Introductory Applied Machine Learning, Tutorial Number 1

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

September 2016

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 $p(Y = ham|X = yes)$.

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.

3. Whizzco decide 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.

$$\boldsymbol{x} = (\text{goal}, \text{football}, \text{golf}, \text{defence}, \text{offence}, \text{wicket}, \text{office}, \text{strategy}) \tag{1}$$

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

```
xP=[1 0 1 1 1 0 1 1;  % Politics
    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]

xS=[1 1 0 0 0 0 0 0; % Sport
    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]
```

Using a Naive Bayes classifier, what is the probability that the document $\boldsymbol{x} = (1, 0, 0, 1, 1, 1, 1, 0)$ is about politics?

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?

[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.