MLMI MPhil

Module MLMI1: Introduction to Machine Learning

Example Sheet 2: Clustering, the Expectation Maximisation Algorithm, Markov Models, and Hidden Markov Models

Straightforward questions are marked †
Hard questions are marked *

Clustering

1. K-means clustering*

   Consider the K-means algorithm that seeks to minimise the cost function

$$C = \sum_{n=1}^{N} \sum_{k=1}^{K} s_{nk} ||x_n - m_k||^2$$

   where m_k is the mean (centre) of cluster k, x_n is the data point n, $s_{nk} = 1$ indicates that the nth data point has been assigned to cluster k, and $s_{nk} = 0$ indicates that it has not been assigned to that cluster. There are N data points and K clusters.

   (a) Given all the assignments $\{s_{nk}\}$ determine the value of m_k that minimises the cost C and give an interpretation in terms of the K-means algorithm.

   (b) You would like to automatically learn the number of clusters K from data. One possibility is to minimise the cost C as a function of K. Explain whether this is a good idea or not, and what the solution to this minimisation is.

   (c) In many real-world applications, data points arrive sequentially and one wants to cluster them as they come in. Devise a sequential variant of the k-means algorithm which takes in one data point at a time and updates the means $\{m_1, ..., m_K\}$ sequentially without revisiting previous data points. Describe your sequential algorithm.

2. The KL Divergence

The KL Divergence between two discrete distributions $p(x = k) = p_k$ and $p(x = k) = q_k$ is defined as $\mathcal{KL}(q, p) = \sum_{k=1}^{K} q_k \log \frac{q_k}{p_k}$.

(a) Prove that the KL Divergence is non-negative and that it attains its unique minimum when $q_k = p_k$.

(b) A machine learner has a target distribution $\mathbf{p} = [p_1, p_2, p_3, p_4, p_5, p_6] = [1, 1, 0, 0, 1, 1]/4$ which they want to fit with approximating distribution q. The choices for q available to the machine learner are: $q_1 = [1, 1, 0, 0, 0, 0]/2$, $q_2 = [0, 1, 1, 0, 0, 0]/2$, $q_3 = [0, 0, 1, 1, 0, 0]/2$, $q_4 = [0, 0, 0, 1, 1, 0]/2$, $q_5 = [0, 0, 0, 0, 1, 1]/2$, and $q_6 = [1, 1, 1, 1, 1, 1]/6$.

i. Determine the distribution(s) q_i that minimises $\mathcal{KL}(q_i, p)$. Comment on your result.

ii. Determine the distribution(s) q_i that minimises $\mathcal{KL}(p, q_i)$. Comment on your result.

Note that, by convention, $0 \times \log 0 = 0$ since $\lim_{\Delta \to 0} \Delta \log \Delta = 0$.

The EM Algorithm

3. Factor Analysis and EM*

The noisy depth sensor from question 2 in example sheet 1 is used to collect measurements of the distances to a set of N objects that are unknown distances d_n metres away. The object depths can be assumed, *a priori*, to be distributed according to independent standard Gaussian distributions $p(d_n) = \mathcal{N}(d_n; 0, 1)$. As before, the depth sensor returns y_n a noisy measurement of the depth, that is also assumed to be Gaussian $p(y_n|d_n, \sigma_y^2) = \mathcal{N}(y_n; d_n, \sigma_y^2)$.

The variance of the σ_y^2 sensor noise is unknown and must be estimated from the measured depths $\{y_n\}_{n=1}^N$ using maximum likelihood.

(a) Derive the steps of the EM algorithm for performing this. Your answer should include the E-Step (you may find your solution to question 2 in example sheet 1 useful here) and the M-Step.

(b) An alternative approach to maximum-likelihood learning would optimise the log-likelihood $p(\{y_n\}_{n=1}^N|\sigma_y^2)$ directly. Compute the objective function for this procedure. How do you think this approach will perform compared to EM?

4. Gaussian Mixture Models and EM** (beyond tripos standard; optional)

A Gaussian Mixture Model for D-dimensional data $\{\boldsymbol{x}_n\}_{n=1}^N$ comprises a categorical distribution over the class membership variables $p(s_n = k|\theta) = \pi_k$ and a general multivariate Gaussian distribution over the observed data given the class membership variables, $p(\boldsymbol{x}_n|s_n = k, \theta) = \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{m}_k, \Sigma_k)$, (i.e. Σ_k is not isotopic). The posterior distribution over the class membership variables has been computed $p(s_n = k|\boldsymbol{x}_n, \theta) = q_{n,k}$.

