# Applied Machine Learning: Tutorial Number 1

School of Informatics, University of Edinburgh
Instructors: Oisin Mac Aodha and Siddharth N.

September 2022

1. This example relates to "spam filtering" for email. 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*.

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. Give two other examples of supervised learning problems.

4. 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 vector of features describing the presence or absence of words.

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

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

```
% Politics
xP=[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]
```

```
% Sport
xS=[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]
```

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?

5. A training set consists of one dimensional examples from two classes. The training examples from class 1 are $\{0.5, 0.1, 0.2, 0.4, 0.3, 0.2, 0.2, 0.1, 0.35, 0.25\}$ and from class 2 are $\{0.9, 0.8, 0.75, 1.0\}$. Fit a (one dimensional) Gaussian using Maximum Likelihood to each of these two classes. You can assume that the variance for class 1 is 0.0149, and the variance for class 2 is 0.0092. Also estimate the class prior probabilities using Maximum Likelihood.

What is the probability that the test point $x = 0.6$ belongs to class 1? Does this answer seem sensible given the observed data?