COMS 4030A
Adaptive Computation and Machine Learning

4. Model Testing

Recall that before training a model on a given dataset, a portion of the data is set aside to be used as testing; this is called the test dataset. The test dataset must be independent of the training process; usually, it comprises anywhere between 20% and 50% of the original dataset. The test dataset is different to the validation dataset, which is used in the training process to avoid overfitting or for tuning hyperparameters.

4.1. **Evaluation metrics.**

In this section we discuss methods that use the test dataset to evaluate how good (or how fit) a model is and, thereby, to decide if it should be kept or discarded.

Consider a classification problem in which we have a dataset consisting of data points and corresponding target classes and we seek to create a model that will correctly predict the class of each data point. Suppose that a model has been trained for this dataset. Evaluating such a model on a test dataset is done by applying the model to every data point in the test dataset and comparing the class predicted by the model with the actual class of the data point in the target dataset.

Note that the output of the model must be converted into a class prediction.
For example, if the model is a neural network and the output obtained for some data point is $(0.7, 0.2, 0.1)$, then the predicted class is the one corresponding to $(1, 0, 0)$.

The following metric can be used on the test dataset:

$$\textbf{Accuracy} = \frac{\text{number of correct predictions}}{\text{total number of predictions}}$$

The model's accuracy on the dataset is a value between 0 and 1, which is often converted to a percentage. Accuracy is not always the best metric and there are other metrics that can be used to determine how good a model is.

Consider the case of **binary classification** in which there are only two possible classes to predict. Let the two classes be 0 and 1.

The results of applying a model on the test dataset can be represented as in the following example:

prediction

actual		1	0
	1	70	30
	0	15	50

The above table is called a **confusion matrix**. The table shows the predicted class versus the actual class for each data point. In the table we can read off that 70 of the data points in class 1 were corrected predicted, 30 of the data points in class 1 were incorrectly predicted as 0, 15 of the data points in class 0 were incorrectly predicted as 1, and 50 of the data points in class 0 were correctly predicted.

There are 165 data points in the test dataset. The accuracy of the model on the test dataset is $\frac{70+50}{165} = 0.7273$, or 72.73%.

To see why accuracy is not always a good metric, consider the next example. Suppose the results of applying the model on the test dataset are given by the following confusion matrix:

prediction

actual		1	0
	1	2	4
	0	9	65

In this example, there are 80 data points in the test set; 2 data points in class 1 were correctly predicted as 1, 4 data points in class 1 were incorrectly predicted as 0, 9 data points in class 0 were incorrectly predicted as 1 and 65 data points in class 0 were correctly predicted as 0.

The accuracy of the model on this test dataset is $\frac{2+65}{80} = 83.75\%$.

Although 83.75% seems like a good accuracy, there is simpler model that does better.

Consider the model that always chooses class 0 regardless of the data point.

Since the test dataset has 74 actual 0's out of 80 datapoints, this model will have accuracy of $\frac{74}{80} = 92.5\%$.

However, such a model does not help with the original dataset, since there is (presumably) a need to predict class 1 in some cases. Thus, accuracy is not a good metric in this case. The reason for this is that the (test) dataset is very unbalanced and the large number of 0 classes dominates the accuracy metric. For datasets with an even distribution of classes, i.e., a **balanced** dataset, accuracy is a good metric.

To discuss other metrics, we introduce some notation. We still consider binary classification using classes 1 and 0. A data point with classification 1 will be called **positive** and a data point with classification 0 will be called **negative**. We define the following:

True Positive (TP): when the predicted class is 1 and the actual class is 1.

True Negative (TN): when the predicted class is 0 and the actual class is 0.

False Positive (FP): when the predicted class is 1 and the actual class is 0.

False Negative (FN): when the predicted class is 0 and the actual class is 1.

