Course: Machine Learning - Foundations
Week 1 (Graded assignment)

---

1. (1 point) [2, 4, -5] belongs to which of the following?

   A. $\mathbb{R}$

   B. $\mathbb{R}^+$

   C. Both $\mathbb{R}^+$ and $\mathbb{R}^-$

   D. $\mathbb{R}^3$

   Answer: D

   Solution:

   The vector $[2, 4, -5]$ contains 3 components and all of them are real numbers.
   So, $\begin{bmatrix} 2 \\ 4 \\ -5 \end{bmatrix} \in \mathbb{R}^3$.

   $\therefore$ Option D is correct.

2. (1 point) Which of the following may **not** be an appropriate choice of loss function for regression?

   A. $\frac{1}{n}\sum_{i=1}^{n}(f(x_i) - y_i)^2$

   B. $\frac{1}{n}\sum_{i=1}^{n}|f(x_i) - y_i|$

   C. $\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}(f(x_i) \neq y_i)$

   Answer: C

   Solution:

   Here, option C that is, $Loss = \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}(f(x_i) \neq y_i)$ may be a good choice for classification, but it is not a good choice for regression.

   You can see that this loss function will increase when the prediction is not equal to a label. However, it does this with a fixed loss of 1. Ideally, we would want the loss increase to be proportionate to the amount of discrepancy between the prediction and the label.

   $\therefore$ Option C is correct.

3. (1 point) Identify which of the following requires use of classification technique.

A. Predicting the amount of rainfall in May 2022 in North India based on precipitation data of the year 2021.

B. Predicting the price of a land based on its area and distance from the market.

C. Predicting whether an email is spam or not.

D. Predicting the number of Covid cases on a given day based on previous month data.

Answer: C

Solution:

Here, in options A, B and D, we can see that we have to predict some kind of real number. Namely, amount of rainfall, price of land and number of cases. These kinds of problems are more suitable to regression. Option C however, is predicting in which category the datapoint (email) falls into. It is an example of binary classification technique.

$\therefore$ Option C is correct.

4. (1 point) (Multiple Select) Mark all incorrect statements in the following

A. $\mathbf{1}(355\%2 = 1) = 1$

B. $\mathbf{1}(788\%2 = 1) = 0$

C. $\mathbf{1}(355\%2 = 0) = 1$

D. $\mathbf{1}(788\%2 = 0) = 1$

Answer: C

Solution:

Let's look at each option one by one.

A. Since 355 is odd, $355\%2 = 1$. So, the statement inside the indicator function is true. That is, $\mathbf{1}(355\%2 = 1) = 1$. Since this option is a true statement, it will not be marked.

B. Since 788 is even, $788\%2 = 0$. So, the statement inside the indicator function is false. That is, $\mathbf{1}(788\%2 = 1) = 0$. Since this option is a true statement, it will not be marked.

C. Since 355 is odd, $355\%2 = 1$. So, the statement inside the indicator function is false. That is, $\mathbf{1}(355\%2 = 0) = 0$. Since this option is a false statement, it will be marked.

D. Since 788 is even, $788\%2 = 0$. So, the statement inside the indicator function is true. That is, $\mathbf{1}(788\%2 = 0) = 1$. Since this option is a true statement, it will not be marked.

$\therefore$ Only option C is correct.

5. (1 point) Which of the following is false regarding supervised and unsupervised machine learning?
   A. Unsupervised machine learning helps you to find different kinds of unknown patterns in data.
   B. Regression and classification are two types of supervised machine learning techniques while clustering and density estimation are two types of unsupervised learning.
   C. In unsupervised learning model, the data contains both input and output variables while in supervised learning model, the data contains only input data.

Answer: C

Solution:

Here, option C is a false statement. It is infact supervised learning model in which, the data contains both input and output variables. Also, it is unsupervised learning model in which, data contains only input data.

$\therefore$ Option C is correct.

6. (1 point) The output of regression model is
   A. is discrete.
   B. is continuous and always within a finite range.
   C. is continuous with any range.
   D. may be discrete or continuous.

Answer: C

Solution:

The output of a regression model, linear regression for example, can be any real number. It is continuous and can be within any range.

$\therefore$ Option C is correct.

7. (1 point) **(Multiple select)** Which of the following is/are supervised learning task(s)?