(a) Derive the M-Step update of the EM algorithm, i.e. the formulae for updating the parameters $\{\pi_k, \boldsymbol{m}_k, \Sigma_k\}_{k=1}^K$ using the posterior probabilities, $q_{n,k}$.

(b) A friend suggests that it might be possible to speed up the EM algorithm by updating the posterior distribution of only a subset of $K < N$ data points during the E-Step that are selected uniformly at random each time, and then updating the parameters using the same expressions derived in part (a).

i. Will this procedure converge to a (local) optimum of the likelihood?

ii. Will this partial E-Step update procedure be computationally more efficient?

You may find the following identities useful,

$$\frac{\mathrm{d}}{\mathrm{d}\alpha} \log\det(\Sigma) = \mathrm{trace}\left(\Sigma^{-1}\frac{\mathrm{d}\Sigma}{\mathrm{d}\alpha}\right), \quad \frac{\mathrm{d}}{\mathrm{d}\alpha}\Sigma^{-1} = -\Sigma^{-1}\frac{\mathrm{d}\Sigma}{\mathrm{d}\alpha}\Sigma^{-1}.$$

Markov Models

5. Markov Models: fitting bi-gram models

   A data scientist observes part of a long sequence that contains $K = 3$ characters: $ABAAABBABCCCBC$. She would like to use a bi-gram model to fit the data with parameters $p(y_1 = k|\theta) = \pi_k^0$ and $p(y_t = k|y_{t-1} = l, \theta) = T_{k,l}$.

   (a) Write down the log-likelihood for the model and optimise it with respect to π^0 and T to find the maximum likelihood parameter estimates.

   (b) Is the maximum-likelihood estimate sensible? How might you improve the estimate?

6. Markov Models: Gaussian AR(1) models

   A data scientist observes a sequence of scalar variables $y_{1:T} = \{y_t\}_{t=1}^T$ generated from a Gaussian AR(1) process $y_t = \lambda y_{t-1} + \epsilon_t$ where $\epsilon_t \sim \mathcal{N}(\mu, \sigma^2)$.

   She knows that the invariant distribution of the process has the following properties

$$\lim_{t\to\infty} \mathbb{E}(y_t) = \mu_\infty, \quad \lim_{t\to\infty} \mathbb{E}(y_t^2) = \sigma_\infty^2 + \mu_\infty^2, \quad \lim_{t\to\infty} \mathbb{E}(y_t y_{t-1}) = \alpha_\infty.$$

   (a) Derive the parameters of the Gaussian AR(1) process $\{\lambda, \mu, \sigma^2\}$ in terms of the properties of the invariant distribution $\{\mu_\infty, \sigma_\infty^2, \alpha_\infty\}$.

   (b) The data scientist reinterprets the original Markov model in terms of a new random variable z_t such that $y_t = z_t + \mu_\infty$. State the form of the distribution $p(z_{1:T})$ required for this model to be equivalent to the original one.

Hidden Markov Models

7. Discrete Valued Hidden Markov Models*

   (a) Provide the probabilistic equations that define a Hidden Markov Model (HMM) for observed data that takes discrete values. Indicate what aspects of the model the following terms refer to: *initial state probabilities*, *transition matrix* and *emission matrix*.

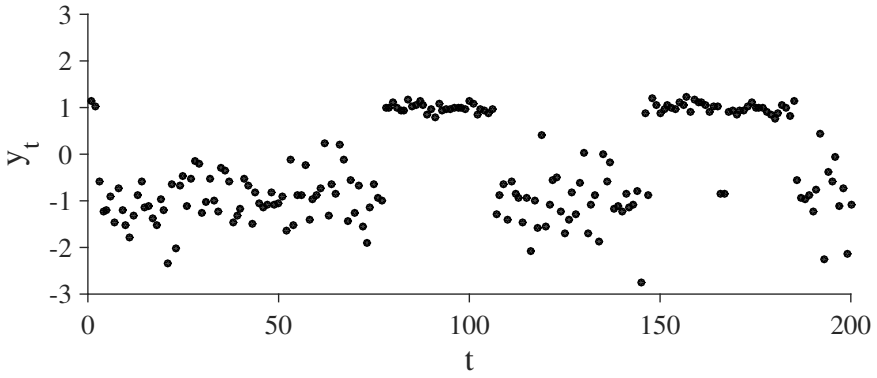
   (b) Consider a dataset consisting of the following string of 160 symbols from the alphabet $\{A, B, C\}$:

