Version RET: final

MACHINE LEARNING AND MACHINE INTELLIGENCE MPHIL

---

Friday 1st November 2019 2 to 3.45

---

**MLMI1**

**INTRODUCTION TO MACHINE LEARNING**

*Answer all questions.*

*All questions carry the same number of marks.*

*The **approximate** percentage of marks allocated to each part of a question is indicated in the right margin.*

*Write your candidate number **not** your name on the cover sheet.*

**STATIONERY REQUIREMENTS**
Single-sided script paper

**SPECIAL REQUIREMENTS TO BE SUPPLIED FOR THIS EXAM**

## You may not start to read the questions printed on the subsequent pages of this question paper until instructed to do so.

1 (a) Define Bayes' rule and explain how to use it to infer an unobserved variable, denoted $c$, from an observed variable, denoted $d$. (You may assume that $c$ and $d$ are both discrete valued variables.) [30%]

(b) A game involves a player secretly tossing a fair coin. If the coin lands tails, $c = 0$, then a fair six sided die is rolled and the resulting number is reported to the other player. If the coin lands heads, $c = 1$, then two fair six sided dice are rolled and the sum of the two resulting numbers is reported. The reported number is denoted $d$.

A player reports the number $d = 6$. What is the probability that the coin landed heads? Explain your reasoning. [70%]

2 An online advertising company records the time it takes for customers to click on an online advert. The times $t_n$ are measured relative to when each customer first views the advert in their web browser and are therefore positive scalars. The times from $N$ customers are collected into a dataset $\{t_n\}_{n=1}^N$. The company would like to model these click times using an exponential distribution

$$p(t_n|\lambda) = \frac{1}{\lambda}\exp(-t_n/\lambda).$$

Here $\lambda$ is a positive parameter that must be learned from data. Each click $t_n$ is assumed to be drawn independently from this distribution.

(a) Define the maximum likelihood estimate for the parameter $\lambda$. [30%]

(b) Compute the maximum likelihood estimate of $\lambda$. [60%]

(c) The company records the click times from a large number of customers. Data are continuously added to the dataset over time. For privacy and storage capacity reasons they would like to avoid storing the individual data points $\{t_n\}_{n=1}^N$. What can they store instead, if they still want to compute the exact maximum likelihood estimate of $\lambda$ as more data arrive? [10%]

3 A single observed scalar variable $y$ is generated by adding standard Gaussian noise $\varepsilon$ to an unobserved parameter $\mu$, that is

$$y = \mu + \varepsilon \quad \text{where} \quad p(\varepsilon) = \mathcal{N}(\varepsilon; 0, 1).$$

The unobserved parameter is drawn from a prior distribution which is a zero mean Gaussian prior distribution with variance $\sigma_\mu^2$, that is

$$p(\mu|\sigma_\mu^2) = \mathcal{N}(\mu; 0, \sigma_\mu^2).$$

(a) Define the *maximum a posteriori* (MAP) estimate of the unobserved parameter $\mu$. [30%]

(b) Compute the MAP estimate of the parameter $\mu$. [50%]

(c) When will the MAP estimate of $\mu$ be equal to the *maximum likelihood estimate*? [20%]

Here, and later in the exam, we have used the following notation to indicate univariate Gaussian distributions:

$$\mathcal{N}(z; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(z-\mu)^2\right).$$

4 A dataset comprises scalar, real valued inputs $x$ at which two outputs $y_1$ and $y_2$ are measured. The first output $y_1$ is binary. The second output $y_2$ is real valued. The full dataset $\{x_n, y_{1,n}, y_{2,n}\}_{n=1}^{N}$ is shown in Fig. 1.

