Exercise sheet: Review of Probability

The following exercises have different levels of difficulty indicated by (*), (**), (***). An exercise with (*) is a simple exercise requiring less time to solve compared to an exercise with (***), which is a more complex exercise.

1. (*) The table below gives details of symptoms that patients presented and whether they were suffering from meningitis Using this dataset, calculate the following probabilities

ID	Headache	Fever	Vomiting	Meningitis
1	true	true	false	false
2	false	true	false	false
3	true	false	true	false
4	true	false	true	false
5	false	true	false	true
6	true	false	true	false
7	true	false	true	false
8	true	false	true	true
9	false	true	false	false
10	true	false	true	true

   (a) $P(\text{Vomiting} = \text{true})$.
   (b) $P(\text{Headache} = \text{false})$.
   (c) $P(\text{Headache} = \text{true}, \text{Vomiting} = \text{false})$.
   (d) $P(\text{Vomiting} = \text{false}|\text{Headache} = \text{true})$.
   (e) $P(\text{Meningitis} = \text{true}|\text{Fever} = \text{true}, \text{Vomiting} = \text{false})$.

2. (*) Consider the experiment of tossing a coin three times. Let X be the RV giving the number of heads obtained. We assume that the tosses are independent and the probability of a head is p. Find the probabilities $P(X = 0), P(X = 1), P(X = 2)$, and $P(X = 3)$.

3. (***) Suppose that the two RVs X and Z are statistically independent. Show that the mean and variance of their sum satisfies

$$E\{X + Z\} = E\{X\} + E\{Z\}$$
$$\text{var}\{X + Z\} = \text{var}\{X\} + \text{var}\{Z\}.$$

4. (*) Consider a discrete RV X whose pmf is given as

$$P(X) = \begin{cases} \frac{1}{3}, & \text{if } x = -1, 0, 1, \\ 0 & \text{otherwise} \end{cases}$$

Find the mean and variance of X.

5. (**) The RV X can take values $x_1 = 1$ and $x_2 = 2$. Likewise, the RV Y can take values $y_1 = 1$ and $y_2 = 2$. The joint pmf of the RVs X and Y is given as

$$P(X,Y) = \begin{cases} k(2x_i + y_j), & \text{for } i = 1,2\,; j = 1,2 \\ 0 & \text{otherwise,} \end{cases}$$

where k is a constant.

   (a) Find the value of k.

   (b) Find the marginal pmf for X and Y.

   (c) Are X and Y independent?

6. (**) The joint pdf of the RVs X and Y is given by

$$p(x,y) = \begin{cases} k(x+y), & \text{for } 0 < x < 2, 0 < y < 2 \\ 0 & \text{otherwise,} \end{cases}$$

where k is a constant.

   (a) Find the value of k.

   (b) Find the marginal pdf for X and Y.

   (c) Are X and Y independent?

7. (**) Suppose that we have three coloured boxes r (red), b (blue), and g (green). Box r contains 3 apples, 4 oranges, and 3 limes, box b contains 1 apple, 1 orange, and 0 limes, and box g contains 3 apples, 3 oranges, and 4 limes. If a box is chosen at random with probabilities $P(r) = 0.2, P(b) = 0.2, P(g) = 0.6$, and a piece of fruit is removed from the box (with equal probability of selecting any of the items in the box), then what is the probability of selecting an apple? If we observe that the selected fruit is in fact an orange, what is the probability that it came from the green box?

Exercise sheet: End-to-end ML project

The following exercises have different levels of difficulty indicated by (*), (**), (***). An exercise with (*) is a simple exercise requiring less time to solve compared to an exercise with (***), which is a more complex exercise.

1. (*) You have built an ML classifier that detects whether a tissue appearing in an image is cancerous or not. Consider the cancerous class as the positive class. The following confusion matrix shows the predicted results obtained in the validation set

	cancerous (predicted)	healthy (predicted)
cancerous (actual)	30	5
healthy (actual)	15	100

   Compute the precision, recall and accuracy of your ML classifier.

2. (*) Table 1 below shows the scores achieved by a group of students on an exam. Using this data, perform the following tasks on the Score feature

   (a) A normalisation in the range $[0, 1]$.

   (b) A normalisation in the range $[-1, 1]$.

   (c) A standardisation of the data.

ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Score	42	47	59	27	84	49	72	43	73	59	58	82	50	79	89	75	70	59	67	35

Table 1: Students' score

3. (*) We designed a model for predicting the number of bike rentals (y) from two attributes, temperature (x_1) and humidity (x_2),

$$y = 500 \times x_1 + 300 \times x_2.$$

   The model was trained with normalised data with values $\min x_1 = -10$ and $\max x_1 = 39$ for x_1, and values $\min x_2 = 20$ and $\max x_2 = 100$. At test time, the model is used to predict the bike rentals for a vector $\mathbf{x}_* = [25,\ 70]^\top$. What is the value of the prediction y?

4. (*) A simple criterion to remove outliers from a dataset is to compute the mean, μ, and the standard deviation, σ, of the variable of interest and consider values outside the range $(\mu - 3\sigma, \mu + 3\sigma)$ as outliers. Applying this criterion to the Scores in Exercise 2, which ones of them can be considered as outliers?

5. (**) Suppose the joint pmf of the two RVs X and Y is given as

$$P(X = x_i, Y = y_j) = \begin{cases} \frac{1}{3}, & \text{for } (x_1 = 0, y_1 = 1), (x_2 = 1, y_2 = 0), (x_3 = 2, y_1 = 1) \\ 0 & \text{otherwise,} \end{cases}$$

(a) Are X and Y independent?

(b) Are X and Y uncorrelated?

6. (**) Two RVs X and Y are uncorrelated if $\sigma_{X,Y} = 0$. Since $\sigma_{X,Y} = E\{XY\} - E\{X\}E\{Y\}$, the two RVs are uncorrelated if $E\{XY\} = E\{X\}E\{Y\}$. Show that if the RVs are independent, then they are also uncorrelated.
[HINT: the expected value $E\{XY\}$ is defined as

$$E\{XY\} = \sum_{\forall x_i} \sum_{\forall y_j} x_i y_j P(x_i, y_j),$$

where $P(x_i, y_j)$ is the joint pmf for the discrete RVs X and Y. A similar definition can be written if X and Y are continuous RVs, replacing the sums for integrals.]

7. (***) Let $Y = aX + b$, where Y and X are RVs and a and b are constants.

(a) Find the covariance of X and Y.

(b) Find the correlation coefficient of X and Y.

Exercise sheet: Decision trees and ensemble methods

The following exercises have different levels of difficulty indicated by (*), (**), (***). An exercise with (*) is a simple exercise requiring less time to solve compared to an exercise with (***), which is a more complex exercise.

1. (*) The table below lists a sample of data from a census. There are four descriptive features and one

ID	AGE	EDUCATION	MARITAL STATUS	OCCUPATION	ANNUAL INCOME
1	39	bachelors	never married	transport	25K-50K
2	50	bachelors	married	professional	25K-50K
3	18	high school	never married	agriculture	$\leq$25K
4	28	bachelors	married	professional	25K-50K
5	37	high school	married	agriculture	25K-50K
6	24	high school	never married	armed forces	$\leq$ 25K
7	52	high school	divorced	transport	25K-50K
8	40	doctorate	married	professional	$\geq$ 50K

   target feature in this dataset: AGE, EDUCATION, MARITAL STATUS and OCCUPATION. The target feature is the ANNUAL INCOME.

   (a) Calculate **information gain** (based on entropy) for the EDUCATION, MARITAL STATUS, and OCCUPATION features.

   (b) Calculate **information gain** using the **Gini index** for the EDUCATION, MARITAL STATUS, and OCCUPATION features.

   (c) When building a decision tree, the easiest way to handle a continuous feature is to define a threshold around which splits will be made. What would be the optimal threshold to split the continuous AGE feature (use information gain based on entropy as the feature selection measure)?

2. (*) The following table lists a dataset of the scores students achieved on an exam described in terms of whether the student studied for the exam (STUDIED) and the energy level of the lecturer when grading the student's exam (ENERGY). Which of the two descriptive features should we use as the

ID	STUDIED	ENERGY	SCORE
1	yes	tired	65
2	no	alert	20
3	yes	alert	90
4	yes	tired	70
5	no	tired	40
6	yes	alert	85
7	no	tired	35

   testing criterion at the root node of a decision tree to predict students' scores?

3. (**) The following table lists a dataset containing the details of five participants in a heart disease study, and a target feature RISK which describes their risk of heart disease. Each patient is described

in terms of four descriptive features: EXERCISE (how regularly do they exercise?), SMOKER (do they smoke?), OBESE (are they overweight?) FAMILY (did any of their parents or siblings suffer from heart disease?).