Using the above notation we can define the **confusion matrix** as the following table:

		prediction	
		1	0
actual	1	TP	FN
	0	FP	TN

The formula for calculating accuracy from TP, TN, FP and FN is as follows:

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN}$$

Other metrics that are based on TP, TN, FP and FN are:

$$\textbf{Precision} = \frac{TP}{TP+FP}$$

$$\textbf{Sensitivity} = \textbf{Recall} = \frac{TP}{TP+FN}.$$

Precision is the positive predictive value.
Sensitivity, or Recall, is the true positive rate.

In the above example,

$$\text{Precision} = \frac{2}{2+9} = 0.18$$

$$\text{Sensitivity} = \frac{2}{2+4} = 0.33.$$

Deciding which of the metrics – precision or sensitivity – to use, depends on the dataset. Precision is used in datasets where you want few false positives, i.e., you want to be sure your positive predictions are correct (e.g., in medical tests, it's important that there are few false positives, or incorrect diagnoses of illnesses). Sensitivity is used if the number of false negatives is more important, that is, you don't want to miss many 1's (e.g., in fraud detection).

A metric that provides a balance between precision and sensitivity is the ***F*1 score**, defined as follows:

$$F1 = \frac{2 * \text{precision} * \text{sensitivity}}{\text{precision} + \text{sensitivity}}$$

The $F1$ score is an average of precision and sensitivity, in fact, it is the **harmonic mean** of precision and sensitivity. The harmonic mean of numbers a and b is defined as the reciprocal of the average of the reciprocals:

$$\text{harmonic mean of } a \text{ and } b = \frac{1}{\frac{1}{2}(\frac{1}{a} + \frac{1}{b})} = \frac{2ab}{a+b}.$$

The $F1$ score in the above example is $\frac{2(0.18)(0.33)}{0.18+0.33} = 0.23$.

In the first example above, precision $= \frac{70}{85} = 0.82$, sensitivity $= \frac{70}{100} = 0.7$ and $F1 = \frac{2(0.82)(0.7)}{0.82+0.7} = 0.76$.

Next, consider the case of multi-class classification. Suppose that a dataset has three classes, say A, B and C. If a model is trained on the training dataset and tested on the test dataset, then the confusion matrix would look something like:

$$\begin{array}{cc} & \text{prediction} \\ \text{actual} & \left[\begin{array}{c|c|c|c} & A & B & C \\ \hline A & 42 & 6 & 2 \\ \hline B & 3 & 50 & 7 \\ \hline C & 0 & 6 & 34 \end{array}\right] \end{array}$$

In the above confusion matrix, the first row indicates that there are 50 data points with target A; of these, 42 are correctly classified as A, 6 are misclassified as B and 2 are misclassified as C. The second row indicates that there are 60 data points with target B; of these, 50 are correctly classified as B, 3 are misclassified as A and 7 are misclassified as C. The third row indicates that there are 40 data points with target C; of these, 34 are correctly classified as C, 0 are misclassified as A and 6 are misclassified as B.
The accuracy of the model is $\frac{42+50+34}{42+50+34+6+2+3+7+0+6} = 0.84$, that is, 84%.

The notions of sensitivity and precision can be defined for each of the classes.
For example, for class A, if we consider A as positive, then

$$\text{precision} = \tfrac{42}{42+3+0} = 0.933$$
$$\text{sensitivity} = \tfrac{42}{42+6+2} = 0.84$$
$$F1 = \tfrac{2*(0.933)(0.84)}{0.933+0.84} = 0.884.$$

For class B, if we consider B as positive, then

$$\text{precision} = \tfrac{50}{6+50+6} = 0.806$$
$$\text{sensitivity} = \tfrac{50}{3+50+7} = 0.833$$
$$F1 = \tfrac{2*(0.806)(0.833)}{0.806+0.833} = 0.819.$$

For class C, if we consider C as positive, then

$$\text{precision} = \tfrac{34}{2+7+34} = 0.791$$
$$\text{sensitivity} = \tfrac{34}{0+6+34} = 0.85$$
$$F1 = \tfrac{2*(0.791)(0.85)}{0.791+0.85} = 0.819.$$

Lastly, for each of sensitivity, precision and $F1$, an average value over the three classes can be obtained, which is denoted as the **macro** score. In the above example,

$$\text{macro-precision} = \tfrac{0.933+0.806+0.791}{3} = 0.819$$
$$\text{macro-sensitivity} = \tfrac{0.84+0.833+0.85}{3} = 0.841$$
$$\text{macro-}F1 = \tfrac{0.884+0.819+0.819}{3} = 0.841.$$

4.2. **Cross-validation.**

Another method for testing a model on a dataset is **k-fold cross-validation**.
(Confusingly, 'validation' in this context actually means testing, so 'k-fold cross testing' would be a better name.)

