Version RET: final

MACHINE LEARNING AND MACHINE INTELLIGENCE MPHIL

Friday 9th November 2018 2.30 to 4.15

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 Two scalar variables x and y are drawn from a joint distribution $p(x,y)$. The marginal mean of y is defined as $\mu_y = \int y\, p(y)\mathrm{d}y$. The conditional mean of y given x is defined as $\mu_{y|x}(x) = \int y\, p(y|x)\mathrm{d}y$.

(a) Show how to compute the marginal mean μ_y from the conditional mean $\mu_{y|x}(x)$ and the joint distribution $p(x,y)$. [50%]

(b) An engineer is studying autonomous vehicle accidents. They find that the mean stopping distance given a vehicle's initial velocity v is $\mu_{d|v}(v) = cv^2$ where c is a known constant. The marginal distribution of initial velocities prior to accidents is well approximated by a Gaussian distribution $p(v) = \mathcal{N}(v;\mu_v,\sigma_v^2)$. Compute the marginal mean stopping distance μ_d. [50%]

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

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

2 An engineer has designed a system to localise intruders in a poorly lit corridor that has high background noise levels. The location of the intruder along the corridor is denoted by the scalar quantity x. The system uses an audio sensor and a visual sensor whose scalar readings are denoted a and v respectively. The sensor readings are the underlying location x corrupted by independent Gaussian noise. That is, $p(a|x) = \mathcal{N}(a;x,\sigma_a^2)$ and $p(v|x) = \mathcal{N}(v;x,\sigma_v^2)$ where σ_a^2 and σ_v^2 are the noise variances in the two sensors. *A priori* the intruder position is assumed to follow a zero mean Gaussian with variance σ_0^2, $p(x) = \mathcal{N}(x;0,\sigma_0^2)$.

(a) Explain how to compute the posterior distribution $p(x|a,v)$ from $p(x)$, $p(a|x)$ and $p(v|x)$, using the rules of probability. [30%]

(b) Compute $p(x|a,v)$ by substituting in the distributional forms for $p(x)$, $p(a|x)$ and $p(v|x)$. The product of Gaussians identity below may be useful. [50%]

(c) Explain what happens to the posterior as $\sigma_a^2 \to \infty$ and provide an intuitive explanation for this behaviour. [20%]

Product of Gaussians identity:

$$\prod_{n=1}^{N} \mathcal{N}(z\,;\,\mu_n,\sigma_n^2) \propto \mathcal{N}\left(z\,;\,\mu = \sigma^2 \sum_{n=1}^{N} \frac{\mu_n}{\sigma_n^2}\,,\;\; \sigma^2 = \left(\sum_{n=1}^{N} \frac{1}{\sigma_n^2}\right)^{-1}\right)$$

3 A regression problem comprises scalar inputs x_n and scalar outputs y_n which are linearly related $y_n = mx_n + c + \varepsilon_n$. The noise is Gaussian, with mean 0, but it has a variance that depends quadratically on the input $p(\varepsilon_n) = \mathcal{N}(\varepsilon_n; 0, \alpha x_n^2)$. Due to physical constraints, the inputs lie in the region $1 < x_n < 100$.

The offset c and noise parameter α are known, but the slope m must be learned from a training dataset $\{y_n, x_n\}_{n=1}^N$.

(a) Compute the likelihood of m. [25%]

(b) Compute the maximum likelihood estimate of m. [65%]

(c) You are allowed to select an input location x at which you will be provided with an output y. Which location is most informative about the parameter m? Explain your reasoning. [10%]

4 A machine learning researcher has been asked to develop a system that takes an image $\mathbf{x}$ as input and returns a binary output label y indicating whether a person is present in the image. They are using Amazon Mechanical Turk to label a large image dataset for this purpose. On each trial they show a worker an image and ask them to return a label. The researcher then uses logistic regression to model the responses $P(y = 1|\mathbf{w}, \mathbf{x}) = \sigma(\mathbf{w}^\top \mathbf{x}) = \frac{1}{1+\exp(-\mathbf{w}^\top \mathbf{x})}$ where $\mathbf{w}$ are the model weights.

(a) Plot $P(y = 1|\mathbf{w}, \mathbf{x})$ as a function of $z = \mathbf{w}^\top \mathbf{x}$. [25%]

(b) Unfortunately, there is a problem: the workers are prone to concentration lapses. With probability Δ on each trial, the worker loses concentration and then selects a label uniformly at random. Modify the logistic regression model to incorporate concentration lapse. Plot the new model as a function of $z = \mathbf{w}^\top \mathbf{x}$ showing the differences from the original model. [50%]

(c) In a second experiment the response buttons are altered. The workers lapse with the same probability, but are found to always select $y = 0$ when they lose concentration. Modify the model again and plot it as a function of $z = \mathbf{w}^\top \mathbf{x}$. [25%]

5 The *binary latent feature model* first draws two binary latent variables s_1 and s_2 from independent Bernoulli distributions. That is $p(s_1 = 1|\theta) = \pi_1$ and $p(s_2 = 1|\theta) = \pi_2$. Observations $\mathbf{y}$, which are real valued and D dimensional, are produced by multiplying the latent variables by the associated weights ($\mathbf{w}_1$ and $\mathbf{w}_2$), adding these contributions, and corrupting with isotropic Gaussian noise of variance σ_y^2. That is $p(\mathbf{y}|s_1, s_2, \theta) = \mathcal{N}(\mathbf{y}; \mathbf{w}_1 s_1 + \mathbf{w}_2 s_2, \mathrm{I}\sigma_y^2)$.