ID	EXERCISE	SMOKER	OBESE	FAMILY	RISK
1	daily	false	false	yes	low
2	weekly	true	false	yes	high
3	daily	false	false	no	low
4	rarely	true	true	yes	high
5	rarely	true	true	no	high

(a) As part of the study researchers have decided to create a predictive model to screen participants based on their risk of heart disease. You have been asked to implement this screening model using a **random forest**. The three tables below list three bootstrap samples that have been generated from the above dataset. Using these bootstrap samples create the decision trees that will be in the random forest model (use entropy based information gain as the feature selection criterion).

ID	EXERCISE	FAMILY	RISK
1	daily	yes	low
2	weekly	yes	high
2	weekly	yes	high
5	rarely	no	high
5	rarely	no	high

Boostrap Sample A

ID	SMOKER	OBESE	RISK
1	false	false	low
2	true	false	high
2	true	false	high
4	true	true	high
5	true	true	high

Boostrap Sample B

ID	OBESE	FAMILY	RISK
1	false	yes	low
1	false	yes	low
2	false	yes	high
4	true	yes	high
5	true	no	high

Boostrap Sample C

(b) Assuming the random forest model you have created uses majority voting, what prediction will it return for the following query:

EXERCISE=rarely, SMOKER=false, OBESE=true, FAMILY=yes.

Exercise sheet: Linear regression

The following exercises have different levels of difficulty indicated by (*), (**), (***). An exercise with (*) is a simple exercise requiring less time to solve compared to an exercise with (***), which is a more complex exercise.

1. (*) Given the two vectors,

$$\mathbf{x} = \begin{bmatrix} 1.3 \\ -2.0 \\ 4.1 \end{bmatrix}, \qquad \mathbf{y} = \begin{bmatrix} 0.4 \\ -0.8 \\ -1.1 \end{bmatrix}.$$

   compute their inner product and their outer product.

2. (**) Let us define a matrix $\mathbf{W}$ of dimensions $n \times m$, a vector $\mathbf{x}$ of dimensions $m \times 1$ and a vector $\mathbf{y}$ of dimensions $n \times 1$. Write the following expression in matrix form

$$\sum_{i=1}^{n}\sum_{j=1}^{m} w_{i,j}x_j + \sum_{j=1}^{m}\sum_{i=1}^{n} y_i w_{i,j}.$$

   [HINT: if necessary define a vector of ones $\mathbf{1}_p = [1 \cdots 1]^\top$ of dimensions $p \times 1$, where p can be any number].

3. (***) Show that using the ML criterion, the optimal value for σ_*^2 is given as in slide 40 of Lecture 4, this is,

$$\sigma_*^2 = \frac{1}{N}(\mathbf{y} - \mathbf{X}\mathbf{w}_*)^\top(\mathbf{y} - \mathbf{X}\mathbf{w}_*).$$

4. (*) You are given a dataset with the following instances, $(x_1, y_1) = (0.8, -1.2), (x_2, y_2) = (-0.3, -0.6)$, and $(x_3, y_3) = (0.1, 2.4)$. Find the optimal value $\mathbf{w}_*$ used in ridge regression with a regularisation parameter $\lambda = 0.1$.

5. (***) Consider a regression problem for which each observed output y_n has an associated weight factor $r_n > 0$, such that the mean of weighted squared errors is given as

$$E(\mathbf{w}) = \frac{1}{N}\sum_{n=1}^{N} r_n(y_n - \mathbf{w}^\top \mathbf{x}_n)^2,$$

   where $\mathbf{w} = [w_0, \ldots, w_D]^\top$ is the vector of parameters, and $\mathbf{x}_n \in \mathbb{R}^{D+1\times 1}$ with $x_{n,0} = 1$.

   (a) Starting with the expression above, write the mean of weighted squared errors in matrix form. You should include each of the steps necessary to get the matrix form solution. [HINT: a diagonal matrix is a matrix that is zero everywhere except for the entries on its main diagonal. The weight factors $r_n > 0$ can be written as the elements of a diagonal matrix $\mathbf{R}$ of size $N \times N$].

(b) Find the optimal value of $\mathbf{w}$, $\mathbf{w}_*$, that minimises the mean of weighted squared errors. The solution should be in matrix form. Use matrix derivatives.