   $$AABBBACABBBACAAAAAAAAAABBBACAAAAABACAAAAAA$$
   $$BBBBACAAAAAAAAAAAAABACABACAABBACAAABBBBACA$$
   $$AABACAAAABACAABACAAABBACAAAABBBBACABBACAA$$
   $$AAAABACABACAAABACAABBBACAAAABACABBACA$$

   Carefully analyse the string. Describe an HMM model for the string. Your description should include the number of states in the HMM, the transition matrix including the values of the elements of the matrix, the emission matrix including the values of its elements, and the initial state probabilities. Explain your reasoning.

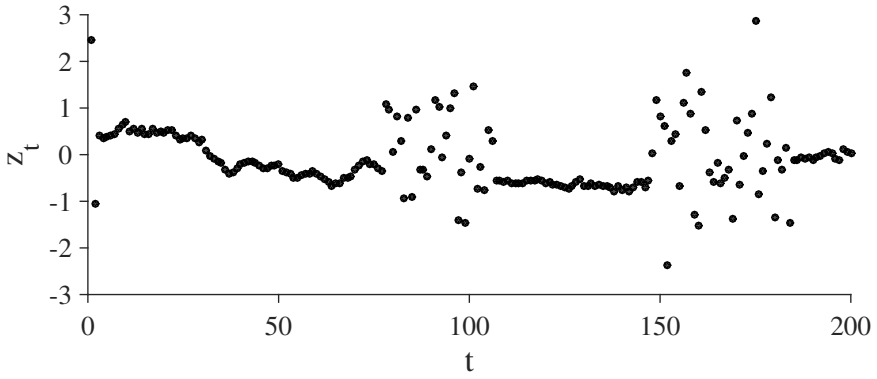
8. Probabilistic Modelling using HMMs for continuous valued observations

(a) A machine learner observes the time-series, y_t, shown below:



Suggest a suitable Hidden Markov Model (HMM) for this sequence and state the model's probabilistic equations. Indicate plausible numerical values for the parameters where possible.

(b) The machine learner is provided with a second set of observations z_t that were measured simultaneously with y_t, shown below:



Extend the HMM you proposed for part (a) so that it can jointly model the first and second set of observations.

9. Inference in HMMs with Discrete Hidden States[†]

A Hidden Markov Model contains a discrete hidden state variable x_t that takes one of two values and a discrete observed state y_t that also takes one of two values. The hidden state has a transition probability,

$$\begin{bmatrix} P(x_t = 1|x_{t-1} = 1) & P(x_t = 1|x_{t-1} = 2) \\ P(x_t = 2|x_{t-1} = 1) & P(x_t = 2|x_{t-1} = 2) \end{bmatrix} = \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix} = \begin{bmatrix} 2/3 & 1/3 \\ 1/3 & 2/3 \end{bmatrix}.$$

The filtering distribution at time $t-1$ is

$$P(x_{t-1}|y_{1:t-1}) = \begin{bmatrix} P(x_{t-1} = 1|y_{1:t-1}) \\ P(x_{t-1} = 2|y_{1:t-1}) \end{bmatrix} = \begin{bmatrix} 1/4 \\ 3/4 \end{bmatrix}.$$

(a) Compute the predictive distribution for the next hidden state variable, $P(x_t|y_{1:t-1})$.

(b) Explain how your solution to part (a) can be used to compute the filtering distribution $P(x_t|y_{1:t})$. What additional piece of information would you require to carry out this computation?

