IAML Tutorial 1

1. Suppose X and Y are two random variables. X takes on the value yes if the word 'password' occurs in an email, and no if the word is not present. Y takes on the values ham and spam. This example relates to spam filtering for e-mail.

Let $p(Y=\text{ham}) = p(Y=\text{spam}) = 0.5$
$p(X=\text{yes} \mid Y=\text{ham}) = 0.02$
$p(X=\text{yes} \mid Y=\text{spam}) = 0.5$

compute $p(Y=\text{ham} \mid X=\text{yes})$

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\bar{A})P(\bar{A})}$$

$P(\bar{A}) = 1 - P(A)$

$$P(Y=\text{ham} \mid X=\text{yes}) = \frac{P(X=\text{yes} \mid Y=\text{ham})\,P(Y=\text{ham})}{P(X=\text{yes} \mid Y=\text{ham})\,P(Y=\text{ham}) + P(X=\text{yes} \mid Y=\text{spam})\,P(Y=\text{spam})}$$

$$= \frac{0.02 \times 0.5}{0.02 \times 0.5 + 0.5 \times 0.5} = 0.0385$$

2. Label the following as supervised / unsupervised

a. The INFCO supermarket collects information on what its customers buy (via loyalty cards). This gives rise to a purchase profile for each customer. It then groups customers on the basis of these profiles, in order to understand the makeup of its customer base.

Unsupervised. No specific notion of input / output, probably no labeled data. INFCO is learning the structure of the data, not trying to predict which customers are likely to pass a bad check.

b. RASHBANK is an investment bank that uses the recent history of stockmarket to predict future on stock performance.

Supervised. There is an input (historical performance), an output (future performance) and a clear error / objective function (expected risk-adjusted gain).

3. Whizzoo decides to make a text classifier. To begin with they attempt to classify documents as either sport or politics. They decide to represent each document as a (row) vector of attributes describing the presence or absence of words.

$x = [\text{goal}, \text{football}, \text{golf}, \text{defence}, \text{offence}, \text{wicket}, \text{office}, \text{strategy}]$

Training data from sport documents and from politics documents are represented using a matrix in which each row represents a (row) vector of the 8 attributes.

$$xP = \begin{bmatrix} 1&0&1&1&1&0&1&1 \\ 0&0&0&1&0&0&1&1 \\ 1&0&0&1&1&0&1&0 \\ 0&1&0&0&1&1&0&1 \\ 0&0&0&1&1&0&1&1 \\ 0&0&0&1&1&0&0&1 \end{bmatrix} \quad 6/13$$

$$xS = \begin{bmatrix} 1&1&0&0&0&0&0&0 \\ 0&0&1&0&0&0&0&0 \\ 1&1&0&1&0&0&0&0 \\ 1&1&0&1&0&0&0&1 \\ 1&1&0&1&1&0&0&0 \\ 0&0&0&1&0&1&0&0 \\ 1&1&1&1&1&0&1&0 \end{bmatrix} \quad 7/13$$

Using a Naive Bayes classifier, what is the probability that the document $x = [1, 0, 0, 1, 1, 1, 1, 0]$ is about politics?

	goal ✓	football	golf	defence ✓	offence ✓	wicket ✓	office ✓	strategy
politics	2/6	1/6	1/6	5/6	5/6	1/6	4/6	5/6
sport	5/7	5/7	2/7	5/7	2/7	1/7	1/7	1/7

$$\begin{aligned} P([x] \mid \text{politics}) &= 2/6 \times (1-1/6) \times (1-1/6) \times 5/6 \times 5/6 \times 1/6 \times 4/6 \times (1-5/6) \\ &= 2/6 \times 5/6 \times 5/6 \times 5/6 \times 5/6 \times 1/6 \times 4/6 \times 1/6 \\ &= \frac{5000}{1679616} \approx 0.0029769 \end{aligned}$$

$$\begin{aligned} P([x] \mid \text{sports}) &= 5/7 \times (1-5/7) \times (1-2/7) \times 5/7 \times 2/7 \times 1/7 \times 1/7 \times (1-1/7) \\ &= 5/7 \times 2/7 \times 5/7 \times 5/7 \times 2/7 \times 1/7 \times 1/7 \times 6/7 \\ &= \frac{3000}{5764801} \approx 0.000520 \end{aligned}$$

$$P(\text{politics} \mid [x]) = \frac{P([x] \mid \text{politics}) P(\text{politics})}{P([x] \mid \text{politics}) P(\text{politics}) + P([x] \mid \text{sport}) P(\text{sport})}$$

$$= \frac{0.0029769 \times 6/13}{0.0029769 \times 6/13 + 0.000520 \times 7/13} \approx 0.881$$

4. You have a collection of 1000 nature photographs which were taken under many different conditions. All of the images are of size 300 x 300 pixels. You wish to develop a binary classifier that labels a photograph as to whether or not it depicts a sunny day on the beach. The images have been pre-processed in the following manner:

- Each image $i \in \{1, \ldots, 1000\}$ is partitioned to nine regions $R_{i,1} \ldots R_{i,9}$. Each region is 100x100 pixels. The regions are arranged in a 3x3 grid, so that the region $R_{i,1}$ is top left corner of image i, the region R_2 is the top middle portion of the image, and so on.
- For each region $R_{i,j}$, we compute the average hue of the pixels within the region $R_{i,j}$. The hue value is quantised into 7 discrete bins: 'red', 'orange', 'yellow', 'green', 'blue', 'indigo', 'violet'.

a. How would you represent this data in terms of att-val pairs?

Naive answer → 9 categorical attributes $X_1 \ldots X_9$, where the possible values are the color labels.
This would work if there's some sort of 'structure' to the regions, e.g. region R_1 represents the 'sun' region.
In practice, there is no structure/ordering to the regions: R_1 in one image = sun (yellow)
R_1 in other image = sky (blue)

Right answer → attributes will reflect presence/absence of particular colours in an image.

b. How many attributes? Are they categorical/ordinal/numeric?

7 (one for each colour) attributes.
Their values are numeric.

c. What values can they take on?

The values are either binary (present/absent) or integer, if we want to allow repetitions of colours.
e.g. an image containing two 'yellow' regions may be deemed different from an image containing one 'yellow' region.