6. (*) A dataset is used to train a linear regression model with polynomial basis functions $\{\phi_i(x) = x^i\}_{i=1}^{M}$, where $M = 4$. Assume that the weight vector after training is equal to $\mathbf{w}_* = [0.5, -0.8, 1.2, 1.3, -0.3]^\top$. What would be the predicted value for this linear model when the input is $x = 2.5$?

7. (***) Show that the optimal solution for $\mathbf{w}_*$ in ridge regression is given as in slide 63 of Lecture 4, this is,

$$\mathbf{w}_* = \left(\mathbf{X}^\top\mathbf{X} + \frac{\lambda N}{2}\mathbf{I}\right)^{-1}\mathbf{X}^\top\mathbf{y}.$$

Exercise sheet: Auto-diff

The following exercises have different levels of difficulty indicated by (*), (**), (***). An exercise with (*) is a simple exercise requiring less time or effort to solve compared to an exercise with (***), which is a more complex exercise.

Let $\mathbf{f}$ be a vector-valued function that maps from $\mathbb{R}^3$ to $\mathbb{R}^2$,

$$y_1 = f_1(x_1, x_2, x_3) = x_1 x_3 + \log(x_2 + x_1) \times e^{-x_3}$$
$$y_2 = f_2(x_1, x_2, x_3) = e^{-x_2} + \cos(x_1 x_3).$$

1. (*) Compute the Jacobian using manual differentiation and evaluate the Jacobian at the point ($x_1 = 3, x_2 = 5, x_3 = 1$)

2. (*) Compute the Jacobian at the same point that in the previous point, but using finite difference approximation.

3. (*) Draw the computational graph.

4. (**) Compute the Jacobian using AD in forward mode. Write the expressions for all the intermediate variables $\dot{v}_i$ in the forward tangent trace.

5. (**) Compute the Jacobian using AD in reverse mode. Write the expressions for all the adjoints $\bar{v}_i$ in the reverse derivative trace.

MLAI Week 6 Exercise: Logistic regression & PyTorch for deep learning

Note: An indicative mark is in front of each question. The total mark is 12. You may mark your own work when we release the solutions.

[3] 1. Figure 1 shows the COVID test results at a centre:
1) convert it to a probability table following a similar example in Lecture 6
2) calculate the observed odds of COVID positive for the age group of 20–29.

Age	COVID
9	1
10	1
15	0
17	1
23	1
25	1
28	0
30	0
33	0
33	1
38	0

Age	COVID
41	0
42	1
46	0
47	1
48	1
49	1
49	0
50	1
51	1
51	0
52	1

Age	COVID
54	1
55	0
58	1
60	1
60	0
62	1
65	0
67	1
71	1
77	1
81	1

Figure 1: Age and COVID test results: 0 = negative, 1 = positive.

[2] 2. Derive π from $\log \frac{\pi}{1-\pi} = \mathbf{w}^\top \mathbf{x}$ (slide 22), i.e. derive the logistic function from the logit function.

[2] 3. The last equation on slide 23 writes the log likelihood in terms of π_i. Rewrite the equation in terms of the weight vector $\mathbf{w}$ and input vector $\mathbf{x}$.

[2] 4. In a binary (two-class) logistic regression model, the weight vector $\mathbf{w} = [4, -2, 5, -3, 11, 9]$. We apply it to some object that we'd like to classify; the vectorized feature representation of this object is $\mathbf{x} = [6, 8, 2, 7, -3, 5]$. What is the probability, according to the model, that this instance belongs to the positive class?

3 5. Consider the fully connected neural network (multilayer perceptron) on slide 33. If we insert two new hidden layers between the old Hidden Layer (4 neurons) and the Output Layer (2 neurons), i.e., New Layer 1 (6 neurons) after old Hidden layer, New Layer 2 (5 neurons) after New layer 1, and Output Layer after New Layer 2, with full connections between all adjacent layers and no other connections. The same activation function sigma (sigmoid) is used in the new hidden layers. How many learnable parameters in total are there for this three-hidden-layer neural network?

MLAI Week 7 Exercise: Neural Networks

Note: An indicative mark is in front of each question. The total mark is 12. You may mark your own work when we release the solutions.

[2] 1. Using the definitions for **o** and **h** on slide 10 of Lecture 7 to show that if the activation function is linear such that $g(a) = a$, then the one-hidden-layer on that slide encodes a linear relationship between the input **x** and output **o**. Include all steps.

