Learning from Data, Tutorial Sheet for week 6

School of Informatics, University of Edinburgh

Instructor: Amos Storkey

1. 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.

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

Training data from sport documents and from politics documents is represented below in MATLAB 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?