Exercise sheet: Review of Probability

Solutions prepared by: Mr Chunchao Ma, Mr Areeb Sherwani. Supervised by M Álvarez

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) $P(\text{Meningitis} = \text{true}|\text{Fever} = \text{true}, \text{Vomiting} = \text{false})$.

Solution:
(a) Vomiting = true corresponds to IDs 3,4,6,7,8 and 10 (6 samples). Since there is no conditioning, we consider the full set of data (10 samples). Therefore
$P(\text{Vomiting} = \text{true}) = 6/10 = 0.60$

(b) Headache = false corresponds to IDs 2, 5 and 9 (3 samples). Since there is no conditioning, we consider the full set of data (10 samples). Therefore
$P(\text{Headache} = \text{false}) = 3/10 = 0.30$

(c) Since we're taking the joint here, we're looking for the IDs for which both statements are correct. That's only ID 1 (1 sample).Since there is no conditioning, we consider the full set of data (10 samples). Therefore
$P(\text{Headache} = \text{true}, \text{Vomiting} = \text{false}) = 1/10 = 0.10$

(d) We're asked to find the probability of picking a symptom where Vomiting is false, given that they already have the symptom of Headache being true. Since we're conditioning on the statement Headache = true, we first identify the IDs for which this is correct. These are 1,3,4,6,7,8 and 10 (7

samples). This is the set we'll be working with. Next, of this set, we identify the IDs for which Vomiting = false. This is only ID 1 (1 sample). Therefore
$P(\text{Vomiting} = \text{false}|\text{Headache} = \text{true}) = 1/7 = 0.14$

(e) We're asked to find the probability of picking a symptom where Meningitis is true, given that they already have the symptom of Fever being true as well as Vomiting being false. Since we're conditioning on the statement Fever = true and Vomiting = false, we first identify the IDs for which this is correct. These are 1,2,5 and 9 (4 samples). This is the set we'll be working with. Next, of this set, we identify the IDs for which Meningitis = true. This is only ID 5 (1 sample). Therefore
$P(\text{Meningitis} = \text{true}|\text{Fever} = \text{true}, \text{Vomiting} = \text{false}) = 1/4 = 0.25$

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 $P(X = 3)$.

Solution:
For this, we need to find the sum of the probabilities of each event that satisfies the condition. We can do this by finding the probability of a single event that satisfies the condition (for example getting HTT satisfies $X = 1$) and multiplying it by all the different combinations that satisfy it (HTT, THT and TTH gives us 3 combinations). The probability for getting n heads in 3 trials is $p^n(1-p)^(3-n)$. The number of combinations of n heads in 3 trials is given by $\binom{3}{n}$. Giving us a final formula of

$$P(X = n) = \binom{n}{x} p^n (1-p)^{3-n}$$

Therefore

$$P(X = 0) = \binom{3}{0} p^0 (1-p)^{3-0} = (1-p)^3$$

$$P(X = 1) = \binom{3}{1} p^1 (1-p)^{3-1} = 3(1-p)^2 p$$

$$P(X = 2) = \binom{3}{2} p^2 (1-p)^{3-2} = 3(1-p)p^2$$

$$P(X = 3) = \binom{3}{3} p^3 (1-p)^{3-3} = p^3$$

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

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

Solution:

Since X and Z are independent, their joint distribution factorises $p(x,z) = p(x)p(z)$, and so

$$\begin{aligned}
\mathbb{E}\{X+Z\} &= \int\int (x+z)p(x,z)\mathrm{d}x\mathrm{d}z = \int\int (x+z)p(x)p(z)\mathrm{d}x\mathrm{d}z \\
&= \int\int xp(x)p(z)\mathrm{d}x\mathrm{d}z + \int\int zp(x)p(z)\mathrm{d}z\mathrm{d}x \\
&= \int xp(x)\mathrm{d}x \int p(z)\mathrm{d}z + \int zp(z)\mathrm{d}z \int p(x)\mathrm{d}x \\
&= \int xp(x)\mathrm{d}x + \int zp(z)\mathrm{d}z \\
&= \mathbb{E}\{X\} + \mathbb{E}\{Z\}
\end{aligned}$$

where we have used $\int p(z)\mathrm{d}z = 1$ and $\int p(x)\mathrm{d}x$.

Similarly for the variances, say we first define $W = X + Z$. We are want to compute

$$\text{var}\{W\} = E\{W - E\{W\}\}^2.$$

By replacing W for $X+Z$ in the expression above, we get

$$\begin{aligned}
\text{var}\{X+Z\} &= E\{X+Z-E\{X+Z\}\}^2 = E\{(X-E\{X\}) + (Z-E\{Z\})\}^2 \\
&= E\{(X-\mathbb{E}\{X\})^2 + (Z-\mathbb{E}\{Z\})^2 + 2(X-\mathbb{E}\{X\})(Z-\mathbb{E}\{Z\})\}.
\end{aligned}$$

Applying the definition of the expected value, we get

$$\begin{aligned}
\text{var}\{X+Z\} &= \int\int \left[(x-\mathbb{E}\{X\})^2 + (z-\mathbb{E}\{Z\})^2 + 2(x-\mathbb{E}\{X\})(z-\mathbb{E}\{Z\})\right] p(x,z)\mathrm{d}x\mathrm{d}z \\
&= \int\int (x-\mathbb{E}\{X\})^2 p(x,z)\mathrm{d}x\mathrm{d}z + \int\int (z-\mathbb{E}\{Z\})^2 p(x,z)\mathrm{d}x\mathrm{d}z \\
&+ 2\int\int (x-\mathbb{E}\{X\})(z-\mathbb{E}\{Z\})p(x,z)\mathrm{d}x\mathrm{d}z.
\end{aligned}$$

Because $p(x,z) = p(x)p(z)$, due to independence, the three double integrals follow as

$$\begin{aligned}
\int\int (x - \mathbb{E}\{X\})^2 p(x,z)\mathrm{d}x\mathrm{d}z &= \int\int (x - \mathbb{E}\{X\})^2 p(x)p(z)\mathrm{d}x\mathrm{d}z \\
&= \int (x - \mathbb{E}\{X\})^2 p(x)\mathrm{d}x \int p(z)\mathrm{d}z = \mathrm{var}\{X\} \\
\int\int (z - \mathbb{E}\{Z\})^2 p(x,z)\mathrm{d}x\mathrm{d}z &= \int\int (z - \mathbb{E}\{Z\})^2 p(x)p(z)\mathrm{d}x\mathrm{d}z \\
&= \int p(x)\mathrm{d}x \int (z - \mathbb{E}\{Z\})^2 p(z)\mathrm{d}z = \mathrm{var}\{Z\} \\
\int\int (x - \mathbb{E}\{X\})(z - \mathbb{E}\{Z\} p(x,z)\mathrm{d}x\mathrm{d}z &= \int\int (x - \mathbb{E}\{X\})(z - \mathbb{E}\{Z\}) p(x)p(z)\mathrm{d}x\mathrm{d}z \\
&= \int (x - \mathbb{E}\{X\}) p(x)\mathrm{d}x \int (z - \mathbb{E}\{Z\}) p(z)\mathrm{d}z \\
&= \left[\int x p(x)\mathrm{d}x - \int \mathbb{E}\{X\} p(x)\mathrm{d}x\right] \left[\int z p(z)\mathrm{d}z - \int \mathbb{E}\{Z\} p(z)\mathrm{d}z\right] \\
&= \left[\mathbb{E}\{X\} - \mathbb{E}\{X\} \int p(x)\mathrm{d}x\right] \left[\mathbb{E}\{Z\} - \mathbb{E}\{Z\} \int p(z)\mathrm{d}z\right] \\
&= 0.
\end{aligned}$$

Putting together these results, we get

$$\mathrm{var}\{X + Z\} = \mathrm{var}\{X\} + \mathrm{var}\{Z\}.$$

For discrete variables the integrals are replaced by summations, and the same results are again obtained.

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.

Solution:

The mean of X is

$$E(X) = \frac{1}{3}(-1 + 0 + 1) = 0$$

The variance of X is

$$\mathrm{Var}(X) = E\left[(X - E(X))^2\right] = E\left(X^2\right) = \frac{1}{3}\left[(-1)^2 + (0)^2 + (1)^2\right] = \frac{2}{3}$$

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?

Solution:

(a) Find the value of k:

$$\begin{aligned}\sum_{x_i}\sum_{y_j} P(x_i, y_j) &= \sum_{x_i=1}^{2}\sum_{y_j=1}^{2} k \times (2x_i + y_j) \\ &= k \times [(2+1) + (2+2) + (4+1) + (4+2)] = k \times 18 = 1\end{aligned}$$

We then obtain $k = \frac{1}{18}$.

(b) The marginal pmf for X is

$$\begin{aligned}P(X) &= \sum_{y_j} P(x_i, y_j) = \sum_{y_j=1}^{2} \frac{1}{18}(2x_i + y_j) \\ &= \frac{1}{18}(2x_i + 1) + \frac{1}{18}(2x_i + 2) = \frac{1}{18}(4x_i + 3) \quad x_i = 1, 2.\end{aligned}$$

We therefore obtain:

$$P(X) = \begin{cases} \frac{1}{18}(4x_i + 3), & \text{for } i = 1, 2 \\ 0 & \text{otherwise} \end{cases}$$

The marginal pmf for Y is

$$\begin{aligned}P(Y) &= \sum_{x_i} P(x_i, y_j) = \sum_{x_i=1}^{2} \frac{1}{18}(2x_i + y_j) \\ &= \frac{1}{18}(2 + y_j) + \frac{1}{18}(4 + y_j) = \frac{1}{18}(2y_j + 6) \quad y_j = 1, 2.\end{aligned}$$

We therefore obtain:

$$P(Y) = \begin{cases} \frac{1}{18}(2y_j + 6), & \text{for } j = 1, 2 \\ 0 & \text{otherwise} \end{cases}$$

(c) Now $P(X)P(Y) \neq P(X, Y)$; Hence X and Y are not 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?

Solution:

(a) Find the value of k:

$$\begin{aligned} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x,y)dxdy &= k\int_0^2\int_0^2 (x+y)dxdy \\ &= k\int_0^2 \left(\frac{x^2}{2}+xy\right)\Big|_{x=0}^{x=2} dy \\ &= k\int_0^2 (2+2y)dy = 8k = 1 \end{aligned}$$

We then obatin $k = \frac{1}{8}$

(b) The marginal pdf for X is

$$\begin{aligned} p(x) &= \int_{-\infty}^{\infty} p(x,y)dy = \frac{1}{8}\int_0^2 (x+y)dy \\ &= \frac{1}{8}\left(xy+\frac{y^2}{2}\right)\Big|_{y=0}^{y=2} = \begin{cases} \frac{1}{4}(x+1) & 0 < x < 2 \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

Since $p(x,y)$ is symmetric with respect to x and y, the marginal pdf of Y is

$$p(y) = \begin{cases} \frac{1}{4}(y+1) & 0 < y < 2 \\ 0 & \text{otherwise} \end{cases}$$

(c)Now $p(x)\,p(y) \neq p(x,y)$; Hence X and y are not 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?

Solution

Based on the question, we know that $P(r) = 0.2, P(b) = 0.2, P(g) = 0.6$. We assume $P(\text{Apple}|r)$ is the probability of selecting an apple in the red colour box; $P(\text{Apple}|b)$ is the probability of selecting an apple in the blue colour box; $P(\text{Apple}|g)$ is the probability of selecting an apple in the green colour box. Then we get

$$P(\text{Apple} \mid r) = \frac{3}{3+4+3} = \frac{3}{10},$$
$$P(\text{Apple} \mid b) = \frac{1}{1+1+0} = \frac{1}{2},$$
$$P(\text{Apple} \mid g) = \frac{3}{3+3+4} = \frac{3}{10}.$$

The probability of selecting an apple is $P(\text{Apple})$:

$$\begin{aligned} P(\text{Apple}) =& P(\text{Apple} \mid r)P(r) + P(\text{Apple} \mid b)P(b) + P(\text{Apple} \mid g)P(g) \\ =& \frac{3}{10} \times 0.2 + \frac{1}{2} \times 0.2 + \frac{3}{10} \times 0.6 = 0.34 \quad . \end{aligned}$$

For the second question, we have observed that the selected fruit is in fact an orange. We want to know $P(g \mid \text{Orange})$, that is, the probability that a selected orange coming comes from a green colour box.

Similarly, we assume $P(\text{Orange}|r)$ is the probability of selecting an orange in red colour box; $P(\text{Orange}|b)$ is the probability of selecting an orange in blue colour box; $P(\text{orange}|g)$ is the probability of selecting an orange in green colour box. Then we get

$$P(\text{Orange} \mid r) = \frac{4}{3+4+3} = \frac{4}{10},$$
$$P(\text{Orange} \mid b) = \frac{1}{1+1+0} = \frac{1}{2},$$
$$P(\text{Orange} \mid g) = \frac{3}{3+3+4} = \frac{3}{10}.$$

Based on Bayes' theorem,

$$P(g \mid \text{Orange}) = \frac{P(\text{Orange} \mid g)P(g)}{P(\text{Orange})}.$$

The probability of selecting an Orange is $P(\text{Orange})$:

$$\begin{aligned} P(\text{Orange}) =& P(\text{Orange} \mid r)P(r) + P(\text{Orange} \mid b)P(b) + P(\text{Orange} \mid g)P(g) \\ =& \frac{4}{10} \times 0.2 + \frac{1}{2} \times 0.2 + \frac{3}{10} \times 0.6 = 0.36 \quad . \end{aligned}$$

We thus obtain:

$$P(g \mid \text{Orange}) = \frac{P(\text{Orange} \mid g)P(g)}{P(\text{Orange})} = \frac{\frac{3}{10} \times 0.6}{0.36} = 0.5 \quad .$$

Exercise sheet: End-to-end ML project

Solutions prepared by: Mr Chunchao Ma, Mr Areeb Sherwani. Supervised by M Álvarez

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.

Solution:

Following Lecture 2, TP stands for **true positive**; TN stands for **true negative**; FP stands for **false positive** and FN stands for **false negative**.

In this confusion matrix, we observe: TP = 30; TN = 100; FP = 15; FN = 5.