Above, the model parameters have been denoted $\theta = \{\pi_1, \pi_2, \mathbf{w}_1, \mathbf{w}_2, \sigma_y^2\}$.

(a) Mathematically define a *mixture of Gaussians model*. Your definition should identify the *mixing proportions*, *component means* and *component variances*. [30%]

(b) Express the binary latent feature model, described at the start of this question, as a mixture of Gaussians stating the *mixing proportions*, *component means* and *component variances* in terms of the parameters θ. [70%]

6 A machine learner will employ the EM algorithm to learn the parameters of the *binary latent feature model* defined in the previous question.

(a) Compute the E-step update by deriving the posterior distribution over the latent variables given an observed variable, $p(s_1, s_2|\mathbf{y}, \theta)$. [50%]

(b) Describe how you would mathematically compute the M-step update. Your description should outline the steps involved, but does not require detailed calculation. [50%]

7 A two dimensional real-valued time series $\{x_t, y_t\}_{t=1}^{T}$ is shown in Fig. 1.

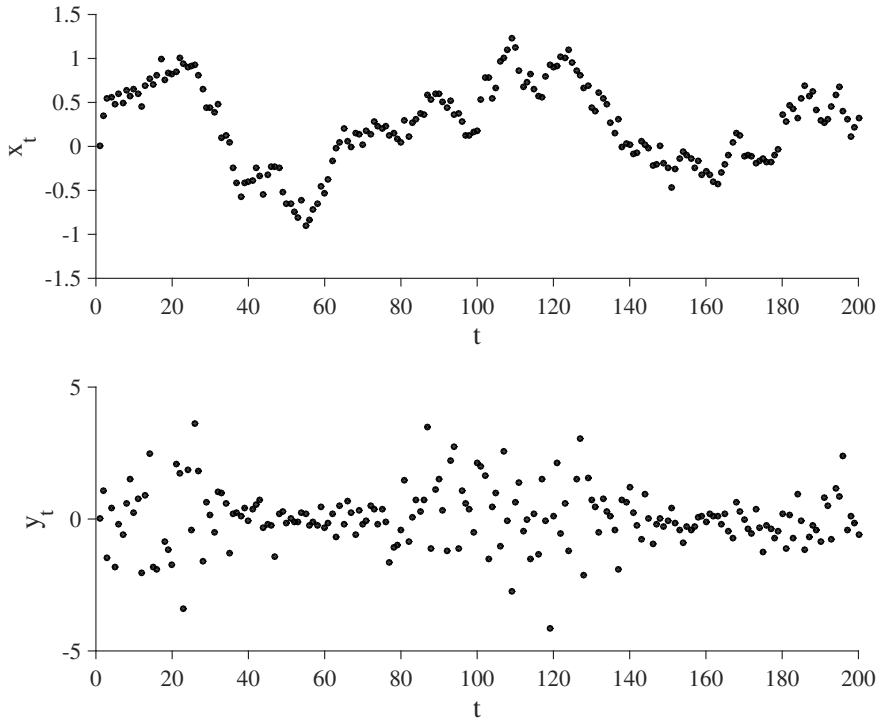

Fig. 1

(a) First, consider $x_{1:T} = \{x_t\}_{t=1}^{T}$. Suggest a first order Markov Model for these data. Explain your reasoning. [50%]

(b) Second, consider $y_{1:T} = \{y_t\}_{t=1}^{T}$. Each variable y_t is produced from the corresponding variable x_t in the plot above. Suggest a suitable model for $p(y_t|x_t)$. Explain your reasoning. [40%]

(c) The model you have devised above can be considered a latent variable model for $y_{1:T}$. Where might such a model be useful? [10%]

8 Consider the following time-series model in which scalar latent variables evolve according to linear Gaussian dynamics

$$x_t = \lambda_1 x_{t-1} + \lambda_2 x_{t-2} + \sigma_x \varepsilon_t \text{ where } p(\varepsilon_t) = \mathcal{N}(\varepsilon_t; 0, 1).$$

The observations are produced by corrupting the latent process with Gaussian noise $y_t = x_t + \eta_t \sigma_y$ where $p(\eta_t) = \mathcal{N}(\eta_t; 0, 1)$.

(a) How does this model differ from the standard linear Gaussian state space model introduced in lectures? [20%]

(b) Rewrite the model into an equivalent form using a two-dimensional hidden state $\mathbf{z}_t = [z_{1,t}, z_{2,t}]^\top$ where

$$\begin{bmatrix} z_{1,t} \\ z_{2,t} \end{bmatrix} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} z_{1,t-1} \\ z_{2,t-1} \end{bmatrix} + \begin{bmatrix} e & f \\ g & h \end{bmatrix} \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{bmatrix}$$

$$y_t = [i, j] \begin{bmatrix} z_{1,t} \\ z_{2,t} \end{bmatrix} + \eta_t \sigma_y$$

and the dynamics noise is standard Gaussian: $p(\varepsilon_{1,t}) = \mathcal{N}(\varepsilon_{1,t}; 0, 1)$ and $p(\varepsilon_{2,t}) = \mathcal{N}(\varepsilon_{2,t}; 0, 1)$.

Your answer should state values for each of the real-valued scalar parameters above $\{a, \ldots, j\}$ in terms of the parameters of the original model $\{\lambda_1, \lambda_2, \sigma_x\}$. It should also describe how the new variables $z_{1,t}$ and $z_{2,t}$ relate to the old ones x_t.

[60%]

(c) You have access to an off-the-shelf implementation of the Kalman filter. Why is it useful to have rewritten the model in this way? [20%]

END OF PAPER

THIS PAGE IS BLANK