This method is often used if the dataset is not sufficiently large to accommodate a large enough training set and test set. An advantage of this method is that every data point is used both in training and in testing in some way.

The method is as follows.

Choose the type of model to be applied.
Choose a number k (usually in the range 5 - 10, but any value will do).
Split the dataset into k approximately equal parts, which are called the **folds**.
Choose k1 of the folds as the training set. The remaining fold will be the test set.
Train the model on the training set. (If necessary, use a portion of the dataset to prevent overtraining).
Use the remaining fold as the test dataset to obtain either a loss value, accuracy, or other metric.

Repeat the above process k times, each time with a different fold selected for testing.
Note that for each split of the data, a new model is trained from scratch and is independent of the models trained on other splits.

The final metric is obtained by averaging the metrics obtained on each fold.

A good final metric obtained by the cross validation method provides confidence in the type of model being used. Note that the above method results in k different models of the selected type. The best model can be chosen as the final model, or a fresh model can be trained.

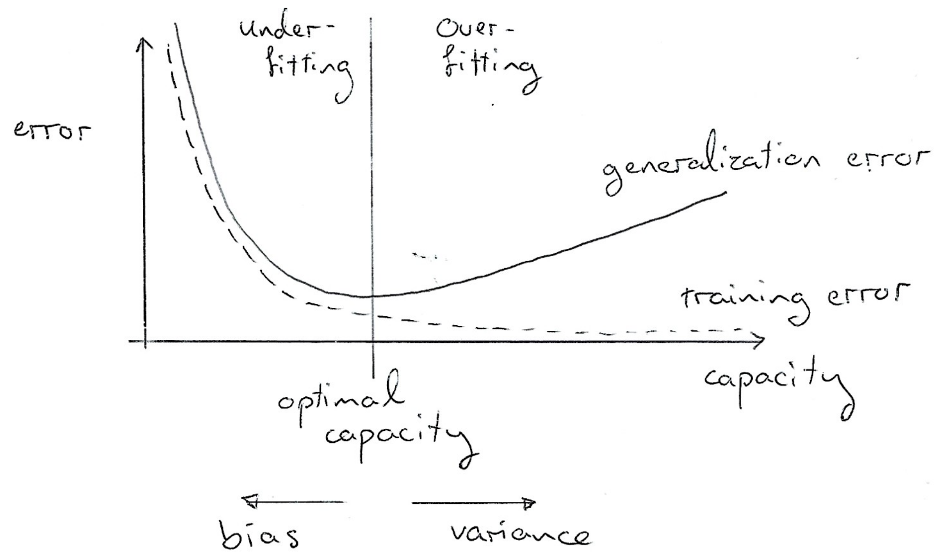
If the final metric obtained from cross validation is not good enough, then it can mean that the type of model is not suitable and should be adjusted. For example, in the case of a neural network, it can mean that the number of nodes or layers in the network should be changed.

4.3. **Underfitting vs Overfitting, Bias vs Variance.**

The following diagram summarises the issues of overfitting and underfitting when choosing a model type. Loosely speaking, the **capacity** of a model type correlates to the complexity of the datasets that it is able to accurately model.

For example, consider a regression problem in one variable. A model type with low capacity is straight line models since they can only accurately fit data that is linear. On more complex datasets, a straight line model may do very poorly; we call this **underfitting**. Quadratic models have greater capacity to model datasets so underfitting is less of a problem, but now the risk of overfitting increases. As the model type becomes more complex the capacity increases, but so does the risk of overfitting. If we use neural networks as an example, the capacity is increased by adding more layers and more nodes to the model. However, it is difficult to know what a good capacity is. The optimal capacity is obtained by minimising the generalisation error, i.e., the error on the test set.