Precision is the ratio of correct positive predictions to the overall number of positive predictions:

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} = \frac{30}{30+15} = \frac{2}{3}$$

Recall is the ratio of correct positive predictions to the overall number of positive examples in the dataset:

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} = \frac{30}{30+5} = \frac{6}{7}$$

Accuracy is the ratio of examples corrected classifed over the total number of examples classifed:

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} = \frac{30+100}{30+100+15+5} = \frac{13}{15}$$

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

Solution:

(a) A range normalisation that generates data in the range [0, 1].

To perform a range normalisation, we need the minimum and maximum of the dataset and the high and low for the target range. From the data we can see that the minimum is 27 and the maximum is 89. In the question we are told that the low value of the target range is 0 and that the high value is 1. Using these values, we normalise an individual value using the following equation:

$$s'_i = \frac{s_i - \min(s)}{\max(s) - \min(s)} \times (\text{high} - \text{low}) + \text{low},$$

where s stands for Score in the table; s_i stands for the i-th score;s'_i stands for the i-th normalised score.

So, the first score in the dataset, 42, would be normalised as follows:

$$\begin{aligned} s'_1 &= \frac{42 - 27}{89 - 27} \times (1 - 0) + 0 \\ &= \frac{15}{62} \\ &= 0.2419 \end{aligned}$$

This is repeated for each instance in the dataset to give the full normalized data set as (we keep two decimals).

ID	1	2	3	4	5	6	7	8	9	10
Score	0.24	0.32	0.52	0.00	0.92	0.35	0.73	0.26	0.74	0.52

ID	11	12	13	14	15	16	17	18	19	20
Score	0.50	0.89	0.37	0.84	1.00	0.77	0.69	0.52	0.65	0.13

(b) A range normalisation that generates data in the range [-1, 1].

This normalisation differs from the previous range normalisation only in that the high and low values are different in this case, -1 and 1. So the first score in the dataset, 42, would be normalized as follows:

$$\begin{aligned} s'_1 &= \frac{42 - 27}{89 - 27} \times (1 - (-1)) + (-1) \\ &= \frac{15}{62} \times 2 - 1 \\ &= -0.5161 \end{aligned}$$

Applying this to each instance in the dataset gives the full normalized dataset as (we keep two decimals).

ID	1	2	3	4	5	6	7	8	9	10
Score	-0.52	-0.35	0.03	-1.00	0.84	-0.29	0.45	-0.48	0.48	0.03

ID	11	12	13	14	15	16	17	18	19	20
Score	0.00	0.77	-0.26	0.68	1.00	0.55	0.39	0.03	0.29	-0.74

(c) A standardisation of the data.

To perform a standardisation, we use the following formula for each instance in the dataset:

$$x'_i = \frac{s_i - \mu}{\sigma},$$

where s_i stands for i-th score; x'_i stands for the i-th standardised score; μ and σ are the mean and standard deviation of Score dataset. So we need the mean, μ and standard deviation, σ for the feature to be standardized. In this case, the mean is calculated from the original dataset as 60.95, and the standard deviation is 17.2519. So the standardized value for the first instance in the dataset can be calculated as

$$\begin{aligned} x'_i &= \frac{42 - 60.95}{17.2519} \\ &= -1.0984 \end{aligned}$$

Standardizing in the same way for the rest of the dataset gives us the following (we use two decimals):

ID	1	2	3	4	5	6	7	8	9	10
Score	-1.10	-0.81	-0.11	-1.97	1.34	-0.69	0.64	-1.04	0.70	-0.11

ID	11	12	13	14	15	16	17	18	19	20
Score	-0.17	1.22	-0.63	1.05	1.63	0.81	0.52	-0.11	0.35	-1.50

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?

Solution:

We first normalise the test vector $\mathbf{x}_* = [25, 70]^\top$ and use the normalised vector into the predictive equation

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

to get the predicted value.

Based on the lecture note 2, the normalisation formula is given as

$$\bar{x}_j = \frac{x_j - \min x_j}{\max x_j - \min x_j}$$

where $\min x_j$ and $\max x_j$ are the minimum and maximum values for the feature in the training set; $\bar{x}_j$ is the normalised valued. It also means that it is normalisation in the range [0,1].

So x_{1*} can be normalised as follows:

$$\bar{x}_{1*} = \frac{x_{1*} - \min x_1}{\max x_1 - \min x_1} = \frac{25 - (-10)}{39 - (-10)} = \frac{35}{49} = \frac{5}{7}$$

Similarly, x_{2*} can be normalised as follows:

$$\bar{x}_{2*} = \frac{x_{2*} - \min x_2}{\max x_2 - \min x_2} = \frac{70 - 20}{100 - 20} = \frac{50}{80} = \frac{5}{8}$$

Then we obtain the normalised data $\mathbf{x}_*$ that is $\bar{\mathbf{x}}_* = [\frac{5}{7}, \frac{5}{8}]^\top$. We input $\bar{\mathbf{x}}_*$ into

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

to obtain,

$$y(\bar{\mathbf{x}}_*) = 500 \times \frac{5}{7} + 300 \times \frac{5}{8} = \frac{2500}{7} + \frac{375}{2} = \frac{7625}{14}.$$

Thus, the value of the prediction y is $\frac{7625}{14} \approx 545$.

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?

   Solution:

   We've already calculated the mean (60.95) and the standard deviation (17.2519). Applying this to our range gives us

$$\begin{aligned}(\mu - 3\sigma, \mu + 3\sigma) &= (60.95 - 3 \times 17.25, 60.95 + 3 \times 17.25) \\ &= (9.19, 112.70)\end{aligned}$$

   We would reject any values that occur outside of this range. However, we can see that no values in Table 1 occur outside of this range. Therefore, we have no 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?

Solution:

(a) If X and Y are independent, their joint distribution factorises $P(X=x, Y=y) = P(X=x)\,P(Y=y)$. We therefore check whether $P(X=x, Y=y) = P(X=x)\,P(Y=y)$ or not.

First, we calculate the marginal pmf's of X ($P(X=x)$). The marginal pmf's of X are

$$\begin{aligned} P(X=0) &= \textstyle\sum_{y_j} P(X=0, y_j) = P(X=0, Y=1) = \frac{1}{3} \\ P(X=1) &= \textstyle\sum_{y_j} P(X=1, y_j) = P(X=1, Y=0) = \frac{1}{3} \\ P(X=2) &= \textstyle\sum_{y_j} P(X=2, y_j) = P(X=2, Y=1) = \frac{1}{3} \end{aligned}$$

Second, we calculate the the marginal pmf's of Y ($P(Y=y)$). The marginal pmf's of Y are

$$\begin{aligned} P(Y=0) &= \textstyle\sum_{x_1} P(x_i, Y=0) = P(X=1, Y=0) = \frac{1}{3} \\ P(Y=1) &= \textstyle\sum_{x_i} P(x_i, Y=1) = P(X=0, Y=1) + P(X=2, Y=1) = \frac{2}{3} \end{aligned}$$

We observe

$$P(X=0, Y=1) = \frac{1}{3} \neq P(X=0)P(Y=1) = \frac{2}{9}$$

Since $P(X=x, Y=y) \neq P(X=x)\,P(Y=y)$, X and Y are not independent.

(b) Two RVs X and Y are uncorrelated if $\sigma_{X,Y} = 0$, where $\sigma_{X,Y} = E\{XY\} - E\{X\}E\{Y\}$. We therefore check whether $\sigma_{X,Y} = 0$ or not.

In order to obtain $\sigma_{X,Y}$, $E\{X\}$, $E\{Y\}$ and $E\{XY\}$ should be obtained firstly. Then

$$\begin{aligned} E\{X\} &= \sum_{x_i} x_i P(X=x_i) = 0 \times P(X=0) + 1 \times P(X=1) + 2 \times P(X=2) \\ &= 0 \times \left(\frac{1}{3}\right) + 1 \times \left(\frac{1}{3}\right) + 2 \times \left(\frac{1}{3}\right) = 1 \\ E\{Y\} &= \sum_{y_j} y_j P(Y=y_j) = 0 \times P(Y=0) + 1 \times P(Y=1) = 0 \times \left(\frac{1}{3}\right) + 1 \times \left(\frac{2}{3}\right) = \frac{2}{3} \\ E\{XY\} &= \sum_{y_j} \sum_{x_1} x_i y_j P(X=x_i, Y=y_j) \\ &= 0 \times 1 \times P(X=0, Y=1) + 1 \times 0 \times P(X=1, Y=0) + 2 \times 1 \times P(X=2, Y=1) \\ &= 0 \times 1 \times \left(\frac{1}{3}\right) + 1 \times 0 \times \left(\frac{1}{3}\right) + 2 \times 1 \times \left(\frac{1}{3}\right) = \frac{2}{3} \end{aligned}$$

We obtain

$$\sigma_{X,Y} = E\{XY\} - E\{X\}E\{Y\} = \frac{2}{3} - 1 \times \left(\frac{2}{3}\right) = 0$$

Thus, X and Y are uncorrelated since $\sigma_{X,Y} = 0$.

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.]

Solution:
We already know that two RVs are uncorrelated if $E\{XY\} = E\{X\}E\{Y\}$. We therefore assume two RVs X and Y are independent, showing whether $E\{XY\} = E\{X\}E\{Y\}$ or not. Since X and Y can be both discrete and continuous RVs, we consider both discrete and continuous cases respectively.

First, we assume X and Y are two discrete RVs and are independent. Since X and Y are independent, their joint distribution factorises $P(x, y) = P(x) P(y)$. We obtain

$$\begin{aligned} E\{XY\} &= \sum_{y_j} \sum_{x_i} x_i y_j P(x_i, y_j) = \sum_{y_j} \sum_{x_i} x_i y_j P(x_i) P(y_j) \\ &= \left[\sum_{x_i} x_i P(x_i)\right] \left[\sum_{y_j} y_j P(y_j)\right] = E\{X\}E\{Y\} \end{aligned}$$

Since $E\{XY\} = E\{X\}E\{Y\}$, X and Y are uncorrelated.

Second, we assume X and Y are two continuous RVs and are independent. Since X and Y are independent, their joint distribution factorises $p(x, y) = p(x) p(y)$. We obtain

$$\begin{aligned} E\{XY\} &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xyp(x, y)dxdy = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xyp(x)p(y)dxdy \\ &= \int_{-\infty}^{\infty} xp(x)dx \int_{-\infty}^{\infty} yp(y)dy = E\{X\}E\{Y\} \end{aligned}$$

Similarly, since $E\{XY\} = E\{X\}E\{Y\}$, X and Y are uncorrelated.

Therefore, if two RVs X and Y are independent, X and Y are uncorrelated.

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.

   Solution:

(a) We assume that the variance of X is denoted as σ_X^2 and the variance of Y is denoted as σ_Y^2. So the covariance of X and Y is given as

$$\begin{aligned}
\text{cov}(X,Y) &= E[(X - E[X])(Y - E[Y])], \\
&= E[(X - E[X])(aX + b - E[aX + b])], \\
&= E[(X - E[X])(aX + b - (aE[X] + b))], \\
&= E[(X - E[X])(aX - aE[X])], \\
&= aE[(X - E[X])(X - E[X])], \\
&= aE[(X - E[X])^2], \\
&= a\sigma_X^2.
\end{aligned}$$

(b) The correlation coefficient of X and Y is given as

$$\rho_{XY} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{a\sigma_X^2}{\sigma_X \sigma_Y}.$$

We need to compute σ_Y, which is the squared root of the variance for Y, which is given by

$$\begin{aligned}
\sigma_Y^2 &= E[Y^2] - (E[Y])^2 \\
&= E[(aX + b)^2] - (aE[X] + b)^2 \\
&= E[a^2X^2 + 2abX + b^2] - \left(a^2E[X]^2 + 2abE[X] + b^2\right) \\
&= a^2E[X^2] + 2abE[X] + b^2 - a^2E[X]^2 - 2abE[X] - b^2 \\
&= a^2\left(E[X^2] - E[X]^2\right) = a^2\sigma_X^2.
\end{aligned}$$

We now have $\sigma_Y = \sqrt{\sigma_Y^2} = |a|\sigma_X$. Going back to the correlation coefficient,

$$\begin{aligned}
\rho_{XY} &= \frac{a\sigma_X^2}{\sigma_X \sigma_Y} = \frac{a\sigma_X^2}{\sigma_X \, |a| \, \sigma_X} = \frac{a\sigma_X^2}{|a| \, \sigma_X^2} \\
&= \frac{a}{|a|} \\
&= \begin{cases} 1, & \text{if } a > 0 \\ -1, & a < 0 \end{cases}
\end{aligned}$$

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)?

