MLPM Tutorial 2

October 15, 2019

1. The Poisson distribution is given by

$$Poi(x|\lambda) = e^{-\lambda}\frac{\lambda^x}{x!}$$

   for $x \in \{1, 2, 3, \ldots\}$ where $\lambda > 0$ is the rate parameter. Derive the MLE for λ. What data is typically modelled with the Poisson distribution?

2. A local supermarket specialising in breakfast cereals decides to analyse the buying patterns of its customers. They make a small survey asking 6 randomly chosen people their age (older or younger than 60 years) and which of the breakfast cereals (Cornflakes, Frosties, Sugar Puffs, Branflakes) they like. Each respondent provides a vector with entries 1 or 0 corresponding to whether they like or dislike the cereal. Thus a respondent with (1101) would like Cornflakes, Frosties and Branflakes, but not Sugar Puffs. The older than 60 years respondents provide the following data (1000), (1001), (1111), (0001). The yonger than 60 years old respondents provide the following data (0110), (1110). A novel customer comes into the supermarket and says she only likes Frosties and Sugar Puffs. Using naive Bayes trained with maximum likelihood, what is the probability that she is younger than 60?

3. The company Whizzco decides 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.

$$\mathbf{x} = [\text{goal}, \text{football}, \text{golf}, \text{defence}, \text{offence}, \text{wicket}, \text{office}, \text{strategy}]$$

Here is the training data from sprot and politics documents represented as matrices, where each row represents the 8 attributes:

$$\mathcal{D}_{politics} = \begin{bmatrix} 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 \end{bmatrix}$$

$$\mathcal{D}_{sport} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 & 1 & 0 \end{bmatrix}$$

Help Whizzco and implement a naive Bayes classifier in your favourite programming language. What is the probability that the document $\mathbf{x} = [1, 0, 0, 1, 1, 1, 1, 0]$ is about politics?

4. Consider a generative classifier for C classes with class conditional density $p(\mathbf{x}|y)$ and uniform class prior $p(y)$. Suppose all the D features are binary, $x_j \in \{0, 1\}$.

   (a) If we assume all the features are conditionally independent (the naive Bayes assumption), how many parameters are required to express the class conditional density?

   (b) Now consider a different model, which we will call the "full" model, in which all the features are fully dependent (i.e., we make no factorisation assumptions). How might we represent $p(\mathbf{x}|y = c)$ in this case? How many parameters are needed to represent the class conditional density?

   (c) Assume the number of features D is fixed. Let there be N training cases. If the sample size N is very small, which model (naive Bayes or full) is likely to give lower test set error, and why? How about training set error?

(d) If the sample size N is very large, which model (naive Bayes or full) is likely to give lower test set error, and why? How about training set error?

(e) What is the computational complexity of fitting the full and naive Bayes models as a function of N and D. (Fitting the model here means computing the MLE or MAP paramater estimates.)

(f) What is the computational complexity of applyig the full and naive Bayes models at test time to a single test case?

(g) Suppose the test case has missing data. Let $\mathbf{x}_v$ be the visibile features of size v and $\mathbf{x}_h$ be the hidden (missing) features of size h, where $v + h = D$. What is the computational complexity of computing $p(y|\mathbf{x}_v, \hat{\theta})$ for the full and naive Bayes models, as a function of v and h.