The issue of underfitting versus overfitting is often phrased as trade-off between **bias** and **variance**. Reducing the capacity of the model means increasing the bias – the trained model is then more biased towards the type of model chosen and, hence, more susceptible to underfitting. Increasing the capacity of the model means there is greater variance in what the trained model can be, but then it is more susceptible to overfitting.



Consider the following dataset that has only one attribute, the COMS2 mark, and one target value which is either $Pass$ or $Fail$ in COMS3:

$$\begin{array}{cc} \text{COMS2} & \text{COMS3} \end{array}$$
$$\left[\begin{array}{c|c} A & Pass \\ C & Fail \\ C & Pass \\ B & Pass \\ B & Fail \\ C & Pass \\ A & Fail \\ B & Pass \end{array}\right]$$

Suppose a student got a B in COMS2 and is currently doing COMS3. Based on the above dataset, what prediction could we make for this student: $Pass$ or $Fail$?
We can calculate the probability of $Pass$ by looking at all the rows in which COMS2 $= B$. There are 3 such rows, and of the 3, there are 2 $Pass$'s, so the probability of $Pass$ given that the student got a B in COMS2 is $\frac{2}{3}$. We write this using conditional probabilities as:

$$P(Pass|\text{COMS2} = B) = \tfrac{2}{3}, \quad \text{or} \quad P(Pass|B) = \tfrac{2}{3}.$$

Alternative, we can use **Bayes' Rule** which states: for events X and Y,

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}.$$

Bayes' rule gives us a way of calculating the probability of X given Y, i.e., $P(X|Y)$, from the probability of Y given X, i.e., $P(Y|X)$. In Bayes' rule,

$P(X)$ is called the **prior** probability,

$P(X|Y)$ is called the **posterior** probability,

$P(Y|X)$ is called the **likelihood** (of Y given X).

In the above example, using Bayes' rule, we calculate:

$$P(Pass|B) = \frac{P(B|Pass)P(Pass)}{P(B)}.$$

Now, $P(B|Pass) = \frac{2}{5}$. This comes from looking at all the rows which have COMS3 $= Pass$, of which there are 5, and counting the number of these rows in which COMS2 $= B$, which is 2.
Also, $P(Pass) = \frac{5}{8}$, since we have 8 datapoints and in 5 of these COMS3 $= Pass$.
Lastly, $P(B) = \frac{3}{8}$, since we have 8 datapoints and in 3 of these COMS2 $= B$.
Putting this into the above equation, we get:

$$P(Pass|B) = \frac{P(B|Pass)P(Pass)}{P(B)} = \frac{(\frac{2}{5})(\frac{5}{8})}{\frac{3}{8}} = \frac{2}{3}.$$

We can similarly work out that $P(Fail|B) = \frac{1}{3}$. Based on these probabilities, we hypothesise that the most likely outcome for the student who got B in COMS2 is $Pass$.

5.1. **MAP hypothesis.**

Consider a dataset S in which the targets are the classes $C_1, \ldots, C_n$. Given a new data point $\boldsymbol{x} = (x_1, \ldots, x_m)$ we want to decide which C_i to classify it as. Suppose we can obtain the following conditional probabilities:

$$P(C_1|\boldsymbol{x}), \ \ldots, \ P(C_n|\boldsymbol{x}).$$

Here, $P(C_i|\boldsymbol{x})$ is the probability that the given datapoint $\boldsymbol{x}$ is classified as C_i.
Based on these probabilities, we would classify $\boldsymbol{x}$ as the C_i for which $P(C_i|\boldsymbol{x})$ is maximum.
This is the **maximum a posteriori** hypothesis, or **MAP** hypothesis.