1. Answer

(a) Based on the lecture slides, we know that computing information gain (based on entropy) involves the following three equations (we also follow the notation on the lecture slides):

$$H(t, \mathcal{D}) = -\sum_{I \in \text{levels}(t)} (P(t = I) \times \log_2(P(t = I)))$$

$$\text{rem}(d, \mathcal{D}) = \sum_{I \in \text{levels}(d)} \underbrace{\frac{|\mathcal{D}_{d=I}|}{|\mathcal{D}|}}_{weighting} \times \underbrace{H\left(t, \mathcal{D}_{d=I}\right)}_{\substack{\text{entropy of} \\ \text{partition } \mathcal{D}_{d=1}}}$$

$$IG(d, \mathcal{D}) = H(t, \mathcal{D}) - \text{rem}(d, \mathcal{D})$$

Step 1: We calculate the entropy for the target feature in this dataset, in this case annual income. Let $S = \{\leq 25K, 25K - 50K, \geq 50K\}$, be the set of different annual incomes in the dataset.

$$\begin{aligned} H(\text{ANNUAL INCOME}, \mathcal{D}) &= -\sum_{l \in S} P(\text{AN.INC.} = l) \times \log_2(P(\text{AN.INC.} = l)) \\ &= -\left(\left(\frac{2}{8} \times \log_2\left(\frac{2}{8}\right)\right) + \left(\frac{5}{8} \times \log_2\left(\frac{5}{8}\right)\right) + \left(\frac{1}{8} \times \log_2\left(\frac{1}{8}\right)\right)\right) \\ &= 1.2988 \text{ bits} \end{aligned}$$

Step 2: We then need to calculate the remainder for each feature. Below is an example for the EDUCATION feature, remember that when we calculate the entropy we need to partition the dataset on each different feature (high school, bachelors, doctorate):

$$\begin{aligned} \text{rem}(\text{EDUCATION}, \mathcal{D}) &= \left(\frac{|\mathcal{D}_{\text{EDUCATION =high school}}|}{|\mathcal{D}|} \times H\left(t, \mathcal{D}_{\text{EDUCATION= high school }}\right)\right) \\ &\quad + \left(\frac{|\mathcal{D}_{\text{EDUCATION =bachelors}}|}{|\mathcal{D}|} \times H\left(t, \mathcal{D}_{\text{EDUCATION =bachelors}}\right)\right) \\ &\quad + \left(\frac{|\mathcal{D}_{\text{EDUCATION =doctorate}}|}{|\mathcal{D}|} \times H\left(t, \mathcal{D}_{\text{EDUCATION =doctorate}}\right)\right) \\ &= -\frac{4}{8} \times \left(\left(\frac{2}{4} \times \log_2\left(\frac{2}{4}\right)\right) + \left(\frac{2}{4} \times \log_2\left(\frac{2}{4}\right)\right) + \left(0 \times \log_2\left(\frac{0}{4}\right)\right)\right) \\ &\quad - \frac{3}{8} \times \left(\left(1 \times \log_2\left(\frac{3}{3}\right)\right) + \left(0 \times \log_2\left(\frac{0}{3}\right)\right) + \left(0 \times \log_2\left(\frac{0}{3}\right)\right)\right) \\ &\quad - \frac{1}{8} \times \left(\left(0 \times \log_2\left(\frac{0}{1}\right)\right) + \left(0 \times \log_2\left(\frac{0}{1}\right)\right) + \left(1 \times \log_2(1)\right)\right) = 0.5 \end{aligned}$$

To compute the remainder for the other features we follow the same process as above.

Step 3: Compute the information gain, e.g with regard to the EDUCATION feature

$$\begin{aligned} IG(\text{EDUCATION } \mathcal{D}) &= H(\text{ANNUAL INCOME}, \mathcal{D}) - \text{rem}(\text{EDUCATION }, \mathcal{D}) \\ &= 1.2988 - 0.5 = 0.7988 \quad \text{bits} \end{aligned}$$

Again we follow the same process for the other features. The table below lists the rest of the calculations for the information gain for the EDUCATION , MARITAL STATUS , and OCUPATION features.

Split by Feature	Level	Instances	Rem.	Info. Gain
EDUCATION	high school	$\mathbf{d}_3, \mathbf{d}_5, \mathbf{d}_6, \mathbf{d}_7$	0.5	0.7988
	bachelors	$\mathbf{d}_1, \mathbf{d}_2, \mathbf{d}_3$		
	doctorate	$\mathbf{d}_8$		
MARITAL STATUS	never married	$\mathbf{d}_1, \mathbf{d}_3, \mathbf{d}_6$	0.75	0.5488
	married	$\mathbf{d}_2, \mathbf{d}_4, \mathbf{d}_5, \mathbf{d}_8$		
	divorced	$\mathbf{d}_7$		
OCCUPATION	transport	$\mathbf{d}_1, \mathbf{d}_7$	0.5944	0.7044
	professional	$\mathbf{d}_2, \mathbf{d}_4, \mathbf{d}_8$		
	agriculture	$\mathbf{d}_3, \mathbf{d}_5$		
	armed forces	$\mathbf{d}_6$		

(b) Based on the lecture slides, we know that information gain can be calculated using the **Gini index** by replacing the entropy measure with the **Gini index**. The Gini index is

$$\text{Gini}(t, \mathcal{D}) = 1 - \sum_{I \in \text{ levels } (t)} P(t = I)^2$$

We use the same idea as in part (a) but replace the entropy measure with the **Gini index**.

Step 1: We calculate the **Gini index** for the target feature in the dataset. Below is an example for the annual income feature. Again let $S = \{\leq 25K, 25K - 50K, \geq 50K\}$ be the set of annual incomes:

$$\begin{aligned}\text{Gini(AN. INC., } \mathcal{D}) &= 1 - \sum_{l \in S} P(\text{AN.INC.} = l)^2 \\ &= 1 - \left(\left(\frac{2}{8}\right)^2 + \left(\frac{5}{8}\right)^2 + \left(\frac{1}{8}\right)^2\right) = 0.5313\end{aligned}$$

Step 2: Remainder for each feature: e.g remainder for EDUCATION feature:

$$\begin{aligned}\text{rem(EDUCATION, } \mathcal{D}) = &\left(\frac{|\mathcal{D}_{\text{EDUCATION =high school}}|}{|\mathcal{D}|} \times \text{Gini}\left(t, \mathcal{D}_{\text{EDUCATION =high school}}\right)\right) \\ &+ \left(\frac{|\mathcal{D}_{\text{EDUCATION =bachelors}}|}{|\mathcal{D}|} \times \text{Gini}\left(t, \mathcal{D}_{\text{EDUCATION =bachelors}}\right)\right) \\ &+ \left(\frac{|\mathcal{D}_{\text{EDUCATION =doctorate}}|}{|\mathcal{D}|} \times \text{Gini}\left(t, \mathcal{D}_{\text{EDUCATION =doctorate}}\right)\right) \\ &= \frac{4}{8} \times \left(1 - \left(\left(\frac{2}{4}\right)^2 + \left(\frac{2}{4}\right)^2 + 0^2\right)\right) \\ &\quad + \frac{3}{8} \times \left(1 - \left(1^2 + 0^2 + 0^2\right)\right) \\ &\quad + \frac{1}{8} \times \left(1 - \left(0^2 + 0^2 + 1^2\right)\right) = 0.25\end{aligned}$$

The remainder for other features follows the same idea as above.

Step 3: Compute the information gain, E.g with regard to the EDUCATION feature

$$\begin{aligned} IG(\text{EDUCATION } \mathcal{D}) = &\text{ Gini (ANNUAL INCOME, } \mathcal{D}) - \text{rem(EDUCATION , } \mathcal{D}) \\ &= 0.5313 - 0.25 = 0.2813 \end{aligned}$$

We follow the same procedure to compute the information gain with respect to the other features. The table below lists the rest of the calculations of information gain for the EDUCATION, MARITAL STATUS, and OCCUPATION features.

Split by Feature	Level	Instances	Partition Gini Index	Rem.	Info. Gain
EDUCATION	high school	$\mathbf{d}_3, \mathbf{d}_5, \mathbf{d}_6, \mathbf{d}_7$	0.5	0.25	0.2813
	bachelors	$\mathbf{d}_1, \mathbf{d}_2, \mathbf{d}_3$	0		
	doctorate	$\mathbf{d}_8$	0		
MARITAL STATUS	never married	$\mathbf{d}_1, \mathbf{d}_3, \mathbf{d}_6$	0.4444	0.3542	0.1771
	married	$\mathbf{d}_2, \mathbf{d}_4, \mathbf{d}_5, \mathbf{d}_8$	0.375		
	divorced	$\mathbf{d}_7$	0		
OCCUPATION	transport	$\mathbf{d}_1, \mathbf{d}_7$	0	0.2917	0.2396
	professional	$\mathbf{d}_2, \mathbf{d}_4, \mathbf{d}_8$	0.4444		
	agriculture	$\mathbf{d}_3, \mathbf{d}_5$	0.5		
	armed forces	$\mathbf{d}_6$	0		

(c) First sort the instances in the dataset according to the AGE feature, as shown in the following table.

ID	AGE	ANNUAL INCOME
3	18	$< 25K$
6	24	$< 25K$
4	28	$25K - 50K$
5	37	$25K - 50K$
1	39	$25K - 50K$
8	40	$> 50K$
2	50	$25K - 50K$
7	52	$25K - 50K$

Based on this ordering, the mid-points in the AGE values of instances that are adjacent in the new ordering but that have different target levels define the possible threshold points. These points are 26, 39.5, and 45.

We calculate the information gain for each of these possible threshold points using the entropy value for the annual income feature that we calculated in part (a) of this question (1.2988 bits), and the remainder, which we determine from the partition entropy along with the size of the partition. This gives us the following values:

Split by Feature	Partition	Instances	Partition Entropy	Rem.	Info. Gain
> 26	$\mathcal{D}_1$	$\mathbf{d}_3, \mathbf{d}_6$	0	0.4875	0.8113
	$\mathcal{D}_2$	$\mathbf{d}_1, \mathbf{d}_2, \mathbf{d}_4, \mathbf{d}_5, \mathbf{d}_7, \mathbf{d}_8$	0.6500		
> 39.5	$\mathcal{D}_3$	$\mathbf{d}_1, \mathbf{d}_3, \mathbf{d}_4, \mathbf{d}_5, \mathbf{d}_6$	0.9710	0.9512	0.3476
	$\mathcal{D}_4$	$\mathbf{d}_2, \mathbf{d}_7, \mathbf{d}_8$	0.9183		
> 45	$\mathcal{D}_5$	$\mathbf{d}_1, \mathbf{d}_3, \mathbf{d}_4, \mathbf{d}_5, \mathbf{d}_6, \mathbf{d}_8$	1.4591	1.0944	0.2044
	$\mathcal{D}_6$	$\mathbf{d}_2, \mathbf{d}_7$	0		

We select the threshold as 26 since the point 26 has the highest information gain from the thresholds.

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?

2. Answer

The target feature in this question (SCORE) is continuous. When a decision tree is predicting a continuous target, we choose, at each node, the descriptive feature that minimises the weighted variance after the dataset has been split based on that feature. We need to compute the variance for each partition of each feature. Following the notation on the lecture slides, the variance at a node can be calculated using the following equation:

$$\text{var}(t, \mathcal{D}) = \frac{\sum_{i=1}^{n} (t_i - \bar{t})^2}{n-1}$$

If the dataset is split by STUDIED, we have two partition domains $\mathcal{D}_1$ and $\mathcal{D}_2$ (see the table below). $\mathcal{D}_1$ and $\mathcal{D}_2$ have four $(\mathbf{d}_1, \mathbf{d}_3, \mathbf{d}_4, \mathbf{d}_6)$ and three instances $(\mathbf{d}_4, \mathbf{d}_5, \mathbf{d}_7)$ separately. Based on the values of SCORE, we obtain $\text{var}(t, \mathcal{D}_1) = 141\frac{2}{3}$ and $\text{var}(t, \mathcal{D}_1) = 108\frac{1}{3}$. Similarly, we can obtain the variance when the data is split by the feature ENERGY (see the table below).

We choose the feature that minimises the weighted variance across the resulting partitions:

$$\mathbf{d}[best] = \underset{d \in \mathbf{d}}{\text{argmin}} \sum_{I \in \text{levels}(d)} \frac{|\mathcal{D}_{d=l}|}{|\mathcal{D}|} \times \text{var}\left(t, \mathcal{D}_{d=I}\right)$$

The table below shows the calculation of the weighted variance for each of the descriptive features in this domain.

Split by Feature	Level	Partition	Instances	$\frac{\lvert\mathcal{D}_{d=l}\rvert}{\lvert\mathcal{D}\rvert}$	var$(t, \mathcal{D})$	Weighted Variance
STUDIED	yes	$\mathcal{D}_1$	$\mathbf{d}_1, \mathbf{d}_3, \mathbf{d}_4, \mathbf{d}_6$	$\frac{4}{7}$	$141\frac{2}{3}$	127.3810
	no	$\mathcal{D}_2$	$\mathbf{d}_4, \mathbf{d}_5, \mathbf{d}_7$	$\frac{3}{7}$	$108\frac{1}{3}$	
ENERGY	alert	$\mathcal{D}_5$	$\mathbf{d}_2, \mathbf{d}_3, \mathbf{d}_6$	$\frac{3}{7}$	1525	829.7619
	tired	$\mathcal{D}_6$	$\mathbf{d}_1, \mathbf{d}_4, \mathbf{d}_5, \mathbf{d}_7$	$\frac{4}{7}$	$308\frac{1}{3}$	

From these calculations we can see that splitting the dataset using the STUDIED feature results in the lowest weighted variance. Consequently, we should use the STUDIED feature at the root node of the tree.

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.

3. Answer

(a) The entropy calculation for Bootstrap Sample A is:

$$H(\text{RISK, Sample A}) = -\sum_{l \in \left\{ \begin{array}{l} low, \\ high \end{array} \right.} P(\text{RISK} = l) \times \log_2(P(\text{RISK} = l))$$
$$= -\left(\left(\frac{1}{5} \times \log_2\left(\frac{1}{5}\right)\right) + \left(\frac{4}{5} \times \log_2\left(\frac{4}{5}\right)\right)\right) = 0.7219 \text{ bits}$$

The information gain for each of the features in Bootstrap Sample A is as follows:

Split by Feature	Level	Instances	Partition Entropy	Rem.	Info. Gain
EXERCISE	daily	$\mathbf{d}_1$	0	0	0.7219
	weekly	$\mathbf{d}_2, \mathbf{d}_2$	0		
	rarely	$\mathbf{d}_5, \mathbf{d}_5$	0		
FAMILY	yes	$\mathbf{d}_1, \mathbf{d}_2, \mathbf{d}_2$	0.9183	0.5510	0.1709
	no	$\mathbf{d}_5, \mathbf{d}_5$	0		

These calculations show that the EXERCISE feature has the highest information gain of the descriptive features in Bootstrap Sample A and should be added as the root node of the decision tree generated from Bootstrap Sample A. Additionally, splitting on EXERCISE generates pure sets so the decision tree does not need to be expanded beyond this initial test. The final tree generated for Bootstrap Sample A: if EXERCISE Level is **daily**, the RISK is **low**; if EXERCISE Level is **weekly** or **rarely**, the RISK is **high**.