   A. Making different groups of customers based on their purchase history.

   B. Predicting whether a loan client may default or not based on previous credit history.

   C. Grouping similar Wikipedia articles as per their content.

   D. Estimating the revenue of a company for a given year based on number of items sold.

Answer: B,D

Solution:

Let's take each option one by one.

A. Making different groups is an example of clustering, which is an unsupervised learning task.
B. Predicting whether a client may default or not is an example of binary classification, which is a supervised learning task.
C. Again, grouping similar articles is an example of clustering, which is an unsupervised learning task.
D. Estimation of revenue which is a real continuous number is an example of regression, a supervised learning task.

$\therefore$ Options B and D are correct.

8. (1 point) Which of the following is used for predicting a continuous target variable?

   A. Classification

   B. Regression

   C. Density Estimation

   D. Dimensionality Reduction

Answer: B

Solution:

Out of the options, the technique used for prediction of a continuous target variable is regression.

$\therefore$ Option B is correct.

9. (1 point) Consider the following: "The _____ is used to fit the model; the _____ is used for model selection; the _____ is used for computing the generalization error."

   Which of the following will fill the above blanks correctly?

   A. Test set; Validation set; training set

   B. Training set; Test set; Validation set

   C. Training set; Validation set; Test set

   D. Test set; Training set; Validation set

   Answer: C

   Solution:

   The training set is used to fit our model. After that, the validation set is used to select the best model. Then, the test set is used for computing the generalization error.

   $\therefore$ Option C is correct.

10. (1 point) Consider the following loss functions:

    1. $\frac{1}{n}\sum_{i=1}^{n} -log(P(X^i))$
    2. $\frac{1}{n}\sum_{i=1}^{n} ||g(f(X^i)) - X^i||^2$
    3. $\frac{1}{n}\sum_{i=1}^{n} (f(X^i) - Y^i)^2$
    4. $\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(f(X^i) \neq Y^i)$

    The above loss functions pertain to which of the following ML techniques (in that order)?

    A. Dimensionality Reduction, Regression, Classification, Density Estimation

    B. Dimensionality Reduction, Classification, Density Estimation, Regression

    C. Density Estimation, Dimensionality Reduction, Regression, Classification

    D. Classification, Density Estimation, Dimensionality Reduction, Regression

    E. Classification, Dimensionality Reduction, Regression, Density Estimation

    Answer: C

    Solution:

    Let's go over them one by one.

1. This is the negative log likelihood loss and is used for density estimation.
2. This is computing the error between the reconstructed datapoint and actual datapoint and is used in dimensionality reduction.
3. This is the squared error loss and it is used for regression.
4. This loss function simply compares if prediction and label are equal or not. This is used in classification.

$\therefore$ Option C is correct.

11. (1 point) Compute the loss when **Pair 1** and **Pair 2** (shown below) are used for dimensionality reduction for the data given in the following Table:

x_1	x_2
1	0.5
2	2.3
3	3.1
4	3.9

Consider the loss function to be $\frac{1}{n}\sum_{i=1}^{n}||g(f(x^i)) - x^i||^2$.

1. **Pair 1:** $f(x) = (x_1 - x_2)$, $g(u) = [u/2, u/2]$
2. **Pair 2:** $f(x) = (x_1 + x_2)/2$, $g(u) = [u/2, u/2]$

Here $f(x)$ is the encoder function and $g(x)$ is the decoder function.
Pair 1: _____
Pair 2: _____

Answer: Pair 1: 7.6 [Range could be 7.22 to 7.98]
Pair 2: 3.8 [Range could be 3.61 to 4]

Solution:

We are given an encoder (f) and a decoder (g) function. To solve this question, we will take each datapoint x^i and encode it using encoder function getting $f(x^i)$ and then decode it to get $g(f(x^i))$. Then the squared error would be given as $||g(f(x^i)) - x^i||^2$. We would then take the average of this error over all datapoints to get the loss.