Observe that, by Bayes' rule, for each class C_i, we have:

$$P(C_i|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|C_i)P(C_i)}{P(\boldsymbol{x})}$$

Thus, to find the MAP hypothesis, we need to find the maximum of:

$$\frac{P(\boldsymbol{x}|C_1)P(C_1)}{P(\boldsymbol{x})}, \ \ldots, \ \frac{P(\boldsymbol{x}|C_n)P(C_n)}{P(\boldsymbol{x})}.$$

Since each of these terms has the same denominator, it is sufficient to find the maximum of:

$$P(\boldsymbol{x}|C_1)P(C_1), \ \ldots, \ P(\boldsymbol{x}|C_n)P(C_n).$$

Example: Consider the following dataset on students doing COMS3:

$$S = \left[\begin{array}{c|c|c|c} \text{COMS2} & \text{doing labs?} & \text{doing tuts?} & \text{COMS3} \\ \hline A & N & Y & Pass \\ C & Y & N & Fail \\ C & N & Y & Pass \\ B & Y & Y & Pass \\ B & N & Y & Fail \\ C & Y & N & Pass \\ A & N & N & Fail \\ B & Y & N & Pass \end{array}\right]$$

In the above dataset, the attributes of the data are: COMS2, doing labs?, and doing tuts?. The target value is in the column COMS3.

Suppose a current student had a B for COMS2, is not doing labs and is not doing tuts, i.e., $\boldsymbol{x} = (B, N, N)$. We want to predict the outcome of COMS3, i.e., $Pass$ or $Fail$, by using the MAP hypothesis. That is, we want to determine which of the following is larger: $P(Pass|B, N, N)$ and $P(Fail|B, N, N)$.
First, apply Bayes' rule:

$$P(Pass|B,N,N) = \frac{P(B,N,N|Pass)P(Pass)}{P(B,N,N)},$$
$$P(Fail|B,N,N) = \frac{P(B,N,N|Fail)P(Fail)}{P(B,N,N)}$$

To determine which of the above is larger, we need only find the larger of:

$$P(B,N,N|Pass)P(Pass) \text{ and } P(B,N,N|Fail)P(Fail)$$

From the target column of the dataset we get $P(Pass) = \frac{5}{8}$ and $P(Fail) = \frac{3}{8}$.

To find $P(B,N,N|Pass)$ and $P(B,N,N|Fail)$ is more difficult since we do not even have a situation like this in our dataset. We use a simplifying assumption here. Assume that the attributes COMS2, doing labs? and doing tuts? are conditionally independent with respect to $Pass$ and $Fail$. This means that

$$P(B,N,N|Pass) = P(\text{COMS2} = B|Pass)P(\text{doing labs?} = N|Pass)P(\text{doing tuts?} = N|Pass).$$

Thus, we can calculate as follows:

$$P(B,N,N|Pass) = \left(\tfrac{2}{5}\right)\left(\tfrac{2}{5}\right)\left(\tfrac{2}{5}\right) = \tfrac{8}{125}$$

hence

$$P(B,N,N|Pass)P(Pass) = \left(\tfrac{8}{125}\right)\left(\tfrac{5}{8}\right) = \tfrac{1}{25}.$$

Similarly, we assume that

$$P(B,N,N|Fail) = P(\text{COMS2} = B|Fail)P(\text{doing labs?} = N|Fail)P(\text{doing tuts?} = N|Fail).$$

so we calculate

$$P(B,N,N|Fail) = \left(\tfrac{1}{3}\right)\left(\tfrac{2}{3}\right)\left(\tfrac{2}{3}\right) = \tfrac{4}{27}$$

hence

$$P(B,N,N|Fail)P(Fail) = \left(\tfrac{4}{27}\right)\left(\tfrac{3}{8}\right) = \tfrac{1}{18}.$$

Since $\frac{1}{18} > \frac{1}{25}$, we hypothesise that the student is more likely to $Fail$, i.e., we classify the student as $Fail$.

The method used in the above example is the Naïve Bayes Classifier.