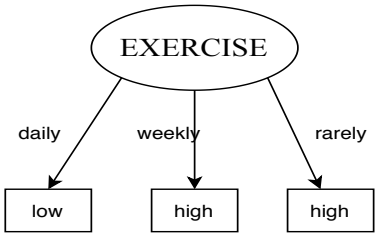


Figure 1: Decision Tree for Boostrap Sample A

By chance, Bootstrap Sample B has the same distribution of target feature values as Bootstrap Sample A, so the entropy calculation for Bootstrap Sample B is the same as the calculation for Bootstrap Sample A:

$$H(\text{RISK, Sample B}) = -\sum_{l \in \left\{ \begin{array}{l} low, \\ high \end{array} \right.} P(\text{RISK} = l) \times \log_2(P(\text{RISK} = l))$$
$$= -\left(\left(\frac{1}{5} \times \log_2\left(\frac{1}{5}\right)\right) + \left(\frac{4}{5} \times \log_2\left(\frac{4}{5}\right)\right)\right) = 0.7219 \text{ bits}$$

The information gain for each of the features in Bootstrap Sample B is as follows:

Split by Feature	Level	Instances	Partition Entropy	Rem.	Info. Gain
SMOKER	true	$\mathbf{d}_2, \mathbf{d}_2, \mathbf{d}_4, \mathbf{d}_5$	0	0	0.7219
	false	$\mathbf{d}_1$	0		
OBESE	true	$\mathbf{d}_4, \mathbf{d}_5$	0	0.5510	0.1709
	false	$\mathbf{d}_1, \mathbf{d}_2, \mathbf{d}_2$	0.9183		

These calculations show that the SMOKER feature has the highest information gain of the descriptive features in Bootstrap Sample B and should be added as the root node of the decision tree generated from Bootstrap Sample B. Additionally, splitting on SMOKER generates pure sets, so the decision tree does not need to be expanded beyond this initial test. The final tree generated for Bootstrap Sample B: if SMOKER Level is **true**, the RISK is **high**; if SMOKER Level is **false**, the RISK is **low**.

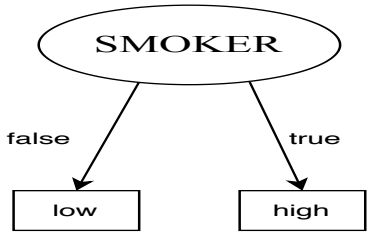


Figure 2: Decision Tree for Boostrap Sample B

The entropy calculation for Bootstrap Sample C is:

$$H(\text{RISK, Boostrap Sample C}) = -\sum_{l \in \begin{cases} low, \\ high \end{cases}} P(\text{RISK} = l) \times \log_2(P(\text{RISK} = l))$$

$$= -\left(\left(\frac{2}{5} \times \log_2\left(\frac{2}{5}\right)\right) + \left(\frac{3}{5} \times \log_2\left(\frac{3}{5}\right)\right)\right) = 0.9710 \text{ bits}$$

The information gain for each of the features in Bootstrap Sample C is as follows:

Split by Feature	Level	Instances	Partition Entropy	Rem.	Info. Gain
OBESE	true	$\mathbf{d}_4, \mathbf{d}_5$	0	0.5510	0.4200
	false	$\mathbf{d}_1, \mathbf{d}_1, \mathbf{d}_2$	0.9183		
FAMILY	true	$\mathbf{d}_1, \mathbf{d}_1, \mathbf{d}_2, \mathbf{d}_4$	1.0	0.8	0.1709
	false	$\mathbf{d}_5$	0		

These calculations show that the OBESE feature has the highest information gain of the descriptive features in Bootstrap Sample C and should be added as the root node of the decision tree generated from Bootstrap Sample C. Splitting Bootstrap Sample C creates one pure partition for OBESE = true ($\mathbf{d}_4$, $\mathbf{d}_5$) where all the instances have RISK = high, and an impure partition for OBESE = false where two instances ($\mathbf{d}_1$, $\mathbf{d}_1$) have RISK = low and for one instance ($\mathbf{d}_2$) RISK = high. Normally this would mean that we would continue to split the impure partition to create pure sets. However, in this

instance there is only one feature that we can still use to split this partition, the FAMILY feature, and all the instances in this partition have the same level for this feature FAMILY = yes. Consequently, instead of splitting this partition further we simply create a leaf node with the majority target level within the partition: RISK =low.So, the final tree generated for Bootstrap Sample C: if OBESE Level is **true**, the RISK is **high**, if OBESE Level is **false**, the RISK is **low**.

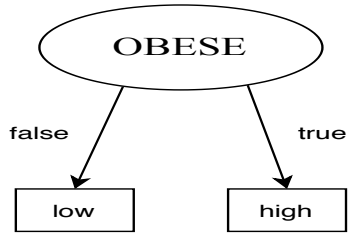


Figure 3: Decision Tree for Boostrap Sample C

(b) Each of the trees in the ensemble will vote as follows:

- Tree 1: EXERCISE=rarely → RISK=high

- Tree 2: SMOKER=false → RISK=low

- Tree 3: OBESE=true → RISK=high

So, the majority vote is for RISK=high, and this is the prediction the model will return for this query.

Exercise sheet: Linear regression

Solutions prepared by Magnus Ross and Mauricio A Álvarez

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.

Answer:
From the lecture slides, we know that the inner product of two vectors $\mathbf{x}$ and $\mathbf{y}$ of dimension m is given by

$$\mathbf{x}^\top \mathbf{y} = \sum_{i=1}^{m} x_i y_i.$$

For the vectors given in the question then, we get that

$$\begin{aligned} \mathbf{x}^\top \mathbf{y} &= \begin{bmatrix} 1.3 & -2.0 & 4.1 \end{bmatrix} \begin{bmatrix} 0.4 \\ -0.8 \\ -1.1 \end{bmatrix} \\ &= (1.3 \times 0.4) + (-2.0 \times -0.8) + (4.1 \times -1.1) \\ &= -2.39 \end{aligned}$$

We can also think of an inner product as a matrix product between a $1 \times m$ matrix and a $m \times 1$ matrix. From the lecture slides, the outer product of two vectors $\mathbf{x}$ and $\mathbf{y}$ of dimension p is given by

$$\mathbf{x}\mathbf{y}^\top = \begin{bmatrix} x_1 y_1 & \cdots & x_1 y_p \\ x_2 y_1 & \cdots & x_2 y_p \\ \vdots & \vdots & \vdots \\ x_m y_1 & \cdots & x_m y_p \end{bmatrix}$$

Note that in the case of the outer product, the dimension of the two vectors need not be the same, whereas in the case of the inner product they must be. For the vectors in the question we get,

$$\begin{aligned} \mathbf{x}\mathbf{y}^\top &= \begin{bmatrix} 1.3 \\ -2.0 \\ 4.1 \end{bmatrix} \begin{bmatrix} 0.4 & -0.8 & -1.1 \end{bmatrix} \\ &= \begin{bmatrix} 1.3 \times 0.4 & 1.3 \times -0.8 & 1.3 \times -1.1 \\ -2.0 \times 0.4 & -2.0 \times -0.8 & -2.0 \times -1.1 \\ 4.1 \times 0.4 & 4.1 \times -0.8 & 4.1 \times -1.1 \end{bmatrix} \\ &= \begin{bmatrix} 0.52 & -1.04 & -1.43 \\ -0.8 & 1.6 & 2.2 \\ 1.64 & -3.28 & -4.5 \end{bmatrix} \end{aligned}$$

We can also think of an outer product as a matrix product between a $m \times 1$ matrix and a $1 \times p$ matrix. The following code snippet computes both the inner and outer product using NumPy:

```
import numpy as np

x = np.array([1.3, -2.0, 4.1])
y = np.array([0.4, -0.8, -1.1])

print(f"Inner product: {np.dot(x, y):.2f}")
print(f"Outer product: {np.outer(x, y)}")
```

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].

Answer:

For the first term, the sum $\sum_{j=1}^{m} w_{i,j}x_j$ can be written as $\mathbf{Wx}$. To obtain $\sum_{i=1}^{n}\sum_{j=1}^{m} w_{i,j}x_j$, we premultiply $\mathbf{Wx}$ by $\mathbf{1}_n^\top$ leading to $\mathbf{1}_n^\top\mathbf{Wx}$. For the second term, the sum $\sum_{i=1}^{n} y_i w_{i,j}$ can be expressed as $\mathbf{y}^\top\mathbf{W}$. To obtain $\sum_{j=1}^{m}\sum_{i=1}^{n} y_i w_{i,j}$, we postmultiply by $\mathbf{1}_m$ leading to $\mathbf{y}^\top\mathbf{W}\mathbf{1}_m$. We can finally write

$$\sum_{i=1}^{n}\sum_{j=1}^{m} w_{i,j}x_j + \sum_{j=1}^{m}\sum_{i=1}^{n} y_i w_{i,j} = \mathbf{1}_n^\top\mathbf{Wx} + \mathbf{y}^\top\mathbf{W}\mathbf{1}_m.$$

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

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

Answer:

In our model, we assume that $\sigma \neq 0$. Based on the lecture notes, we know that that the log likelihood function is given by,

$$LL\left(\mathbf{w}, \sigma^2\right) = -\frac{N}{2}\log(2\pi) - \frac{N}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\mathbf{w})^\top(\mathbf{y} - \mathbf{X}\mathbf{w}),$$

which we need to maximise with respect to σ_*^2 As we have seen, we can find the optimal $\mathbf{w}$ that maximises $LL\left(\mathbf{w}, \sigma^2\right)$ by taking the gradient $\frac{dLL\left(\mathbf{w},\sigma^2\right)}{d\mathbf{w}}$, equating to zero. Similarly, we can find the optimal σ that maximises $LL\left(\mathbf{w}, \sigma^2\right)$ by taking the gradient $\frac{dLL\left(\mathbf{w},\sigma^2\right)}{d\sigma}$, equating to zero and then solving the resulting equation for σ_* (if we get the optimal value for σ_*, we can easily get the optimal value for σ_*^2).

Taking the gradient of each term in $L\left(\mathbf{w}, \sigma^2\right)$ wrt σ, we get

$$\frac{d}{d\sigma}\left[-\frac{N}{2}\log(2\pi)\right] = 0$$

$$\frac{d}{d\sigma}\left[-\frac{N}{2}\log\sigma^2\right] = -\frac{N}{2}\frac{d}{d\sigma}\left[\log\sigma^2\right] = -\frac{N}{2}\frac{d}{d\sigma}\left[2\log\sigma\right] = -\frac{N}{2}2\frac{d}{d\sigma}\left[\log\sigma\right] = -N\frac{d}{d\sigma}\left[\log\sigma\right] = -\frac{N}{\sigma}$$

$$\begin{aligned}
\frac{d}{d\sigma}\left[-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\mathbf{w})^\top(\mathbf{y}-\mathbf{X}\mathbf{w})\right] &= -(\mathbf{y}-\mathbf{X}\mathbf{w})^\top(\mathbf{y}-\mathbf{X}\mathbf{w})\frac{d}{d\sigma}\left[\frac{1}{2\sigma^2}\right] \\
&= -(\mathbf{y}-\mathbf{X}\mathbf{w})^\top(\mathbf{y}-\mathbf{X}\mathbf{w})\frac{1}{2}\frac{d}{d\sigma}\left[\frac{1}{\sigma^2}\right] \\
&= -(\mathbf{y}-\mathbf{X}\mathbf{w})^\top(\mathbf{y}-\mathbf{X}\mathbf{w})\frac{1}{2}\frac{d}{d\sigma}\left[\sigma^{-2}\right] \\
&= -(\mathbf{y}-\mathbf{X}\mathbf{w})^\top(\mathbf{y}-\mathbf{X}\mathbf{w})\frac{1}{2}\left[(-2)\sigma^{-3}\right] \\
&= (\mathbf{y}-\mathbf{X}\mathbf{w})^\top(\mathbf{y}-\mathbf{X}\mathbf{w})\left[\sigma^{-3}\right] \\
&= (\mathbf{y}-\mathbf{X}\mathbf{w})^\top(\mathbf{y}-\mathbf{X}\mathbf{w})\frac{1}{\sigma^3}
\end{aligned}$$

Putting these terms together, we get

$$\frac{d}{d\sigma}LL\left(\mathbf{w}, \sigma^2\right) = 0 - \frac{N}{\sigma} + (\mathbf{y}-\mathbf{X}\mathbf{w})^\top(\mathbf{y}-\mathbf{X}\mathbf{w})\frac{1}{\sigma^3}$$

Now, equating to zero and solving for σ^2, we get

$$\begin{aligned}
0 - \frac{N}{\sigma} + (\mathbf{y}-\mathbf{X}\mathbf{w})^\top(\mathbf{y}-\mathbf{X}\mathbf{w})\frac{1}{\sigma^3} &= 0 \\
-\frac{N}{\sigma} + (\mathbf{y}-\mathbf{X}\mathbf{w})^\top(\mathbf{y}-\mathbf{X}\mathbf{w})\frac{1}{\sigma^3} &= 0 \\
(\mathbf{y}-\mathbf{X}\mathbf{w})^\top(\mathbf{y}-\mathbf{X}\mathbf{w})\frac{1}{\sigma^3} &= \frac{N}{\sigma} \\
(\mathbf{y}-\mathbf{X}\mathbf{w})^\top(\mathbf{y}-\mathbf{X}\mathbf{w}) &= N\sigma^2 \quad \text{(We assume: } \sigma \neq 0\text{)} \\
(\mathbf{y}-\mathbf{X}\mathbf{w})^\top(\mathbf{y}-\mathbf{X}\mathbf{w})\frac{1}{N} &= \sigma^2 \\
\sigma^2 &= \frac{1}{N}(\mathbf{y}-\mathbf{X}\mathbf{w})^\top(\mathbf{y}-\mathbf{X}\mathbf{w})
\end{aligned}$$

We already have $\mathbf{w}_*$, the optimal value for $\mathbf{w}$, so we just plug this in to obtain,

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

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 $\lambda = 0.1$.