Pair 1:

	x^i	$f(x^i)$	$g(f(x^i))$	$g(f(x^i)) - x^i$	$\|\|g(f(x^i)) - x^i\|\|^2$
x^1	$[1, 0.5]$	0.5	$[0.25, 0.25]$	$[-0.75, -0.25]$	0.625
x^2	$[2, 2.3]$	-0.3	$[-0.15, -0.15]$	$[-2.15, -2.45]$	10.625
x^3	$[3, 3.1]$	-0.1	$[-0.05, -0.05]$	$[-3.05, -3.15]$	19.2245
x^4	$[4, 3.9]$	0.1	$[0.05, 0.05]$	$[-3.95, -3.85]$	30.425

$$Loss = \frac{1}{4}\sum_{i=1}^{4} ||g\left(f\left(x^i\right)\right) - x^i||^2 = \frac{1}{4}\left(0.625 + 10.625 + 19.2245 + 30.425\right) = 15.224$$

Pair 2:

	x^i	$f\left(x^i\right)$	$g\left(f\left(x^i\right)\right)$	$g\left(f\left(x^i\right)\right) - x^i$	$\|\|g\left(f\left(x^i\right)\right) - x^i\|\|^2$
x^1	$[1, 0.5]$	0.75	$[0.375, 0.375]$	$[-0.625, -0.125]$	0.406
x^2	$[2, 2.3]$	2.15	$[1.075, 1.075]$	$[-0.925, -1.225]$	2.356
x^3	$[3, 3.1]$	3.05	$[1.525, 1.525]$	$[-1.475, -1.575]$	4.66
x^4	$[4, 3.9]$	3.95	$[1.975, 1.975]$	$[-2.025, -1.925]$	7.81

$$Loss = \frac{1}{4}\sum_{i=1}^{4} ||g\left(f\left(x^i\right)\right) - x^i||^2 = \frac{1}{4}\left(0.406 + 2.356 + 4.66 + 7.81\right) = 3.8$$

$\therefore$ The answer is 15.224 and 3.8.

12. (1 point) Consider the following 4 training examples. We want to learn a function

x	y
-1	0.0319
0	0.8692
1	1.9566
2	3.0343

$f(x) = ax + b$ which is parameterized by (a,b). Using average squared error as the loss function, which of the following parameters would be best to model the given data?

A. (1, 1)

B. (1, 2)

C. (2, 1)

D. (2, 2)

Answer: A

Solution: For each of the parameters given, we have a different function to estimate y. For each function we will estimate each label y^i.

Then the loss will be given by $\frac{1}{4}\sum_{i=1}^{4} ||f(x^i) - y^i||^2$.

x	y	$x+1$	$x+2$	$2x+1$	$2x+2$
-1	0.0319	0	1	-1	0
0	0.8692	1	2	1	2
1	1.9566	2	3	3	4
2	3.0343	3	4	5	6

Let's go over each option one by one.

A. $f(x) = x + 1$

$$\begin{aligned} Loss &= (0 - 0.0319)^2 + (1 - 0.8692)^2 + (2 - 1.9566)^2 + (3 - 3.0343)^2 \\ &= (-0.0319)^2 + 0.13^2 + 0.0434^2 + (-0.0343)^2 \\ &= 0.005 \end{aligned}$$

B. $f(x) = x + 2$

$$\begin{aligned} Loss &= (1 - 0.0319)^2 + (2 - 0.8692)^2 + (3 - 1.9566)^2 + (4 - 3.0343)^2 \\ &= 0.9681^2 + 1.13^2 + 1.0434^2 + 0.9657^2 \\ &= 1.058 \end{aligned}$$

C. $f(x) = 2x + 1$

$$\begin{aligned} Loss &= (-1 - 0.0319)^2 + (1 - 0.8692)^2 + (3 - 1.9566)^2 + (5 - 3.0343)^2 \\ &= (-1.0319)^2 + 0.13^2 + 1.0434^2 + 1.9657^2 \\ &= 1.51 \end{aligned}$$

D. $f(x) = 2x + 2$

$$\begin{aligned} Loss &= (0 - 0.0319)^2 + (2 - 0.8692)^2 + (4 - 1.9566)^2 + (6 - 3.0343)^2 \\ &= (-0.0319)^2 + 1.13^2 + 2.0434^2 + 2.9657^2 \\ &= 3.55 \end{aligned}$$

Since, the loss for $f(x) = x + 1$ is the smallest, the parameters $(1, 1)$ are the best fit for this model.

$\therefore$ Option A is correct.

