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:

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:

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:

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 $P\left(X=x,Y=y\right)=P\left(X=x\right)P\left(Y=y\right)$. We therefore check whether $P\left(X=x,Y=y\right)=P\left(X=x\right)P\left(Y=y\right)$ or not.

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

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

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

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

We observe

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

Since $P\left(X=x,Y=y\right)\neq P\left(X=x\right)P\left(Y=y\right)$, 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

$$\begin{aligned}E\{X\}&=\sum_{x_i}x_iP\left(X=x_i\right)=0\times P(X=0)+1\times P(X=1)+2\times P(X=2)\\ &=0\times\left(\frac{1}{3}\right)+1\times\left(\frac{1}{3}\right)+2\times\left(\frac{1}{3}\right)=1\\ E\{Y\}&=\sum_{y_j}y_jP\left(Y=y_j\right)=0\times P(Y=0)+1\times P(Y=1)=0\times\left(\frac{1}{3}\right)+1\times\left(\frac{2}{3}\right)=\frac{2}{3}\\ E\{XY\}&=\sum_{y_j}\sum_{x_1}x_iy_jP\left(X=x_i,Y=y_j\right)\\ &=0\times1\times P(X=0,Y=1)+1\times0\times P(X=1,Y=0)+2\times1\times P(X=2,Y=1)\\ &=0\times1\times\left(\frac{1}{3}\right)+1\times0\times\left(\frac{1}{3}\right)+2\times1\times\left(\frac{1}{3}\right)=\frac{2}{3}\end{aligned}$$

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,

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