5.2. **Naïve Bayes Classifier.**

Suppose we have a dataset S in which the targets are the classes $C_1, \ldots, C_n$, and, for a new data point $\boldsymbol{x} = (x_1, \ldots, x_m)$, we want to decide which C_i to classify it as. In order to use the MAP hypothesis, we must find the maximum of:

$$P(\boldsymbol{x}|C_1)P(C_1), \ldots, P(\boldsymbol{x}|C_n)P(C_n).$$

If we make the simplifying assumption that the attributes x_i are all conditionally independent with respect to the C_is, then:

$$P(\boldsymbol{x}|C_i) = P(x_1|C_i)P(x_2|C_i)\cdots P(x_m|C_i).$$

With this simplification, we classify $\boldsymbol{x}$ as the class C_i which gives the maximum of

$$P(\boldsymbol{x}|C_1)P(C_1),\ \ldots,\ P(\boldsymbol{x}|C_n)P(C_n).$$

This is known as the **Naïve Bayes Classifier**.

The word 'naïve' refers to the naïve assumption that the attributes are conditionally independent. This is usually not the case.

Example: Consider the dataset in the previous example, and now suppose that a student got C for COMS2, is doing labs and is not doing tuts, i.e., $\boldsymbol{x} = (C, Y, N)$. In the dataset, there are two previous cases of such students - one of which *Passed* and one which *Failed* COMS3.

So we cannot make a prediction based on these two cases. Let's use the Naïve Bayes Classifier to make a classification.

We want to find the larger of:

$$P(C, Y, N|Pass)P(Pass) \text{ and } P(C, Y, N|Fail)P(Fail).$$

Using the conditional independence assumption we calculate:

$$\begin{aligned}
&P(C, Y, N|Pass)P(Pass) \\
&= P(\text{COMS2} = C|Pass)P(\text{doing labs?} = Y|Pass)P(\text{doing tuts?} = N|Pass)P(Pass) \\
&= \left(\tfrac{2}{5}\right)\left(\tfrac{3}{5}\right)\left(\tfrac{2}{5}\right)\left(\tfrac{5}{8}\right) \\
&= \tfrac{3}{50}
\end{aligned}$$

$$\begin{aligned}
&P(C, Y, N|Fail)P(Fail) \\
&= P(\text{COMS2} = C|Fail)P(\text{doing labs?} = Y|Fail)P(\text{doing tuts} = N|Fail)P(Fail) \\
&= \left(\tfrac{1}{3}\right)\left(\tfrac{1}{3}\right)\left(\tfrac{2}{3}\right)\left(\tfrac{3}{8}\right) \\
&= \tfrac{1}{36}
\end{aligned}$$

Thus, the Naïve Bayes Classifier would classify this student as $Pass$, since $\frac{3}{50} > \frac{1}{36}$.

5.3. **Applying Bayes' Rule.**

Bayes' rule can be used directly to make predictions about a dataset. The method of applying Bayes' rule is usually described using urns filled with balls of different colours. We give such an example below. In the exercises below there is an example that shows how this method may be applied to a data problem.

Suppose we have two urns containing some *Blue* balls and some *Red* balls, as follows:

Urn 1 contains 5 balls, 3 or which are *Red* and 2 of which are *Blue*.

Urn 2 contains 8 balls, 3 or which are *Red* and 5 of which are *Blue*.

A friend will select one of the urns by (secretly) flipping a coin.
If the coin lands heads up, he will choose Urn 1 and if it lands tails up, he will choose Urn 2.
You must guess which of the urns he chose.

You know that the urn was chosen by a coin-flip, so before you receive any new information, your best guess is that there is a $\frac{1}{2}$ probability that Urn 1 was chosen and a $\frac{1}{2}$ probability that Urn 2 was chosen. This is the **prior** probability distribution: $P(\text{Urn 1}) = \frac{1}{2}$ and $P(\text{Urn 2}) = \frac{1}{2}$.