13. (1 point) Consider the following input data points:

X	y
[2]	5.8
[3]	8.3
[6]	18.3
[7]	21
[8]	22

What will be the amount of loss when the functions $g = 3x_1 + 1$ and $h = 2x_1 + 2$ are used to represent the regression line. Consider the average squared error as loss function.

g: ____

h: _____

Answer: g: 2.964 [Range could be 2.82 to 3.11]
h: 11.924 [Range could be 11.32 to 12.52]

Solution:

The average squared loss for regression line $f(x)$ is given by $\frac{1}{5}\sum_{i=1}^{5}(y - f(x^i))^2$

$\boldsymbol{x}$	y	$g(\boldsymbol{x})$	$h(\boldsymbol{x})$	$(y - g(\boldsymbol{x}))^2$	$(y - h(\boldsymbol{x}))^2$
2	5.8	7	6	1.44	0.04
3	8.3	10	8	2.89	0.09
6	18.3	19	14	0.49	18.49
7	21	22	16	1	25
8	22	25	18	9	16
				14.82	59.62

We can see that loss for $g(x) = 3x_1 + 1$ is $\frac{14.82}{5} = 2.964$ and the loss for $h(x) = 2x_1 + 2$ is $\frac{59.62}{5} = 11.924$.

$\therefore$ Answer is 2.964 and 11.924.

14. (2 points) Consider the following input data points:

X	y
$[4, 2]$	+1
$[8, 4]$	+1
$[2, 6]$	-1
$[4, 10]$	-1
$[10, 2]$	+1
$[12, 8]$	-1

What will be the average misclassification error when the functions $g(\mathbf{X}) = sign(x_1 - x_2 - 2)$ and $h(\mathbf{X}) = sign(x_1 + x_2 - 10)$ are used to classify the data points into classes $+1$ or -1.

g: _____

h: _____

Answer: g: 1/6(Range 0.158 to 0.175)
h: 1/2 (Range 0.475 to 0.525)

Solution:

The average misclassification error for a function $f(x)$ is given by $\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}(f(X^i) \neq y^i)$

$\boldsymbol{x}$	y	$g(\boldsymbol{x})$	$h(\boldsymbol{x})$	$\mathbf{1}(y \neq g(\boldsymbol{x}))$	$\mathbf{1}(y \neq h(\boldsymbol{x}))$
$(4,2)$	1	1	-1	0	1
$(8,4)$	1	1	1	0	0
$(2,6)$	-1	-1	-1	0	0
$(4,10)$	-1	-1	1	0	1
$(10,2)$	1	1	1	0	0
$(12,8)$	-1	1	1	1	1
				1	3

So, the loss for $g(x) = sign(x_1 - x_2 - 2)$ is $\frac{1}{6}$ and the loss for $h(x) = sign(x_1 + x_2 - 10)$ is $\frac{1}{2}$.

$\therefore$ Answer is $\frac{1}{6}$ and $\frac{1}{2}$.

15. (1 point) $f(x_1, x_2, x_3) = \frac{x_1+2x_2}{2}$ is used as encoder function and $g(u) = [u, 2u, 3u]$ is used as decoder function for dimensionality reduction of following data set.

X
[1,2,3]
[2,3,4]
[-1,0,1]
[0,1,1]

Give the reconstruction error for this encoder decoder pair. The reconstruction error is the mean of the squared distance between the reconstructed input and input.

Answer: 34.5 (Range 32.78 to 36.22)

Solution:

We are given an encoder (f) and a decoder (g) function. To solve this question, we will take each datapoint x^i and encode it using encoder function getting $f(x^i)$ and then decode it to get $g(f(x^i))$. Then, the squared error would be given as $||g(f(x^i)) - x^i||^2$. We would then take the average of this error over all datapoints to get the loss.

$\boldsymbol{x}^i$	$f(\boldsymbol{x}^i)$	$g(f(\boldsymbol{x}^i))$	$\|\|g(f(\boldsymbol{x}^i)) - \boldsymbol{x}^i\|\|^2$
$(1,2,3)$	2.5	$(2.5, 5, 7.5)$	31.5
$(2,3,4)$	4	$(4,8,12)$	93
$(-1,0,1)$	-0.5	$(-0.5, -1, -1.5)$	7.5
$(0,1,1)$	1	$(1,2,3)$	6
			138

So, the loss will be $\frac{138}{4} = 34.5$.

$\therefore$ Answer is 34.5.