Answer:
From the Lectures we have that the optimum value for $\mathbf{w}$ in ridge regression is given by,

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

recall that $\mathbf{I}$ here is the identity matrix, and N is the number of data, in this case 3. We need to compute this equation for the dataset given in the question. We can write the dataset as,

$$\mathbf{X} = \begin{bmatrix} 1 & 0.8 \\ 1 & -0.3 \\ 1 & 0.1 \end{bmatrix}, \qquad \mathbf{y} = \begin{bmatrix} -1.2 \\ -0.6 \\ 2.4 \end{bmatrix}.$$

Where we add 1's to the first column to account for the intercept in the regression model. We need to compute the formula for this data. First let's compute some of the necessary terms, recalling that matrix multiplication is computed via

$$\begin{aligned} &\mathbf{AB} = \mathbf{C}, \\ &C_{i,j} = \sum_k A_{ik} B_{kj}, \end{aligned}$$

so,

$$\begin{aligned} \mathbf{X}^\top \mathbf{X} + \frac{\lambda N}{2} &= \left(\begin{bmatrix} 3 & 0.6 \\ 0.6 & 0.74 \end{bmatrix} + \frac{0.1 \times 3}{2} \times \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right) \\ &= \begin{bmatrix} 3.15 & 0.6 \\ 0.6 & 0.89 \end{bmatrix}, \\ \mathbf{X}^\top \mathbf{y} &= \begin{bmatrix} 0.6 \\ -0.54 \end{bmatrix}. \end{aligned}$$

Next we will need the formula to invert a 2×2 matrix, which is

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}.$$

Finally we can obtain $\mathbf{w}^*$,

$$\mathbf{w}^* = \begin{bmatrix} 0.364 & -0.246 \\ -0.246 & 1.289 \end{bmatrix} \begin{bmatrix} 0.6 \\ -0.54 \end{bmatrix} = \begin{bmatrix} 0.351 \\ -0.843 \end{bmatrix} \tag{1}$$

This question can also be solved using a maths package like NumPy (or even Matlab), the following snippet gives the solution using NumPy, the code uses `np.eye(p)` for representing $\mathbf{I}_p$ and `np.linalg.solve(A, b)` for solving the linear system $\mathbf{Ax} = \mathbf{b}$. See also the Lab Notebook for Week 4.

```python
import numpy as np

X = np.array([[1.0, 1.0, 1.0], [0.8, -0.3, 0.1]]).T
y = np.array([-1.2, -0.6, 2.4])

N = 3
l = 0.1

A = X.T @ X + ((N * l)/2) * np.eye(2)
b = X.T@y

# Solve Ax = b equivient to x = A^{-1} b
w_star = np.linalg.solve(A, b)
print(f"Optimum w: {w_star}")
```

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 $N \times N$].
Answer:
We start by writing the sum as

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

Using the HINT given above, each term of the sum can be expressed as

$$\sum_{n=1}^{N} r_n y_n^2 = \mathbf{y}^\top \mathbf{R}\mathbf{y}$$

$$-2\sum_{n=1}^{N} r_n y_n \mathbf{w}^\top \mathbf{x}_n = -2\mathbf{w}^\top \sum_{n=1}^{N} r_n y_n \mathbf{x}_n = -2\mathbf{w}^\top \mathbf{X}^\top \mathbf{R}\mathbf{y}$$

$$\sum_{n=1}^{N} r_n \left(\mathbf{w}^\top \mathbf{x}_n\right)^2 = \sum_{n=1}^{N} r_n \mathbf{w}^\top \mathbf{x}_n \mathbf{x}_n^\top \mathbf{w} = \mathbf{w}^\top \left(\sum_{n=1}^{N} r_n \mathbf{x}_n \mathbf{x}_n^\top\right) \mathbf{w} = \mathbf{w}^\top \mathbf{X}^\top \mathbf{R}\mathbf{X}\mathbf{w},$$

where $\mathbf{y} = [y_1, \ldots, y_N]^\top$ and $\mathbf{X}$ is a *design matrix*. Putting these terms together in the expression for the mean of weigthed squared errors, we get

$$\begin{aligned} E(\mathbf{w}) = \frac{1}{N}\sum_{n=1}^{N} r_n(y_n - \mathbf{w}^\top \mathbf{x}_n)^2 &= \frac{1}{N}\left[\mathbf{y}^\top\mathbf{R}\mathbf{y} - 2\mathbf{w}^\top\mathbf{X}^\top\mathbf{R}\mathbf{y} + \mathbf{w}^\top\mathbf{X}^\top\mathbf{R}\mathbf{X}\mathbf{w}\right] \\ &= \frac{1}{N}\left[\mathbf{y}^\top\mathbf{R}\mathbf{y} - \mathbf{y}^\top\mathbf{R}\mathbf{X}\mathbf{w} - \mathbf{w}^\top\mathbf{X}^\top\mathbf{R}\mathbf{y} + \mathbf{w}^\top\mathbf{X}^\top\mathbf{R}\mathbf{X}\mathbf{w}\right] \\ &= \frac{1}{N}\left[\mathbf{y}^\top\mathbf{R}\left(\mathbf{y} - \mathbf{X}\mathbf{w}\right) - \mathbf{w}^\top\mathbf{X}^\top\mathbf{R}\left(\mathbf{y} - \mathbf{X}\mathbf{w}\right)\right] \\ &= \frac{1}{N}\left[\left(\mathbf{y}^\top\mathbf{R} - \mathbf{w}^\top\mathbf{X}^\top\mathbf{R}\right)\left(\mathbf{y} - \mathbf{X}\mathbf{w}\right)\right] \\ &= \frac{1}{N}\left(\mathbf{y} - \mathbf{X}\mathbf{w}\right)^\top\mathbf{R}\left(\mathbf{y} - \mathbf{X}\mathbf{w}\right). \end{aligned}$$

(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.
Answer:
We start with the mean of weighted squared errors in matrix form as

$$E(\mathbf{w}) = \frac{1}{N}\left(\mathbf{y}^\top\mathbf{R}\mathbf{y} - 2\mathbf{w}^\top\mathbf{X}^\top\mathbf{R}\mathbf{y} + \mathbf{w}^\top\mathbf{X}^\top\mathbf{R}\mathbf{X}\mathbf{w}\right).$$

The two main results that we need are

$$\frac{d\mathbf{w}^\top\mathbf{a}}{d\mathbf{w}} = \mathbf{a}, \qquad \frac{d\mathbf{w}^\top\mathbf{A}\mathbf{w}}{d\mathbf{w}} = 2\mathbf{A}\mathbf{w}.$$

The derivative of $E(\mathbf{w})$ wrt $\mathbf{w}$ is then given as

$$\frac{dE(\mathbf{w})}{d\mathbf{w}} = -\frac{2}{N}\mathbf{X}^\top\mathbf{R}\mathbf{y} + \frac{2}{N}\mathbf{X}^\top\mathbf{R}\mathbf{X}\mathbf{w}.$$

Making the expression above equal to zero, we get

$$\mathbf{X}^\top\mathbf{R}\mathbf{X}\mathbf{w} = \mathbf{X}^\top\mathbf{R}\mathbf{y}.$$

The optimal value $\mathbf{w}^*$ is then given as

$$\mathbf{w}^* = \left(\mathbf{X}^\top\mathbf{R}\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{R}\mathbf{y}.$$

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 $x = 2.5$?
Answer:

For basis function regression, we must first compute the value of each basis function at the input location, and as usual add the bias. In this case we have,

$$\boldsymbol{\phi}(\mathbf{x}) = \begin{bmatrix} 1 \\ 2.5^1 \\ 2.5^2 \\ 2.5^3 \\ 2.5^4 \end{bmatrix} = \begin{bmatrix} 1 \\ 2.5 \\ 6.25 \\ 15.625 \\ 39.0625 \end{bmatrix}$$

The prediction can then be computed by taking the inner product with the weight vector giving,

$$\mathbf{w}_*^\top \boldsymbol{\phi}(\mathbf{x}) = \begin{bmatrix} 0.5 & -0.8 & 1.2 & 1.3 & -0.3 \end{bmatrix} \begin{bmatrix} 1. \\ 2.5 \\ 6.25 \\ 15.625 \\ 39.0625 \end{bmatrix} = 14.594$$

Alternatively the following Python code computes the answer:

```python
import numpy as np

w_star = np.array([0.5, -0.8, 1.2, 1.3, -0.3])
phi = np.power(2.5 * np.ones(5), np.arange(5))

print(f"Prediction: {np.dot(w_star, phi)}")
```

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

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

Answer:

Based on the Lecture notes, in ridge regression, we consider the objective function as

$$h(\mathbf{w}) = \frac{1}{N}\sum_{n=1}^{N}\left(y_n - \mathbf{w}^\top \mathbf{x}_n\right)^2 + \frac{\lambda}{2}\mathbf{w}^\top \mathbf{w}$$

Using what we reviewed in the section on vector/matrix notation, it can be shown that this expression can be written in a vectorial form as

$$h(\mathbf{w}) = \frac{1}{N}(\mathbf{y} - \mathbf{X}\mathbf{w})^\top (\mathbf{y} - \mathbf{X}\mathbf{w}) + \frac{\lambda}{2}\mathbf{w}^\top \mathbf{w}$$

The term $\frac{1}{N}(\mathbf{y} - \mathbf{X}\mathbf{w})^\top (\mathbf{y} - \mathbf{X}\mathbf{w})$ can be expressed as

$$\frac{1}{N}(\mathbf{y} - \mathbf{X}\mathbf{w})^\top (\mathbf{y} - \mathbf{X}\mathbf{w}) = \frac{1}{N}\mathbf{y}^\top \mathbf{y} - \frac{1}{N}\mathbf{w}^\top \mathbf{X}^\top \mathbf{y} - \frac{1}{N}\mathbf{y}^\top \mathbf{X}\mathbf{w} + \frac{1}{N}\mathbf{w}^\top \mathbf{X}^\top \mathbf{X}\mathbf{w}$$

We can find the $\mathbf{w}$ that maximises $h(\mathbf{w})$ by taking the gradient $\frac{dh(\mathbf{w})}{d\mathbf{w}}$, equating to zero and solving for $\mathbf{w}$. Taking the gradient of each term in $h(\mathbf{w})$ wrt $\mathbf{w}$, we get

$$\begin{aligned}
\frac{d}{d\mathbf{w}}\left[\frac{1}{N}\mathbf{y}^\top\mathbf{y}\right] &= 0 \\
\frac{d}{d\mathbf{w}}\left[-\frac{1}{N}\mathbf{w}^\top\mathbf{X}^\top\mathbf{y}\right] &= -\frac{1}{N}\mathbf{X}^\top\mathbf{y} \\
\frac{d}{d\mathbf{w}}\left[-\frac{1}{N}\mathbf{y}^\top\mathbf{X}\mathbf{w}\right] &= -\frac{1}{N}\mathbf{X}^\top\mathbf{y} \\
\frac{d}{d\mathbf{w}}\left[\frac{1}{N}\mathbf{w}^\top\mathbf{X}^\top\mathbf{X}\mathbf{w}\right] &= \frac{2}{N}\mathbf{X}^\top\mathbf{X}\mathbf{w} \\
\frac{d}{d\mathbf{w}}\left[\frac{\lambda}{2}\mathbf{w}^\top\mathbf{w}\right] &= \frac{\lambda}{2}2\mathbf{w} = \lambda\mathbf{w}
\end{aligned}$$

Putting these terms together, we get

$$\begin{aligned}
\frac{d}{d\mathbf{w}}h\left(\mathbf{w}\right) &= 0 - \frac{1}{N}\mathbf{X}^\top\mathbf{y} - \frac{1}{N}\mathbf{X}^\top\mathbf{y} + \frac{2}{N}\mathbf{X}^\top\mathbf{X}\mathbf{w} + \lambda\mathbf{w} \\
&= -\frac{2}{N}\mathbf{X}^\top\mathbf{y} + \frac{2}{N}\mathbf{X}^\top\mathbf{X}\mathbf{w} + \lambda\mathbf{w} \\
&= -\frac{2}{N}\mathbf{X}^\top\mathbf{y} + \left(\frac{2}{N}\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}\right)\mathbf{w},
\end{aligned}$$

where $\mathbf{I}$ is an identity matrix of the same dimensions that $\mathbf{w}$. Now, equating to zero and solving for $\mathbf{w}$, we get

$$\begin{aligned}
-\frac{2}{N}\mathbf{X}^\top\mathbf{y} + \left(\frac{2}{N}\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}\right)\mathbf{w} &= 0 \\
\left(\frac{2}{N}\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}\right)\mathbf{w} &= \frac{2}{N}\mathbf{X}^\top\mathbf{y} \\
\frac{N}{2}\left(\frac{2}{N}\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}\right)\mathbf{w} &= \frac{N}{2}\frac{2}{N}\mathbf{X}^\top\mathbf{y} \\
\left(\mathbf{X}^\top\mathbf{X} + \frac{\lambda N}{2}\mathbf{I}\right)\mathbf{w} &= \mathbf{X}^\top\mathbf{y} \\
\mathbf{w} &= \left(\mathbf{X}^\top\mathbf{X} + \frac{\lambda N}{2}\mathbf{I}\right)^{-1}\mathbf{X}^\top\mathbf{y}
\end{aligned}$$

Thus, the optimal solution $\mathbf{w}_*$ in ridge regression 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

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 ($x_1 = 3, x_2 = 5, x_3 = 1$)

   Answer The Jacobian $\mathbf{J}$ has dimensions 2×3 and is given as

   $$\mathbf{J} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \frac{\partial f_1}{\partial x_3} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \frac{\partial f_2}{\partial x_3} \end{bmatrix},$$

   where the derivatives in the first row are given as

   $$\frac{df_1}{dx_1} = x_3 + e^{-x_3} \frac{1}{(x_2 + x_1)},$$
   $$\frac{df_1}{dx_2} = e^{-x_3} \frac{1}{x_2 + x_1},$$
   $$\frac{df_1}{dx_3} = x_1 - \log(x_2 + x_1) e^{-x_3},$$

   and the derivatives in the second row are given as

   $$\frac{df_2}{dx_1} = -x_3 \sin(x_1 x_3)$$
   $$\frac{df_2}{dx_2} = -e^{-x_2}$$
   $$\frac{df_2}{dx_3} = -x_1 \sin(x_1 x_3).$$

   The Jacobian at the point ($x_1 = 3, x_2 = 5, x_3 = 1$) is given as

   $$\mathbf{J}(x_1 = 3, x_2 = 5, x_3 = 1) = \begin{bmatrix} 1.0459849301 & 0.0459849301 & 2.2350162077 \\ -0.1411200081 & -0.006737947 & -0.4233600242 \end{bmatrix}.$$

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

   Answer We can compute the finite difference approximations for the partial derivatives using the Python code shown in following snippet:

```python
import numpy as np

def f1(x1, x2, x3):
  return x1*x3 + np.log(x2+x1)*np.exp(-x3)

def f2(x1, x2, x3):
  return np.exp(-x2) + np.cos(x1*x3)

x1_0 = 3
x2_0 = 5
x3_0 = 1
epsilon = 1e-6

df1dx1_numerical = (f1(x1_0+epsilon, x2_0, x3_0) - f1(x1_0, x2_0, x3_0))/epsilon
df1dx2_numerical = (f1(x1_0, x2_0+epsilon, x3_0) - f1(x1_0, x2_0, x3_0))/epsilon
df1dx3_numerical = (f1(x1_0, x2_0, x3_0+epsilon) - f1(x1_0, x2_0, x3_0))/epsilon

df2dx1_numerical = (f2(x1_0+epsilon, x2_0, x3_0) - f2(x1_0, x2_0, x3_0))/epsilon
df2dx2_numerical = (f2(x1_0, x2_0+epsilon, x3_0) - f2(x1_0, x2_0, x3_0))/epsilon
df2dx3_numerical = (f2(x1_0, x2_0, x3_0+epsilon) - f2(x1_0, x2_0, x3_0))/epsilon
```

The Jacobian computed using finite differences is approximated as

$$\mathbf{J}(x_1 = 3, x_2 = 5, x_3 = 1) \approx \begin{bmatrix} 1.0459849276 & 0.0459849274 & 2.2350165896 \\ -0.1411195131 & -0.0067379436 & -0.4233555692 \end{bmatrix}.$$

3. (**) Draw the computational graph.

