University of York

MLAP

Machine Learning and Applications

Open Examination

Examination number:

Y3606797

Contents

1 Task 1

1.1 What sort of model?

The naive Bayes model makes the assumption that all features are conditionally independent given the class. The given Bayesian network explicitly models the conditional independence in it's structure.

1.2 EM in general

The EM algorithm is an iterative approach to maximising the likelihood when we have hidden variables. We have a model $p(v, h|\theta)$ and we wish to find θ by maximising the marginal likelihood $p(v|\theta)$. EM achieves this by replacing he marginal likelihood with a lower bound, the bound depends on θ and the set of variational distributions $\{q\}$, we optimise this bound. This is done by first fixing θ and then optimising $w.r.t$ $\{q\}$, then fixing $\{q\}$ and optimising $w.r.t$ θ. These are known as the 'E' and 'M' steps and are repeated until convergence. The EM algorithm may not always return the correct MLEs because it can get stuck on local optima.

1.3 EM for the current model

Firstly we assume some initial parameters for the distribution, so we have θ^0_{X1}, θ^0_{X2}, θ^0_{X3}, θ^0_H and θ^0_C. Then we compute the distribution for the hidden data H:

$$q^{n=1}_{t=1}(H) = p(H|C = 0, X1 = 0, X2 = 0, X3 = 1, \theta^0),$$

$$q^{n=2}_{t=1}(H) = p(H|C = 1, X1 = 1, X2 = 0, X3 = 0, \theta^0)...$$

and so on for each data point in the training data. This is first E-step in the algorithm.

The first M-step involves maximising the energy term by choosing new values for θ. The energy term is given by the following equation:

$$E(\theta) = \sum_{n=1}^{N} \{\langle \log p(C^n|H^n)\rangle_{q^n_t(H)} + \langle \log p(X1^n|H^n)\rangle_{q^n_t(H)} + \langle \log p(X2^n|H^n)\rangle_{q^n_t(H)} + \langle \log p(X3^n|H^n)\rangle_{q^n_t(H)} + \langle \log p(H^n)\rangle_{q^n_t(H)}\} \quad (1)$$

To calculate the M-step update for each table we have to maximise each term of the above equation individually. The contribution of the term $p(C = i|H = j)$ is given by:

$$\sum_n \mathbb{I}[C^n = i] q^n(H = j) \log p(C = i|H = j))$$

where the indicator function $\mathbb{I}[C^n = i]$ equals 1 if C is in state i and is zero otherwise. We normalise the table by adding a Lagrange term:

$$\sum_n \mathbb{I}[C^n = i] q^n(H = j) \log p(C = i|H = j)) + \lambda\{1 - \sum_k p(C = k|H = j)\}$$

We differentiate with respect to $p(C = i|H = j)$ and equate to zero to find the maximum, this gives us:

$$\sum_n \mathbb{I}[C^n = i] \frac{q^n(H = j)}{p(C = i|H = j)} = \lambda$$

Hence

$$p(C = i|H = j) = \frac{\sum_n \mathbb{I}[C^n = i] q^n(H = j)}{\sum_{n,k} \mathbb{I}[C^n = k] q^n(H = j)}$$

This result highlights a relationship between the table updates and the standard MLE. That is, when we have missing data we replace the functions such as $\mathbb{I}[C^n = i]$ which represent the real counts of the data, by the assumed distributions q. We use this relationship to calculate the table updates for each other term in the energy equation.

We can then perform the E-step again using our updated parameters such that we have new assumed distributions for the hidden variable:

$$q_t^{n=1}(H) = p(H|C = 0, X1 = 0, X2 = 0, X3 = 1, \theta^{t-1}),$$

$$q_t^{n=2}(H) = p(H|C = 1, X1 = 1, X2 = 0, X3 = 0, \theta^{t-1})...$$

We perform the steps E and M iteratively until the algorithm converges to a local optima.