in the programming tasks)
in the end we train one model

7. Overfitting and Underfitting

It is easy to understand Overfitting and Underfitting but it is hard to detect them.
Consider two scenarios in a classification task:
[✓ (1)] the training accuracy is 100% and the testing accuracy is 50%   overfitting
(2) the training accuracy is 80% and the testing accuracy is 70%
In which scenario is overfitting likely present?

Consider two new scenarios in a classification task:
(1) the training accuracy is 80% and the testing accuracy is 70%
[✓ (2)] the training accuracy is 50% and the testing accuracy is 50%
In which scenario is underfitting likely present?   underfitting

Keep in mind that, in real applications, the numbers in different scenarios may be very similar.
We can always increase model complexity to avoid underfitting.
We need to find the model with the "right" complexity (i.e. the best hyper-parameters) to reduce overfitting if possible.

parameters that can't be directly lerned

8. Training, Validation, and Testing for Classification and Regression

—(1) What are hyper-parameters of a regression/classification model?

—(2) Why do we need a validation set? Why don't we just find the optimal hyper-parameters on the training set?

(3) Why don't we optimize hyper-parameters using the testing set?

validation set to update hyper parameters we retn cant on model

because we check traing and test on testing

it would be overfitting to know final exam

9. SVM

(1) Why maximizing the margin in the input space will improve classifier robustness against noise?

maximizing margin will give use the best decion boundaries against noise

(2) Will the margin in the input space be maximized by a nonlinear SVM?

we can use a linear model to represent non linear decision boundary.

10. Information theory (Graduate Student only)

The PMF for a discrete random variable X is $[p_1, p_2, p_3, \dots p_K]$. Show that entropy is at its maximum when the PMF is a uniform distribution, i.e. $p_k = 1/K$

Hint: you can use Jensen's inequality

11. Information theory (bonus points)

For a pair of discrete random variables X and Y (scalars, not vectors) with the joint distribution $p(x, y)$, the joint entropy $H(X, Y)$ is defined as

$$H(X,Y) = -\sum_x \sum_y p(x,y) log(p(x,y))$$

which can also be expressed using mathematical expectation:

$$H(X,Y) = -E[log(p(x,y))]$$

The entropy of X (PDF is $p(x)$) is defined as

$$H(X) = -\sum_x p(x) log(p(x))$$

The entropy of Y (PDF is $p(y)$) is defined as

$$H(Y) = -\sum_y p(y) log(p(y))$$

note: $p(x)$ and $p(y)$ represent different PDFs.

The mutual information is defined as

$$I(X,Y) = \sum_x \sum_y p(x,y) log\left(\frac{p(x,y)}{p(x)p(y)}\right)$$

Jensen's inequality says: $E[f(X)] \geq f(E[X])$ where $f(x)$ is a convex function.

Prove that $H(X,Y) \leq H(X) + H(Y)$

Hint: show that $I(X,Y) = H(X) + H(Y) - H(X,Y)$, and then show $I(X,Y) \geq 0$ using Jensen's inequality.

Part-2: Programming on classification and regression

Read the instructions in HW2P2T1.ipynb, HW2P2T2.ipynb, HW2P2T3.ipynb

Grading: (points for each question/task)

	Undergraduate Student	Graduate Student
Question 1	10	10
Question 2	5	5
Question 3	5	5
Question 4	5	5

Question 5	3	3
Question 6	2	2
Question 7	5	5
Question 8	5	5
Question 9	5	5
Question 10	Bonus (5 points)	5
Question 11	Bonus (5 points)	Bonus (5 points)
HW2P2T1	30	25
HW2P2T2	25	25
HW2P2T3	Bonus (5 points)	Bonus (5 points)

following lemma:

10: $p(x)$ and $q(x)$ be pmf on interval I

in real numbers $p \geq 0$ $q > 0$ on I

where $- \int p \log p \, dx \leq - \int p \log q \, dx$

if both integrals exist and there is equality if and only if $p(x) = q(x)$ for all x

now let p be pmf of $\{x_1 \ldots x_n\}$

with $p_i = p(x)$ $q_i = 1/n$ for all i

$$- \sum_{i=1}^{n} p \log q_i = \sum_{i=1}^{n} p_i \log n = \log n$$

which is entropy of q therefore

our lemma says $h(p) \leq h(q)$ if and

only if p is uniform