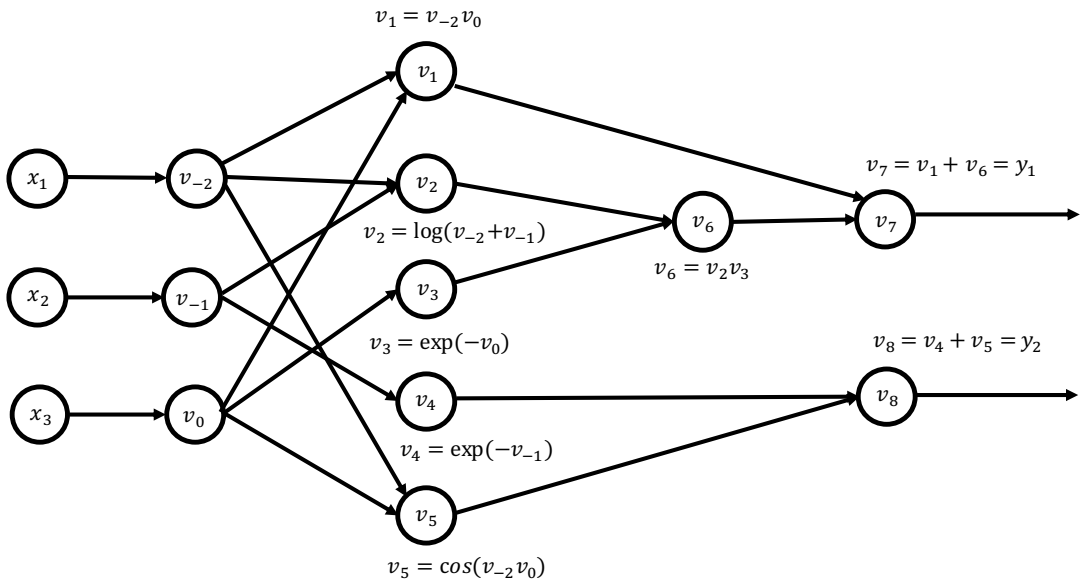


Figure 1: Computational graph for the vector-valued function

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.

Forward primal trace	Forward tangent trace
$v_{-2} = x_1$	$\dot{v}_{-1} = \dot{x}_1$
$v_{-1} = x_2$	$\dot{v}_{-1} = \dot{x}_2$
$v_0 = x_3$	$\dot{v}_0 = \dot{x}_3$
$v_1 = v_{-2}v_0$	$\dot{v}_1 = v_0$
$v_2 = \log(v_{-2} + v_{-1})$	$\dot{v}_2 = 1/(v_{-2} + v_{-1})$
$v_3 = \exp(-v_0)$	$\dot{v}_3 = 0$
$v_4 = \exp(-v_{-1})$	$\dot{v}_4 = 0$
$v_5 = \cos(v_{-2}v_0)$	$\dot{v}_5 = -v_0 \sin(v_{-2}v_0)$
$v_6 = v_2v_3$	$\dot{v}_6 = v_3\dot{v}_2$
$v_7 = v_1 + v_6$	$\dot{v}_7 = \dot{v}_1 + \dot{v}_6$
$v_8 = v_4 + v_5$	$\dot{v}_8 = \dot{v}_5$
$y_1 = v_7$	$\dot{y}_1 = \dot{v}_7$
$y_2 = v_8$	$\dot{y}_2 = \dot{v}_8$

Table 1: Forward tanget trace for $\frac{dy_1}{dx_1}$ and $\frac{dy_2}{dx_1}$

Forward primal trace		Forward tangent trace	
$v_{-2} = x_1$	$= 3$	$\dot{v}_{-1} = \dot{x}_1$	$= 1$
$v_{-1} = x_2$	$= 5$	$\dot{v}_{-1} = \dot{x}_2$	$= 0$
$v_0 = x_3$	$= 1$	$\dot{v}_0 = \dot{x}_3$	$= 0$
$v_1 = v_{-2}v_0$	$= 3$	$\dot{v}_1 = v_0$	$= 1$
$v_2 = \log(v_{-2} + v_{-1})$	$= 2.079$	$\dot{v}_2 = 1/(v_{-2} + v_{-1})$	$= 0.125$
$v_3 = \exp(-v_0)$	$= 0.367$	$\dot{v}_3 = 0$	
$v_4 = \exp(-v_{-1})$	$= 0.006$	$\dot{v}_4 = 0$	
$v_5 = \cos(v_{-2}v_0)$	$= -0.989$	$\dot{v}_5 = -v_0 \sin(v_{-2}v_0)$	$= -0.141$
$v_6 = v_2v_3$	$= 0.764$	$\dot{v}_6 = v_3\dot{v}_2$	$= 0.045$
$v_7 = v_1 + v_6$	$= 3.764$	$\dot{v}_7 = \dot{v}_1 + \dot{v}_6$	$= 1.045$
$v_8 = v_4 + v_5$	$= -0.983$	$\dot{v}_8 = \dot{v}_5$	$= -0.141$
$y_1 = v_7$	$= 3.764$	$\dot{y}_1 = \dot{v}_7$	$= 1.045$
$y_2 = v_8$	$= -0.983$	$\dot{y}_2 = \dot{v}_8$	$= -0.141$

Table 2: Derivatives for $\frac{dy_1}{dx_1}$ and $\frac{dy_2}{dx_1}$, at $(x_1 = 3, x_2 = 5, x_3 = 1)$

Answer Let us compute the forward tangent trace for $\frac{dy_1}{dx_1}$ and $\frac{dy_2}{dx_1}$. Table 1 shows the forward primal trace and the forward tangent trace.

We use table 1 to compute the derivatives $\frac{dy_1}{dx_1}$ and $\frac{dy_2}{dx_1}$ at $(x_1 = 3, x_2 = 5, x_3 = 1)$. Notice how table 2 provides the first column of the Jacobian.

The other two columns of the Jacobian can be computed using a similar procedure. This is left to the student to complete.

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

Answer. Let us compute the partial derivatives $\frac{\partial y_1}{\partial x_1}$, $\frac{\partial y_1}{\partial x_2}$ and $\frac{\partial y_1}{\partial x_3}$ using the reverse mode. The

computation of $\frac{\partial y_2}{\partial x_1}$, $\frac{\partial y_2}{\partial x_2}$ and $\frac{\partial y_2}{\partial x_3}$ is left to the student.

The adjoint $\bar{y}_1$ is simply $\bar{y}_1 = 1$.

Looking at the computational graph, we now compute $\bar{v}_7 = \frac{\partial y_1}{\partial v_7} = 1$.

The adjoints we need to compute are then

$$\begin{aligned}
\bar{v}_6 &= \bar{v}_7 \frac{\partial v_7}{\partial v_6} = \bar{v}_7(1) = 1 \\
\bar{v}_1 &= \bar{v}_7 \frac{\partial v_7}{\partial v_1} = \bar{v}_7(1) = 1 \\
\bar{v}_2 &= \bar{v}_6 \frac{\partial v_6}{\partial v_2} = \bar{v}_6 v_3 = (1)(0.367) = 0.367 \\
\bar{v}_3 &= \bar{v}_6 \frac{\partial v_6}{\partial v_3} = \bar{v}_6 v_2 = (1)(2.079) = 2.079 \\
\bar{v}_{-2} &= \bar{v}_1 \frac{v_1}{v_{-2}} + \bar{v}_2 \frac{v_2}{v_{-2}} = \bar{v}_1 v_0 + \bar{v}_2 \frac{1}{v_{-2} + v_{-1}} = (1)(1) + (0.367)/(8) = 1.0458 \\
\bar{v}_{-1} &= \bar{v}_2 \frac{v_2}{v_{-1}} = \bar{v}_2 \frac{1}{v_{-2} + v_{-1}} = (0.367)/(8) = 0.0459 \\
\bar{v}_0 &= \bar{v}_1 \frac{v_1}{v_0} + \bar{v}_3 \frac{v_3}{v_0} = \bar{v}_1 v_{-2} + \bar{v}_3(-\exp(-v_0)) = (1)(3) - (2.079)\exp(-1) = 2.235.
\end{aligned}$$

Finally, we get

$$\begin{aligned}
\bar{x}_1 &= \bar{v}_{-2} \frac{\partial v_{-2}}{\partial x_1} = \bar{v}_{-2} = 1.0458 \\
\bar{x}_2 &= \bar{v}_{-1} \frac{\partial v_{-1}}{\partial x_2} = \bar{v}_{-1} = 0.0459 \\
\bar{x}_3 &= \bar{v}_0 \frac{\partial v_0}{\partial x_3} = \bar{v}_0 = 0.0459
\end{aligned}$$

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.

Solution:

Conversion to probability table [2 marks].

		COVID Positive	
Age Group	# in Group	#	%
0-9	1	1	100
10-19	3	2	67
20-29	3	2	67
30-39	4	1	25
40-49	7	4	57
50-59	7	5	71
60-69	5	3	60
70-79	2	2	100
80-89	1	1	100

Compute the odds [1 mark].

Using the probability computed in the table above for the 20-29 age group of $\frac{2}{3}$, we can compute the odds as follows:

$$\begin{aligned} \text{Odds} &= \frac{\pi}{1-\pi} \\ &= \frac{\frac{2}{3}}{1-\frac{2}{3}} \\ &= 2 \end{aligned} \tag{1}$$

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.

Solution:

starting at the expression of the logit function,

$$log\left(\frac{\pi}{1-\pi}\right) = \mathbf{w^T x} \tag{2}$$

taking the exponent of both sides

$$\frac{\pi}{1-\pi} = e^{\mathbf{w^T x}} \tag{3}$$

also,

$$\begin{aligned} \frac{1-\pi}{\pi} &= e^{-\mathbf{w^T x}} \\ \frac{1}{\pi} - 1 &= e^{-\mathbf{w^T x}} \\ \frac{1}{\pi} &= 1 + e^{-\mathbf{w^T x}} \\ \pi &= \frac{1}{1+e^{-\mathbf{w^T x}}} \end{aligned} \tag{4}$$

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 $\mathbf{x}$.

Solution:

$$\log P(y \mid X) = \sum_{i=1}^{n} \log P(y_i \mid x_i) \tag{5}$$

$$= \sum_{i=1}^{n} y_i \log\left(\frac{1}{1+e^{-\mathbf{w^T x_i}}}\right) + \sum_{i=1}^{n}(1-y_i)\log\left(1 - \frac{1}{1+e^{-\mathbf{w^T x_i}}}\right) \tag{6}$$

$$= \sum_{i=1}^{n} y_i \log\left(\frac{1}{1+e^{-\mathbf{w^T x_i}}}\right) + \sum_{i=1}^{n}(1-y_i)\log\left(\frac{1+e^{-\mathbf{w^T x_i}}-1}{1+e^{-\mathbf{w^T x_i}}}\right) \tag{7}$$

$$\log P(y \mid X) = \sum_{i=1}^{n} y_i \log\left(\frac{1}{1+e^{-\mathbf{w^T x_i}}}\right) + \sum_{i=1}^{n}(1-y_i)\log\left(\frac{1}{e^{\mathbf{w^T x}}+1}\right) \tag{8}$$

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?

Solution:

We can compute this probability using the following expression,

$$P(y=1|x) = \frac{1}{1+e^{-\mathbf{w^T x}}}. \tag{9}$$

This first requires computation of $\mathbf{w^T x}$, which in this case is,

$$\begin{aligned}\mathbf{w^T x} &= (4\times 6) + (-2\times 8) + (5\times 2) + (-3\times 7) + (11\times -3) + (9\times 5)\\ &= 9.\end{aligned} \tag{10}$$

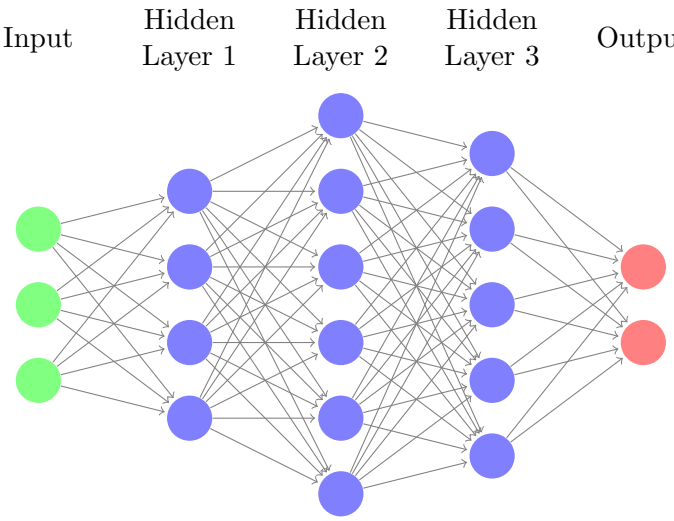
We can then substitute this value into (9) to obtain the final answer (given to 5 significant figures),

$$\begin{aligned}P(y=1|x) &= \frac{1}{1+e^{-9}}\\ &= 0.99988.\end{aligned} \tag{11}$$

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?