Now your friend (secretly) takes a random ball from the chosen urn, looks at it, and returns it to the urn. He tells you that it was a *Red* ball. You must now estimate the probability that Urn 1 was chosen and the probability that Urn 2 was chosen. Naïvely, you can guess, correctly, that Urn 1 is more likely since it has a higher proportion of *Red* balls than Urn 2 has. We use Bayes' Rule to calculate the exact probabilities as follows:

$$P(\text{Urn 1}|Red) = \frac{P(Red|\text{Urn 1})P(\text{Urn 1})}{P(Red)}$$

We can work out that $P(Red|\text{Urn 1}) = \frac{3}{5}$ since 3 of the 5 balls in Urn 1 are *Red*. Also $P(\text{Urn 1}) = \frac{1}{2}$ as discussed above. To find $P(Red)$ we use marginalisation:

$$\begin{aligned} P(Red) &= P(Red|\text{Urn 1})P(\text{Urn 1}) + P(Red|\text{Urn 2})P(\text{Urn 2}) \\ &= \left(\tfrac{3}{5}\right)\left(\tfrac{1}{2}\right) + \left(\tfrac{3}{8}\right)\left(\tfrac{1}{2}\right) \\ &= \tfrac{39}{80}. \end{aligned}$$

Thus,

$$\begin{aligned} P(\text{Urn 1}|Red) &= \frac{P(Red|\text{Urn 1})P(\text{Urn 1})}{P(Red)} \\ &= \frac{\left(\frac{3}{5}\right)\left(\frac{1}{2}\right)}{\frac{39}{80}} \\ &= \tfrac{24}{39}. \end{aligned}$$

It follows that $P(\text{Urn 2}|Red) = 1 - \frac{24}{39} = \frac{15}{39}$.

Thus, our posterior probability distribution is: $P(\text{Urn 1}) = \frac{24}{39}$ and $P(\text{Urn 2}) = \frac{15}{39}$.

Let us suppose now that the friend (secretly) takes a second random ball from the chosen urn, and again, the ball is *Red*. This is more evidence that the chosen urn is Urn 1, but we can calculate the exact probability using Bayes' Rule again, i.e.:

$$P(\text{Urn 1}|Red) = \frac{P(Red|\text{Urn 1})P(\text{Urn 1})}{P(Red)}$$

We still have that $P(Red|\text{Urn } 1) = \frac{3}{5}$. Our prior probability distribution now has $P(\text{Urn } 1) = \frac{15}{39}$. To get $P(Red)$ we marginalise again:

$$\begin{aligned} P(Red) &= P(Red|\text{Urn } 1)P(\text{Urn } 1) + P(Red|\text{Urn } 2)P(\text{Urn } 2) \\ &= \left(\tfrac{3}{5}\right)\left(\tfrac{24}{39}\right) + \left(\tfrac{3}{8}\right)\left(\tfrac{15}{39}\right) \\ &= 0.513 \end{aligned}$$

Thus,

$$\begin{aligned} P(\text{Urn } 1|Red) &= \frac{P(Red|\text{Urn } 1)P(\text{Urn } 1)}{P(Red)} \\ &= \frac{(\frac{3}{5})(\frac{24}{39})}{0.513} \\ &= 0.720. \end{aligned}$$

It follows that $P(\text{Urn } 2|Red) = 1 - 0.720 = 0.280$.
Thus, our posterior probability distribution is $P(\text{Urn } 1) = 0.720$ and $P(\text{Urn } 2) = 0.280$.
This makes sense since it much more likely that Urn 1 is the chosen urn.

5.4. **Marginalisation.**

The formula we used for marginalisation to get $P(Red)$ above is given more generally as follows.

For an event X that can be classified as any one of the classes $C_1, \ldots, C_n$, we have that

$$P(X) = \sum_{i=1}^{n} P(X|C_i)P(C_i).$$

We say that we are **marginalising** X over the C_i's.

The above formula can be obtained as follows:
Event X must be classified as one of the classes $C_1, \ldots, C_n$, and it cannot be classified as more than one, so

$$P(C_1|X) + P(C_2|X) + \cdots + P(C_n|X) = 1$$

hence, by Bayes' Rule:

$$\frac{P(X|C_1)P(C_1)}{P(X)} + \frac{P(X|C_2)P(C_2)}{P(X)} + \cdots + \frac{P(X|C_n)P(C_n)}{P(X)} = 1$$

and the marginalisation formula follows if we multiply through by $P(X)$.