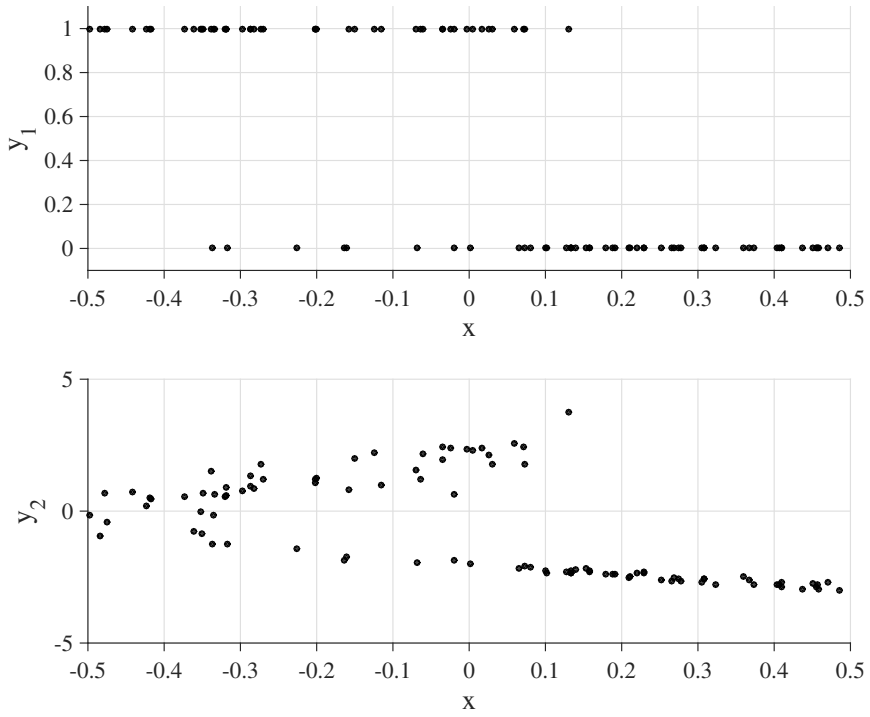


Fig. 1

(a) Consider only the first output $y_1$ as a function of the input $x$ which is shown in the top panel in the figure. Suggest a model for predicting the output from the input, that is $p(y_1|x,\theta)$. Explain your reasoning. [30%]

(b) Now consider the second output $y_2$ shown in the bottom panel in the figure. Suggest a suitable model for $p(y_2|y_1,x,\theta)$. Explain your reasoning. [70%]

5 The k-means algorithm is applied to a dataset using the Euclidean distance. The dataset comprises $N = 6$ data points $\{\mathbf{x}_n\}_{n=1}^{N}$. Each data point $\mathbf{x}_n$ is two dimensional and has real valued elements. These data are plotted in Fig. 2.

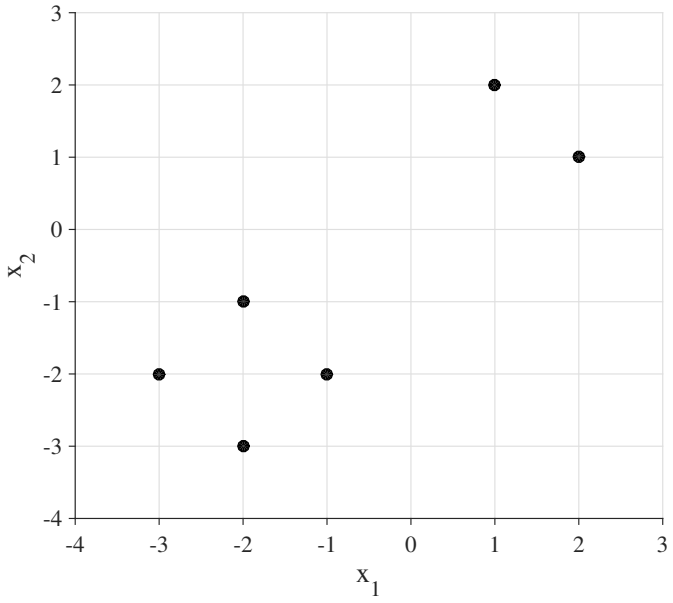


Fig. 2

(a) Describe the *assignment* and *update steps* of the k-means algorithm. [30%]

(b) K-means is run on the data set using $K = 2$ clusters. The initial value of the cluster centres are $\mathbf{m}_1 = [-3, -3]$ and $\mathbf{m}_2 = [-1, -1]$. By drawing annotated plots based on Fig. 2, describe what happens after i) first assignment step, ii) the first update step, iii) the second assignment step, and iv) the second update step. Your answer should state the value of the cluster centres, $\mathbf{m}_1$ and $\mathbf{m}_2$, after each update step. [70%]