Solution:

The neural network described in the question has the following structure:



Firstly we must count all of the weights which connect the layers of our model,

$$\begin{aligned} \text{Number of weights} &= (3 \times 4) + (4 \times 6) + (6 \times 5) + (5 \times 2) \\ &= 76. \end{aligned} \tag{12}$$

Next, we count up all of the bias parameters,

$$\begin{aligned} \text{Number of biases} &= 4 + 6 + 5 + 2 \\ &= 17. \end{aligned} \tag{13}$$

The sum of these two values is the total number of model parameters, therefore the answer is $76 + 17 = 93$.

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.

Solution:
$\boldsymbol{h} = g((W^{(1)})^T)\boldsymbol{x} + b^{(1)})$
$\boldsymbol{o} = g((W^{(2)})^T)\boldsymbol{h} + b^{(2)})$
$\boldsymbol{o} = g((W^{(2)})^T)g((W^{(1)})^T)\boldsymbol{x} + b^{(1)}) + b^{(2)})$
g is defined as $g(a) = a$
$\boldsymbol{o} = (W^{(2)})^T((W^{(1)})^T)\boldsymbol{x} + b^{(1)}) + b^{(2)})$
$\boldsymbol{o} = (W^{(2)})^T(W^{(1)})^T\boldsymbol{x} + (W^{(2)})^T b^{(1)} + b^{(2)}$
Substitute $W = (W^{(2)})^T(W^{(1)})^T; b = (W^{(2)})^T b^{(1)} + b^{(2)}$
$\boldsymbol{o} = W\boldsymbol{x} + b$

[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?

Solution:
$$\text{Image} = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 \end{bmatrix}$$
$$\text{Kernel} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$
Apply kernel to image in the first position (the red indicated where the kernel is placed and the kernel values at their respective positions):
$$\text{Image} = \begin{bmatrix} 1_{\times 1} & 1_{\times 0} & 1_{\times 0} & 0 & 0 \\ 0_{\times 0} & 1_{\times 1} & 1_{\times 0} & 1 & 0 \\ 0_{\times 0} & 0_{\times 0} & 1_{\times 1} & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 \end{bmatrix}$$
The Convolved Feature after the first convolution, summing all the products between the image and the kernel values:

First Convolved Feature $= 1 + 0 + 0 + 0 + 1 + 0 + 0 + 0 + 1 = 3$
$$\text{Convolved Features} = \begin{bmatrix} 3 & \\ & \\ & \end{bmatrix}$$

The next step is to shift the kernel and perform the same operation:

$$\text{Image} = \begin{bmatrix} 1 & 1_{\times 1} & 1_{\times 0} & 0_{\times 0} & 0 \\ 0 & 1_{\times 0} & 1_{\times 1} & 1_{\times 0} & 0 \\ 0 & 0_{\times 0} & 1_{\times 0} & 1_{\times 1} & 1 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 \end{bmatrix}$$

Calculate the new convolved feature:

$$\text{Convolved Features} = \begin{bmatrix} 3 & 3 & \\ & & \\ & & \end{bmatrix}$$

Perform the same operation for the entire image, resulting in the final Convolved Features:

$$\text{Convolved Features} = \begin{bmatrix} 3 & 3 & 3 \\ 1 & 3 & 2 \\ 1 & 1 & 2 \end{bmatrix}.$$

The kernel can detect diagonal edges in the image.

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?

Solution:

Size of output:

Size of output: (Image Length - Filter Size + 2× Padding) / Stride + 1

Image Length = 512
Filter Size = 5
Stride = 7
Padding = 2

After applying the first 5×5 filter:

Output Size After First Filter = $(512 - 5 + 2 \times 2)/7 + 1 = 74$

Final Output Shape = Number of Filters × Output Size × Output Size

Final Output Shape = $100 \times 74 \times 74$

Number of parameters:

Number of parameters = (Filter Width × Filter Height × Filters in Previous Layer + 1) × Number of Filters

Filter Width = 5
Filter Height = 5
Filters in Previous Layer = 3
Number of Filters = 100

Number of parameters = $(5 \times 5 \times 3 + 1) \times 100 = 7600$

6 4. For the AlexNet depicted in Slide 35 of Lecture 6, there are about 60 million learnable pa-

rameters. 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.

Solution:

The AlexNet consists of convolutional layers, pooling layers and fully connected layers. The pooling layer does not have any learnable parameters.

The number of parameters in the convolutional layer is:

$$W_c = K^2 \times C \times N, \tag{1}$$

where the K is the size of the kernel, C is the number of channels in the input and N is the number of kernels. In addition to the weights, there are also N bias values. The final number of parameters is $P_c = N + W_c$.

There are also two types of fully connected (FC) layer: the first is where the last pooling layer is connected to a FC layer, and the other is where a FC layer is connected to another FC layer.

The number of parameters in the first case is:

$$W_{fc} = O^2 \times N \times F, \tag{2}$$

where the O is the size of the convolved output, N is the number of kernels in the previous convolutional layer and F is the number of neurons in the layer. The convolved output is flatted to a vector of length $O \times O \times N$. In addition to the weights, there are also F bias values. The total number of parameters in this layer is $P_c = F + W_{fc}$.

In the case where a fc layer is connected to another fc layer:

$$W_{fc} = F_{-1} \times F, \tag{3}$$

where, F_{-1} is the number of neurons in the previous layer and F is the number of neurons in the current layer. The total number of parameters in this layer is $P_c = F + W_{fc}$.

For example in the first layer:

$$P_1 = 11^2 \times 3 \times 96 + 96 = 34944. \tag{4}$$

The second layer:

$$P_2 = 5^2 \times 96 \times 256 + 256 = 614656. \tag{5}$$

After performing the appropriate operations at each layer the total number of parameters in AlexNet is: $62,378,344$.

Parameters in each layer:

- Conv Layer 1: 34944
- Conv Layer 2: 614656

- Conv Layer 3: 885120
- Conv Layer 4: 1327488
- Conv Layer 5: 884992
- FC layer 1: 37752832
- FC layer 2: 16781312
- FC layer 3: 4097000

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?

Solution:

For a 30-bit image, each pixel in the image can be represented by a 30-bit integer. The 30-bit integer can have 2^{30} possible values. Therefore, the total number of distinct images with a size of 1200×1200 (1200×1200 pixels) is calculated as follows.

$$\#\text{Distinct Images} = (2^{30})^{(1200\times1200)} = 2^{30\times1200\times1200} \tag{1}$$

[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 $\hat{\mathbf{x}}_2$. Assume that all data points are centred already.

Solution:

Let $\mathbf{R} = \begin{pmatrix} 0.4729 & -0.8817 \\ -0.8817 & -0.4719 \\ 0 & 0 \end{pmatrix}$, $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2)$. The projection of data points $\mathbf{x}_1$ and $\mathbf{x}_2$ from the 3-dimensional space to the 2-dimensional subspace spanned by the eigenvectors is calculated as follows.

$$\hat{\mathbf{X}} = (\hat{\mathbf{x}_1}, \hat{\mathbf{x}_2}) = \mathbf{R}^\top\mathbf{X} = \begin{pmatrix} 0.4729 & -0.8817 & 0 \\ -0.8817 & -0.4719 & 0 \end{pmatrix} \begin{pmatrix} 2 & 4 \\ 3 & 1 \\ 3 & 0 \end{pmatrix} = \begin{pmatrix} -1.6993 & 1.0099 \\ -3.1791 & -3.9987 \end{pmatrix} \tag{2}$$

"Assume that all data points are centred already." was added while preparing the solution. Centring, i.e. subtracting the mean (from the training data), before projection is a good and common practice.

[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?

Solution:

We define the centre values of cluster 1 and cluster 2 as c_1 and c_2. Let $\mathbf{X} = \{0, 2, 4, 6, 24, 26\}$. There are two steps in each iteration of K-means algorithm, aiming to find:

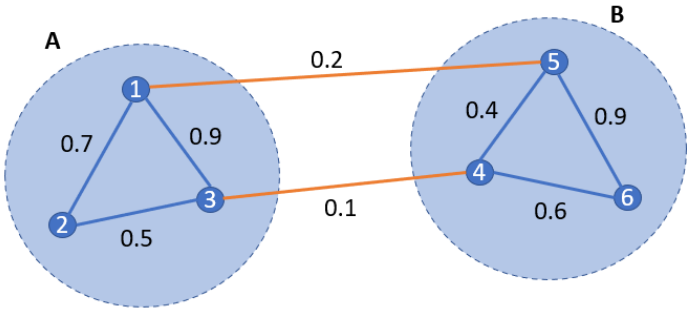
$$\min \sum_{j=1}^{2} \sum_{x^{(i)} \in \mathbf{X} \text{ allocated to } j} \left(x^{(i)} - c_j \right)^2 \tag{3}$$

In the first step we group the data points to a cluster whose centre they are closest to in terms of distance. With an initialisation of $c_1 = 3$ and $c_2 = 4$ in the first iteration, we allocate 0 and 2 to cluster 1 and 4, 6, 24, 26 to cluster 2. In the second step, we set new centres to the clusters found in the first step. In this step, we simply set the centre to the mean value of data points in the same cluster:

$$c_i = \mathbf{E}[X_i] = \frac{1}{|X_i|} \sum_{j}^{|X_i|} x_i^{(j)} \quad , x_i^{(j)} \in X_i, \ i = 1, 2 \tag{4}$$

where $X_1 = \{0, 2\}$, $X_2 = \{4, 6, 24, 26\}$. Therefore the centres of cluster 1 and cluster 2 are updated to $c_1 = 1$ and $c_2 = 15$ in Eqn. (4). In the next iteration, we repeat step 1 and step 2. With the $c_1 = 1$ and $c_2 = 15$ from previous iteration, cluster 1 and cluster 2 will contain data points $X_1 = \{0, 2, 4, 6\}$, $X_2 = \{24, 26\}$ in the second iteration. The new centres for cluster 1 and cluster 2 are updated to $c_1 = 3$ and $c_2 = 25$.

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



Solution:
$Ncut(A, B) = cut(A, B) \frac{Vol(A)+Vol(B)}{Vol(A)Vol(B)}$

We first show the similarity matrix for the data $S = \begin{pmatrix} 1 & 0.7 & 0.9 & 0 & 0.2 & 0 \\ 0.7 & 1 & 0.5 & 0 & 0 & 0 \\ 0.9 & 0.5 & 1 & 0.1 & 0 & 0 \\ 0 & 0 & 0.1 & 1 & 0.4 & 0.6 \\ 0.2 & 0 & 0 & 0.4 & 1 & 0.9 \\ 0 & 0 & 0 & 0.6 & 0.9 & 1 \end{pmatrix}$.
Note that the similarity of data between itself is 1 because the Gaussian Kernel in Eqn. (5) is equal to 1 when $x_i = x_j$. The similarity of data points without direct link in the graph is set to 0 because we assume $|x_i - x_j|^2 \to \infty$.

$$\mathbf{W}(i, j) = \exp \frac{-|x_i - x_j|^2}{\sigma^2} \tag{5}$$

By the definition of volume in page 38 of week 8 slide, $Vol(A) = \sum s_{ij},\ i = 1, 2, 3; j = 1, 2, 3, 4, 5, 6$ with s_{ij} is an element in S and i and j are the entries of S. So $Vol(A) = 7.5$. Similarly, $Vol(B) = \sum s_{ij},\ i = 4, 5, 6; j = 1, 2, 3, 4, 5, 6$ with $Vol(B) = 7.1$. Finally, the $cut(A, B) = \sum s_{ij},\ i = 1, 2, 3; j = 4, 5, 6$ with $cut(A, B) = 0.3$. Note that S is symmetrical, so $cut(A, B)$ can be also calculated by $cut(A, B) = \sum s_{ij},\ i = 4, 5, 6; j = 1, 2, 3$ with the same value.

With $Vol(A), Vol(B), cut(A, B)$ calculated above, the normalised cut $Ncut(A, B)$ is obtained by:

$$Ncut(A, B) = cut(A, B)\frac{Vol(A) + Vol(B)}{Vol(A)Vol(B)} = 0.3 \times \frac{7.5 + 7.1}{7.5 \times 7.1} = 0.08225 \tag{6}$$

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).

Solution: The most elegant proof is from https://people.eecs.berkeley.edu/~jordan/courses/294-fall09/lectures/dimensionality/paper-1x2.pdf.

