Applied Machine Learning: Tutorial Number 1 - Solutions

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

Solution:

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

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. Give two other examples of supervised learning problems.

Solution:

- Image classification
- Speech to text
- Music genre prediction
- ...

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?

Solution:

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:

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

and therefore:

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

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?

Solution:

The maximum likelihood estimator for the mean of each Gaussian is given by $\frac{\sum_i x_i}{n}$:

$$\begin{aligned} \hat{\mu}_1 &= 0.26\,(\text{add up the 10 numbers and divide by 10}), \\ \hat{\mu}_2 &= 0.8625\,(\text{add up the 4 numbers and divide by 4}), \end{aligned}$$

with variances as in the question:

$$\begin{aligned} \hat{\sigma}_1^2 &= 0.0149, \\ \hat{\sigma}_2^2 &= 0.0092\,. \end{aligned}$$

Class probabilities are:

$$\begin{aligned} p(c_1) &= \frac{10}{14} = 0.7143, \\ p(c_2) &= 1 - p(c_1) = \frac{4}{14} = 0.2857\,. \end{aligned}$$

Now, the probability that a point x belongs to class 1 is given by:

$$p(c_1 \mid x) = \frac{p_1 p(x \mid c_1)}{p_1 p(x \mid c_1) + p_2 p(x \mid c_2)}\,,$$

where,

$$p(x \mid c_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left[-\frac{1}{2}\frac{(x-\mu_k)^2}{\sigma_k^2}\right]\,.$$

Crunching the numbers we obtain $p(c_1 \mid x = 0.6) = 0.6305$.

Note that $\hat{\mu}_2$ is nearer to $x = 0.6$ than $\hat{\mu}_1 = 0.26$, but that $\hat{\sigma}_1^2 = 0.0149$ is broader than $\hat{\sigma}_2^2 = 0.0092$. The prior for c_1 is also larger than c_2 which influences the final prediction.