EXERCISES

(1) Consider the following dataset from Lecture 5:

$$S = \left[\begin{array}{c|c|c|c} F_1 & F_2 & F_3 & \text{target} \\ \hline T & 0 & B & Yes \\ F & 2 & A & No \\ F & 1 & C & Yes \\ T & 2 & B & Yes \\ F & 0 & A & No \\ F & 2 & C & No \\ F & 0 & A & Yes \\ T & 1 & B & No \\ T & 0 & B & Yes \\ F & 1 & C & Yes \\ T & 2 & A & No \\ T & 2 & C & No \end{array}\right]$$

Use the Naïve Bayes Classifier to classify input $(F, 1, A)$ as either Yes or No. Do the same for inputs $(F, 0, A)$ and for $(T, 2, C)$. Compare your answers with those obtained using decision trees in Lecture 5.

(2) Consider the following dataset from Lecture 5:

$$S = \left[\begin{array}{c|c|c|c} F_1 & F_2 & F_3 & \text{target} \\ \hline a & 0 & Y & A \\ b & 0 & N & C \\ c & 1 & Y & B \\ b & 1 & Y & B \\ c & 0 & N & A \\ a & 0 & N & C \\ c & 1 & N & B \\ a & 1 & Y & A \\ b & 0 & Y & B \\ a & 1 & N & C \end{array}\right]$$

Use the Naïve Bayes Classifier to classify input $(b, 1, N)$ as either A, B or C.

Do the same with input $(c, 1, Y)$. Compare your answers with those obtained by decision trees in Lecture 5.

(3) Consider an urn problem similar to the example in the notes, but where Urn 1 has 1 *Red* ball and 3 *Blue* balls, and Urn 2 has 7 *Red* balls and 3 *Blue* balls.

(a) A friend uses a fair coin to choose an urn. Give the prior probability distribution on the urns.

(b) Your friend secretly takes a ball out of the chosen urn (so you can't see which urn) and tells you the ball is *Blue* and then returns it to the urn. Use Bayes' rule to work out $P(\text{Urn } 1|Blue)$ and $P(\text{Urn } 2|Blue)$.

(c) Your friend then draws another ball out of the chosen urn, and tells you it is *Red*. Now work out $P(\text{Urn } 1|Red)$ and $P(\text{Urn } 2|Red)$.

(4) Suppose you have a dataset compiled from 200 randomly chosen people. For each person in the dataset you have some personal information and you also know if the person prefers cricket or soccer. There are 80 people in the dataset that prefer cricket and 120 people that prefer soccer.

(a) Suppose you want to know if a new random person prefers cricket or soccer. Based on the sample dataset, what is your prior probability that the person prefers cricket and the prior probability that the person prefers soccer?

(b) You are now informed that the new person is female. In your dataset, you know that 45 of the 80 people that prefer cricket are female, and 50 of the 120 people that prefer soccer are female. Use Bayes' rule to update your probability that the person prefers cricket and the probability that the person prefers soccer.

(c) You now receive some more information - the person is under 25 years old. In your dataset, of the 80 that prefer cricket, 30 are under 25 and 15 of them are female, and of the 120 that prefer soccer, 50 are under 25 and 30 of them are female. Use Bayes' rule to determine the probability that the person prefers cricket and the probability that the person prefers soccer.

(5) Suppose you are given a dataset S with N datapoints. Suppose S has attributes x_1, x_2, x_3 and target t, where attribute x_1 can take values in $\{T, F\}$, attribute x_2 can take values in $\{a, b, c\}$ and attribute x_3 can take values in $\{0, 1, 2, 3\}$. The target t can take values in $\{C_1, C_2, C_3\}$. Describe how to implement the Naïve Bayes Classifier for any input $\boldsymbol{x} = (x_1, x_2, x_3)$, where $x_1 \in \{T, F\}$, $x_2 \in \{a, b, c\}$ and $x_3 \in \{0, 1, 2, 3\}$.