Introductory Applied Machine Learning, Tutorial Number 1

School of Informatics, University of Edinburgh, Instructor: Nigel Goddard

September 2017

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 this word is not present. Y takes on the values of *ham* and *spam*. This example relates to "spam filtering" for email.

Let $p(Y = ham) = p(Y = spam) = 0.5$, and $p(X = yes|Y = ham) = 0.02$, $p(X = yes|Y = spam) = 0.5$. Compute $p(Y = ham|X = yes)$.

Solution:

$$\begin{aligned} p(Y{=}ham|X{=}yes) &= \frac{p(X{=}\,yes|Y{=}ham)P(Y{=}ham)}{p(X{=}yes|Y{=}ham)P(Y{=}ham) + p(X{=}yes|Y{=}spam)P(Y{=}spam)} && (1) \\ &= \frac{0.02 \times 0.5}{0.02 \times 0.5 + 0.5 \times 0.5} && (2) \\ &= 0.0385 && (3) \end{aligned}$$

If it helps you can put up the joint probability distribution, which is

	$Y = ham$	$Y = spam$
$X = yes$	0.01	0.25
$X = no$	0.49	0.25

2. Label the following situations as either supervised or unsupervised learning:

(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.

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

Solution:

(a) 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 pass a bad check.

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

3. Class conditional probabilities for each word are:

	goal	football	golf	defence	offence	wicket	office	strategy
politics	$\frac{2}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{5}{6}$	$\frac{5}{6}$	$\frac{1}{6}$	$\frac{4}{6}$	$\frac{5}{6}$
sport	$\frac{5}{7}$	$\frac{5}{7}$	$\frac{2}{7}$	$\frac{5}{7}$	$\frac{2}{7}$	$\frac{1}{7}$	$\frac{1}{7}$	$\frac{1}{7}$

Based on the data:

$$\begin{aligned} p(\text{politics}) &= \frac{6}{13} = 0.462, \\ p(\text{sport}) &= \frac{7}{13} = 0.538\,. \end{aligned}$$

For $\mathbf{x} = (1,0,0,1,1,1,1,0)^T$, the document contains the words goal, defence, offence, wicket and office, so:

$$\begin{aligned} p_{\text{NB}}(\mathbf{x}\,|\,\text{politics}) &= \frac{2}{6} \times \frac{5}{6} \times \frac{5}{6} \times \frac{5}{6} \times \frac{5}{6} \times \frac{1}{6} \times \frac{4}{6} \times \frac{1}{6} = \frac{5000}{1679616} = 0.0029769 \\ p_{\text{NB}}(\mathbf{x}\,|\,\text{sport}) &= \frac{5}{7} \times \frac{2}{7} \times \frac{5}{7} \times \frac{5}{7} \times \frac{2}{7} \times \frac{1}{7} \times \frac{1}{7} \times \frac{6}{7} = \frac{3000}{5764801} = 0.000520\,, \end{aligned}$$

and therefore:

$$p(\text{politics}\,|\,\mathbf{x}) = \frac{p(\text{politics})p(\mathbf{x}\,|\,\text{politics})}{p(\text{politics})p(\mathbf{x}\,|\,\text{politics}) + p(\text{sport})p(\mathbf{x}\,|\,\text{sport})} = 0.831\,.$$

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

- Each image $i \in \{1 \ldots 1000\}$ is partitioned nine regions $R_{i,1} \ldots R_{i,9}$. Each region is 100×100 pixels. The regions are arranged in a 3×3 grid, so that the region R_{i1} is the top-left corner of image i, the region R_{i2} is the top middle portion of the image, and so on.
- For each region $R_{i,j}$, we compute the average *hue*[1] of pixels within the region $R_{i,j}$. The hue value is quantised into 7 discrete bins: "red", "orange", "yellow", "green", "blue", "indigo" and "violet".

(a) How would you represent this data in terms of attribute-value pairs?

(b) How many attributes are there? Are they categorical, ordinal or numeric?

(c) What values can they take on?

Solution: The naive (and incorrect) solution is to use 9 categorical attributes $X_1 \ldots X_9$, where the possible values are the colour labels. This would work if there was a natural "structure" to the regions, e.g. if region R_1 represented the same thing in all images (e.g. the "sun" region). In practice, there is no structure or ordering to the regions: in one image the top-left region R_1 might contain the "sun" while in another R_1 could contain clear blue sky.

The correct answer is:

(a) Attributes will reflect presence or absence of particular colours in the image.

(b) There are 7 attributes (one per colour value), their values are numeric.

(c) The values are either binary (presence / absence) 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.

[1] The *hue* is a scalar representation of color. It ranges from 0° to 360°. For example, colors with hues around 0° look red, hues around 120° look blue, and hues around 240° look green.