10. Forecasting in Linear Gaussian State Space Models[*]

A simple linear Gaussian state space model with scalar hidden state variables x_t has been used to model scalar observations y_t,

$$p(x_t|x_{t-1}, \lambda, \sigma^2) = \mathcal{N}(x_t; \lambda x_{t-1}, \sigma^2), \quad p(y_t|x_t, \sigma_y^2) = \mathcal{N}(y_t; x_t, \sigma_y^2).$$

The Kalman filter recursions have been used to process T observations, $y_{1:T}$, in order to return the posterior distribution over the Tth latent state, $p(x_T|y_{1:T}) = \mathcal{N}(x_T; \mu_T, \sigma_T^2)$.

(a) Explain how to transform the posterior distribution over the Tth latent state into a forecast for the observations one time step into the future, i.e. express $p(y_{T+1}|y_{1:T})$ in terms of μ_T and σ_T^2.

(b) Now provide a forecast for the observations τ time steps into the future by expressing $p(y_{T+\tau}|y_{1:T})$ in terms of μ_T and σ_T^2.

(c) What happens to $p(y_{T+\tau}|y_{1:T})$ as $\tau \to \infty$?

Selected solutions and hints

1. a) $\mathbf{m}_k = \sum_{n=1}^N s_{n,k}\mathbf{x}_n / \left(\sum_{n=1}^N s_{n,k}\right)$

2. b) $\{\mathcal{KL}(q_i, p)\}_{i=1}^6 = \{\log 2, \infty, \infty, \infty, \log 2, \infty\}$,
   c) $\{\mathcal{KL}(q_i, p)\}_{i=1}^6 = \{\infty, \infty, \infty, \infty, \infty, \log 3/2\}$

3. a) E-Step: $q_n(d_n) = \mathcal{N}(\mu_{d_n|y_n}, \sigma^2_{d_n|y_n})$ where $\mu_{d_n|y_n} = y_n/(1+\sigma_y^2)$, $\sigma^2_{d_n|y_n} = \sigma_y^2/(1+\sigma_y^2)$
   M-Step: $\sigma_y^2 = \frac{1}{N}\sum_{n=1}^N y_n^2 + \frac{1}{N}\sum_n(\mu^2_{d_n|y_n} + \sigma^2_{d_n|y_n}) - \frac{2}{N}\sum_n y_n \mu_{d_n|y_n}$

4. a) $\boldsymbol{m}_k = \frac{\sum_n q_{n,k}\boldsymbol{x}_n}{\sum_n q_{n,k}}$ $\Sigma_k = \frac{\sum_n q_{n,k}(\boldsymbol{x}_n - \boldsymbol{m}_k)(\boldsymbol{x}_n - \boldsymbol{m}_k)^\top}{\sum_n q_{n,k}}$ $\pi_k = \frac{1}{N}\sum_n q_{n,k}$
   b) i) consider the free-energy before and after the partial update, ii) consider what happens in the limit $N \to \infty$ and whether it is necessary to see all of the data before you make the first M-Step.

5. a) $\log p(y_{1:T}|\theta) = \log \pi_1^0 + 2\log(T_{11}T_{12}T_{32}T_{33}) + 3\log T_{21} + \log(T_{22}T_{23})$

$$T = \begin{bmatrix} 2/5 & 2/5 & 0 \\ 3/5 & 1/5 & 1/3 \\ 0 & 2/5 & 2/3 \end{bmatrix}, \quad \pi^0 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

6. a) $\lambda = (\alpha_\infty - \mu_\infty^2)/\sigma_\infty^2$, $\mu = \mu_\infty(1 + (\mu_\infty^2 - \alpha_\infty)/\sigma_\infty^2)$, $\sigma^2 = \sigma_\infty^2\left(1 - \left(\frac{\alpha_\infty - \mu_\infty^2}{\sigma_\infty^2}\right)^2\right)$

7. b) pay close attention to repeated patterns and remember that some parts of a HMM can be deterministic

8. b) consider whether the low variance z_t regions are correlated through time and whether a standard HMM could model this

9. a) $p(x_t|y_{1:t-1}) = [5, 7]^\top/12$

10. b) $p(y_{T+\tau}|y_{1:T}) = \mathcal{N}(y_{T+\tau}; \lambda^\tau \mu_T, \sigma_y^2 + \lambda^{2\tau}\sigma_T^2 + \sigma^2\sum_{t'=0}^{\tau-1}\lambda^{2t'})$

Richard E. Turner