6 A discrete valued variable $x$ can take values $x = 1$, $x = 2$, or $x = 3$. $P(x)$ and $Q(x)$ are two distributions over $x$. The KL divergence between $P(x)$ and $Q(x)$ is defined as

$$\mathrm{KL}(P(x)||Q(x)) = \sum_{k=1}^{3} P(x=k) \log \frac{P(x=k)}{Q(x=k)}.$$

$Q(x)$ is restricted to be of the following form: $Q(x=1) = Q(x=3) = \alpha$ and $Q(x=2) = 1-2\alpha$. The parameter $\alpha$ may be varied.

(a) Consider the case where $P(x=1) = 0.1$, $P(x=2) = 0.8$, and $P(x=3) = 0.1$. What value of $\alpha$ minimises the KL divergence? Explain your reasoning. [30%]

(b) Consider the case where $P(x=1) = 0.6$, $P(x=2) = 0.2$, and $P(x=3) = 0.2$. What value of $\alpha$ minimises the KL divergence now? Justify your solution mathematically. [70%]

7 A *Hidden Markov Model* has a binary hidden state $x_t$ and a continuous scalar observed state $y_t$. The initial hidden state and transition distribution are given by,

$$\begin{bmatrix} p(x_1=1) \\ p(x_1=0) \end{bmatrix} = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}, \quad \begin{bmatrix} p(x_t=1|x_{t-1}=1) & p(x_t=1|x_{t-1}=0) \\ p(x_t=0|x_{t-1}=1) & p(x_t=0|x_{t-1}=0) \end{bmatrix} = \begin{bmatrix} 0.8 & 0.2 \\ 0.2 & 0.8 \end{bmatrix}.$$

The observations are generated by Gaussian distributions whose mean depends on the hidden state and whose variances are equal to 1,

$$p(y_t|x_t) = \mathcal{N}(y_t; 2x_t - 1, 1).$$

(a) Compute the joint distribution over the first two hidden state variables, $p(x_1, x_2)$. [30%]

(b) Compute the joint distribution over the first two observed states $p(y_1, y_2)$ using your solution to part (a). [60%]

(c) What type of distribution is the joint distribution over all observed variables $p(y_1, y_2, \ldots, y_T)$? Explain your reasoning. [10%]

8 A public bike hire scheme has three locations at which bikes can be collected and returned. The probabilities of a bike being picked up from location $s_{t-1} = k$ and returned to location $s_t = l$ are given by the following matrix,

$$\begin{bmatrix} p(s_t = 1|s_{t-1} = 1) & p(s_t = 1|s_{t-1} = 2) & p(s_t = 1|s_{t-1} = 3) \\ p(s_t = 2|s_{t-1} = 1) & p(s_t = 2|s_{t-1} = 2) & p(s_t = 2|s_{t-1} = 3) \\ p(s_t = 3|s_{t-1} = 1) & p(s_t = 3|s_{t-1} = 2) & p(s_t = 3|s_{t-1} = 3) \end{bmatrix} = \begin{bmatrix} 0.85 & 0 & 0 \\ 0.1 & 0.85 & 0 \\ 0.05 & 0.15 & 1 \end{bmatrix}.$$

(a) Draw the *state transition diagram* for this system. [30%]

(b) By inspection of the state transition diagram determine the *stationary distribution* of the system. Discuss implications for the bike hire scheme. [30%]

(c) The public bike hire scheme would like there to be, on average, an equal number of bikes at each of the three locations. To do this, they employ people to move bikes from location 3 to locations 1 and 2 which modifies the transition probabilities. Compute the new probabilities $p(s_t = 1|s_{t-1} = 3)$, $p(s_t = 2|s_{t-1} = 3)$, and $p(s_t = 3|s_{t-1} = 3)$ that are required to achieve their goal. [40%]

**END OF PAPER**

THIS PAGE IS BLANK