[1] 2. In Slide 38: we change the 3×3 kernel to $\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$. What will be the 3×3 convolved features? What features can this kernel detect?

[3] 3. We have a $512 \times 512 \times 3$ colour image. We apply 100 5×5 filters with stride 7, and pad 2 to obtain a convolution output. What is the output volume size? How many parameters are needed for such a layer?

[6] 4. For the AlexNet depicted in Slide 35 of Lecture 6, there are about 60 million learnable parameters. With the help of the illustration `https://static.packt-cdn.com/products/9781789956177/graphics/assets/ec08175c-5282-4be2-b6e7-6f2d99272166.png`, compute the exact number of learnable parameters in AlexNet, showing the steps.

MLAI Week 8 Exercise: Unsupervised Learning

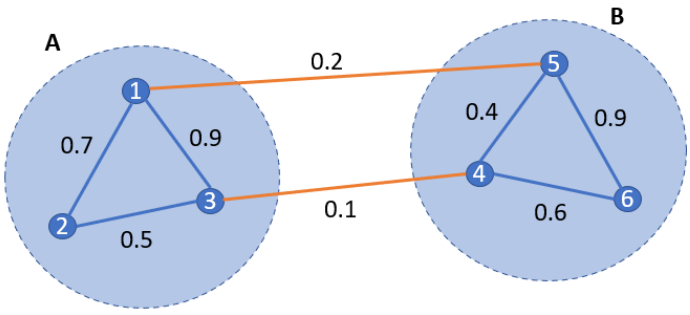
Note: An indicative mark is in front of each question. The total mark is 13. You may mark your own work when we release the solutions.

[1] 1. Consider 30-bit deep colour images of size 1200×1200. How many possible images of this size and bit depth are there?

[2] 2. We are using PCA to reduce data dimensionality from 3 to 2. The top two eigenvectors are $\begin{pmatrix} 0.4729 & -0.8817 \\ -0.8817 & -0.4719 \\ 0 & 0 \end{pmatrix}$ where each column is an eigenvector. Use this PCA transformation to reduce the dimensionality of two data points $\mathbf{x}_1 = (2, 3, 3)^\top$ and $\mathbf{x}_2 = (4, 1, 0))^\top$ to 2 as $\hat{\mathbf{x}}_1$ and $\hat{\mathbf{x}}_2$. Show the procedures to compute $\hat{\mathbf{x}}_1$ and $\hat{\mathbf{x}}_2$.

[2] 3. Given a dataset $\{0, 2, 4, 6, 24, 26\}$, initialise the k-means clustering algorithm with 2 cluster centres $c_1 = 3$ and $c_2 = 4$. What are the values of c_1 and c_2 after one iteration of k-means? What are the values of c_1 and c_2 after the second iteration of k-means?

[2] 4. For the graph below, compute the normalised cut $Ncut(A, B)$.



[3] 5. An alternative to derive PCA is to minimize the reconstruction error (Slide 26) for all N data samples $\mathbf{x}^{(i)}, i = 1, \cdots, N$, assuming that the mean $\boldsymbol{\mu} = \sum_i \mathbf{x}^{(i)}$ is zero. Take this approach to derive the first principal component (as the first eigenvector of the data matrix).

[3] 6. In spectral clustering, show that the smallest eigenvalue for the formulated generalized eigenvalue problem on Slide 41 is 0 with the corresponding generalized eigenvector $\mathbf{y} = \mathbf{1}$, hence the same "representation/embedding" for all nodes.

MLAI Week 9 Exercise: Generative Models

Note: An indicative mark is in front of each question. The total mark is 7. You may mark your own work when we release the solutions.

[1] 1. Slide 19: if the observed data point is $(x = -0.9, y = -0.1)$ instead, sketch what the likelihood will look like.

[2] 2. Slide 20: if the second observed data point is $(x = -0.7, y = 0.8)$ instead, sketch what the posterior will look like on this slide, assuming the first observed data point is still as it is $(x = 0.9, y = 0.1)$.

[1] 3. Slide 26: What is/are the sufficient statistics for a Bernoulli distribution?

[3] 4. Slide 36: show how to obtain a variable z with a normal distribution of mean μ and standard deviation (std) σ from a standard normal distribution with a mean of zero and std of 1 and verify the mean and std of z are indeed μ and σ respectively.