Let us denote an **orthonormal** projection vector as $\mathbf{u}$. It will project an input vector $\mathbf{x}$ to a scalar $y = \mathbf{u}^\top \mathbf{x}$. Using this scalar to reconstruct $\mathbf{x}$ as $\hat{\mathbf{x}} = \mathbf{u}y = \mathbf{u}\mathbf{u}^\top \mathbf{x}$.

$$\begin{aligned}
& \text{Reconstruction error} \\
= & \sum_{i=1}^{N} \left\| \mathbf{x}^{(i)} - \hat{\mathbf{x}}^{(i)}) \right\|^2 && (7) \\
= & \sum_{i=1}^{N} \left\| \mathbf{x}^{(i)} - \mathbf{u}\mathbf{u}^\top \mathbf{x}^{(i)} \right\|^2 && (8) \\
= & \sum_{i=1}^{N} \left(\mathbf{x}^{(i)} - \mathbf{u}\mathbf{u}^\top \mathbf{x}^{(i)})^\top (\mathbf{x}^{(i)} - \mathbf{u}\mathbf{u}^\top \mathbf{x}^{(i)} \right) && (9) \\
= & \sum_{i=1}^{N} \left(\mathbf{x}^{(i)\top} - \mathbf{x}^{(i)^\top} \mathbf{u}\mathbf{u}^\top)(\mathbf{x}^{(i)} - \mathbf{u}\mathbf{u}^\top \mathbf{x}^{(i)} \right) && (10) \\
= & \sum_{i=1}^{N} \left(\mathbf{x}^{(i)\top} \mathbf{x}^{(i)} - \mathbf{x}^{(i)^\top} \mathbf{u}\mathbf{u}^\top \mathbf{x}^{(i)} - \mathbf{x}^{(i)\top} \mathbf{u}\mathbf{u}^\top \mathbf{x}^{(i)} + \mathbf{x}^{(i)^\top} \mathbf{u}\mathbf{u}^\top \mathbf{u}\mathbf{u}^\top \mathbf{x}^{(i)} \right) \text{ note } \mathbf{u}^\top\mathbf{u}{=}1 \\
= & \sum_{i=1}^{N} \left(\mathbf{x}^{(i)\top} \mathbf{x}^{(i)} - \mathbf{x}^{(i)^\top} \mathbf{u}\mathbf{u}^\top \mathbf{x}^{(i)} \right) && (11) \\
= & \text{ constant} - \sum_{i=1}^{N} \left(\mathbf{x}^{(i)^\top} \mathbf{u}\mathbf{u}^\top \mathbf{x}^{(i)} \right) && (12) \\
= & \text{ constant} - \sum_{i=1}^{N} \left(\mathbf{u}^\top \mathbf{x}^{(i)} \right)^2 && (13)
\end{aligned}$$

Note $\mathbf{u}^\top \mathbf{x}^{(i)}$ is the projection $y^{(i)} = \mathbf{u}^\top \mathbf{x}^{(i)}$ so the summation in Eqn. (13) is the

variance. Maximising the variance minimises the reconstruction error so we have the same solution as that by variance maximisation.

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.

Solution:
$(D - W)y = \lambda Dy$
$(D - W)y = \lambda D^{\frac{1}{2}} D^{\frac{1}{2}} y$
$D^{\frac{-1}{2}}(D - W)y = \lambda D^{\frac{-1}{2}} D^{\frac{1}{2}} D^{\frac{1}{2}} y$
$D^{\frac{-1}{2}}(D - W)D^{\frac{-1}{2}} D^{\frac{1}{2}} y = \lambda I D^{\frac{1}{2}} y$
Make the substitution of $z = D^{\frac{1}{2}} y$
$D^{\frac{-1}{2}}(D - W)D^{\frac{-1}{2}} z = \lambda z$
If we set y to $\mathbf{1}$ we get
$z = D^{\frac{1}{2}}\mathbf{1}$
$D^{\frac{-1}{2}}(D - W)D^{\frac{-1}{2}} D^{\frac{1}{2}}\mathbf{1} = \lambda D^{\frac{1}{2}}\mathbf{1}$
$D^{\frac{-1}{2}}(D - W)I\mathbf{1} = \lambda D^{\frac{1}{2}}\mathbf{1}$
If we observe $(D - W)\mathbf{1}$, we can see that it's a summation of the rows of $D - W$
In a row of the Laplacian, $D-W$, we have the degree of the ith node, d_i on the diagonal, and all of the negative weights of the edges connected to node i filling the rest of the row. Therefore, adding across a row gives us:
$d_i + \sum_{j=1}^{n}(-w_{i,j}) = \sum_{j=1}^{n} w_{i,j} + \sum_{j=1}^{n}(-w_{i,j}) = 0$
Which means $(D-W)\mathbf{1} = \mathbf{0}$ and therefore the eigenvector corresponds to the eigenvalue $\lambda = 0$.

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.

Solution: *Note: The solution provided here is made to be comprehensive to give you more insights. However, your answer will be considered as correct as long as you can sketch correctly to show what the likelihood looks like typically.*

Likelihood function (`https://en.wikipedia.org/wiki/Likelihood_function`) tells how likely the observed data can be generated by a given model. Given observed data $\mathcal{D}$, the likelihood is expressed as $p(\mathcal{D} \mid \mathbf{w})$ or more specificly $p(\mathbf{y} \mid \mathbf{w}, \mathbf{x})$ with $\mathbf{y} = f(\mathbf{x}, \mathbf{w})$. With the linear model defined by:

$$y = w_0 + w_1 x \tag{1}$$

To make the model generate the observed data point $(x = -0.9, y = -0.1)$, w_0 and w_1 need to satisfy:

$$-0.9 = w_0 - 0.1 w_1 \tag{2}$$

Eqn. (2) tells a model with any pair of w_0 and w_1 that satisfies it can confidently generate a data point like $(x = -0.9, y = -0.1)$, therefore the the likelihood w.r.t such pairs of w_0 and w_1 is 1. We can draw a line in the space formed by w_0 and w_1, showing the models with the parameters on the line having a high likelihood.

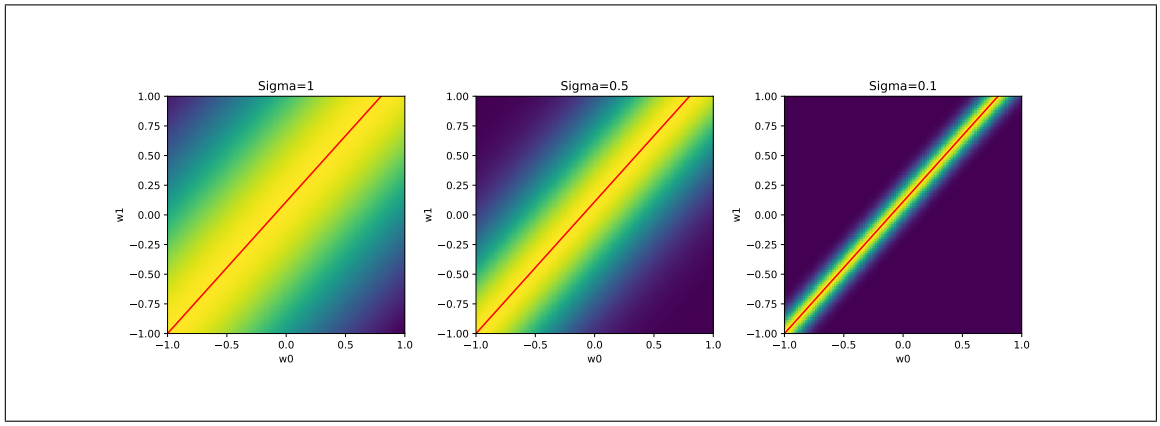
However, our belief is the output of the model is disturbed by noise which is independent from the input, thus the linear model becomes:

$$y = w_0 + w_1 x + \epsilon \tag{3}$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$. With the linear model expressed by Eqn. (3), the likelihood function can be written as:

$$p(y \mid w_0, w_1, x) = \mathcal{N}(w_0 + w_1 x, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(y - w_0 - w_1 x)^2} \tag{4}$$

With $(x = -0.9, y = -0.1)$ given, we can calculate Eqn. (4) in the space formed by w_0 and w_1. We show the sketch of the likelihood in the following figure under different σ value. Eqn. (2) is shown as the red straight line in the figure to illustrate the likelihood when Eqn. (1) is used as the model. Note that σ reflects the uncertainty of the output generated by Eqn. (3). As such, the smaller σ is, the more confident we are that the observed data is generated by the models that lies on the line expressed by Eqn. (2).

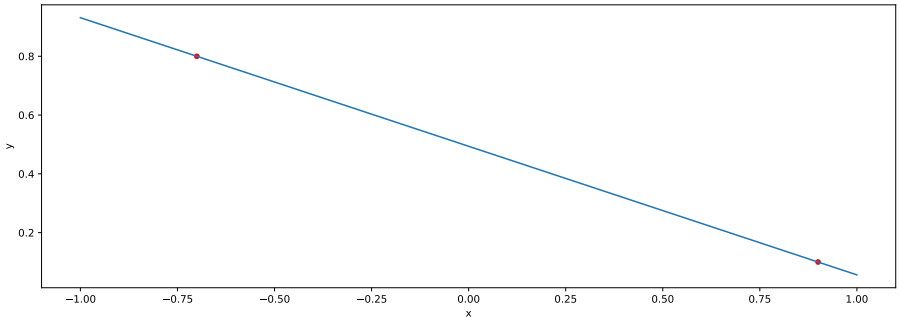


2 2. Slide 20: if the second observed data point is $(x = -0.7, y = 0.8)$ instead, sketch what the posterior will look like on this slide, assuming the first observed data point is still as it is $(x = 0.9, y = 0.1)$.

Solution: *Note: The solution provided here is made to be comprehensive to give you more insights. However, your answer will be considered as correct as long as you can sketch correctly to show what the posterior looks like typically. You can also get the likelihood as above and multiply with the current prior.*

The posterior is a PDF (probability density function) of the parameters of model conditioned by the evidence (observed data $\mathcal{D}$), represented by $p(\mathbf{w} \mid \mathcal{D})$ or $p(\mathbf{w} \mid \mathbf{y}, \mathbf{x})$. It tells the probability of the model given the observed evidence.

Under the assumption of a linear model defined by Eqn. (1), if we observed two data points $(x = -0.7, y = 0.8)$ and $(x = 0.9, y = 0.1)$, we could draw a straight line connecting the two points to justify this is exactly the model we get after seeing the data points.



Therefore, the parameters of the model can be calculated by

$$\begin{aligned} -0.7w_1 + w_0 &= 0.8 \\ 0.9w_1 + w_0 &= 0.1 \end{aligned} \tag{5}$$

resulting in $w_0 = 0.4938, w_1 = -0.4375$. We can draw this point in the space formed by w_0 and w_1 to illustrate the posterior PDF at this point has a probability density of infinite value (CDF is 1 under an infinitely small area).

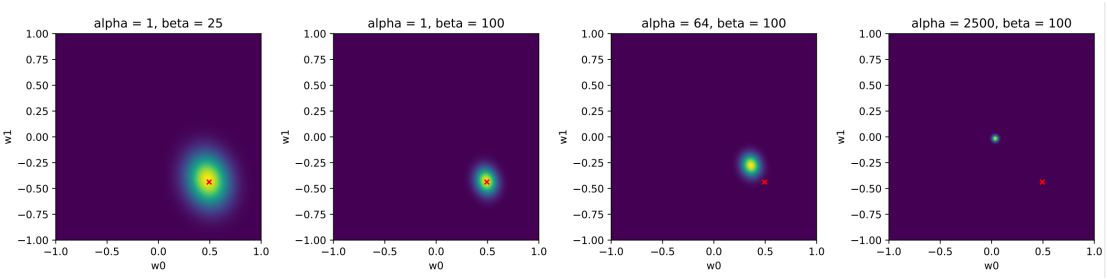
Similar to the first question, we believe the model should include noise and be represented by Eqn. (3). Therefore, the parameters of model can be relaxed from a

fixed line. Intuitively, the posterior will look like to have an area centred by the point $w_0 = 0.4938, w_1 = -0.4375$ with high probability density. Next, let us show if our intuition is correct in a diagram.

With a Gaussian prior and Gaussian noise assumption:

$$\begin{aligned} p(\mathbf{w} \mid \alpha) &= \mathcal{N}\left(\mathbf{w} \mid \mathbf{0}, \alpha^{-1}\mathbf{I}\right) \\ p(\epsilon) &= \mathcal{N}\left(\epsilon \mid 0, \beta^{-1}\right) \end{aligned} \tag{6}$$

where β^{-1} is the inverse variance, we can obtain a closed-form of the posterior. We borrow the formulas and implementation from A2 and A3 in lab9, using two data points $(x = -0.7, y = 0.8)$ and $(x = 0.9, y = 0.1)$, the posterior in the space of w_0 and w_1 is shown in the following diagram. The red cross dot represent the point $w_0 = 0.4938, w_1 = -0.4375$. Different combinations of α and β are used in the evaluation.



The diagram shows that with a larger β (inverse variance), the area has a smaller size. Because the output of the model is less uncertain with a large β, the possible models are more similar to the fixed line represented by the red cross dot. Moreover, the diagram shows that our aforementioned intuition is incorrect when α is large. This is because the prior $p(\mathbf{w} \mid \alpha) = \mathcal{N}\left(\mathbf{w} \mid \mathbf{0}, \alpha^{-1}\mathbf{I}\right)$ with a large α will become a dominant factor over the likelihood in the posterior expressed by Eqn. (7). The possible parameters generated by the Gaussian prior is mostly confined to a small area with mean value as centre given a large α. The results show that our belief on the prior is critical to the model learned from the evidence.

$$p(\mathbf{w} \mid \mathbf{y}, \mathbf{x}, \alpha) \propto p(\mathbf{y} \mid \mathbf{w}, \mathbf{x}) p(\mathbf{w} \mid \alpha) \tag{7}$$

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

Solution:
Let $X_1, \cdots, X_n$ be iid Bernoulli random variables with parameter π, $0 < \pi < 1$. Then $\sum_{i=1}^{n} X_i$ is a sufficient statistic for π.

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.

Solution: $x \sim \mathcal{N}(0, 1)$
We know that:

$\mathbb{E}[x] = 0$
$Var(x) = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = 1$
We can obtain $z \sim \mathcal{N}(\mu, \sigma)$ by setting $z = \mu + \sigma x$.

Let us prove it now.

Mean:
$\mathbb{E}[z] = \mu + \sigma\mathbb{E}[x]$
$= \mu + \sigma(0) = \mu$
$Var(z) = \mathbb{E}[z^2] - \mathbb{E}[z]^2$
$= \mathbb{E}[(\mu + \sigma x)^2] - (\mu)^2$
$= (\mu^2 + 2\mu\sigma\mathbb{E}[x]) + \sigma^2\mathbb{E}[x^2] - (\mu)^2$
Looking at the defined formula for $Var(x)$ above. We know that:
$\mathbb{E}[x^2] = 1 - \mathbb{E}[x]^2 = 1$
Therefore:
$Var(z) = \mu^2 + 2\mu\sigma(0) + \sigma^2(1) - \mu^2$
$= \sigma^2$