Version RET: Final

MACHINE LEARNING, SPEECH & LANGUAGE TECHNOLOGY MPHIL

Wednesday 1st November 2017 11 to 12.45

MLSALT1

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 continuous random variable x is drawn from a distribution $p(x)$ which is uniform over the interval 1 to 2.

(a) Compute the mean and the variance of x. [50%]

(b) N data points drawn from $p(x)$ are fitted with a distribution $q(x|\sigma^2)$ which is a zero mean Gaussian of variance σ^2,

$$q(x|\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}x^2\right).$$

What value does the maximum likelihood estimate for σ^2 tend to as $N \to \infty$? [50%]

2 An urn contains three balls. Each ball is either black or white. Two of the balls are known to be of the same colour and one is known to be a different colour. A ball is pulled out of the urn and found to be black.

(a) What is the probability that the urn originally contained two black balls and one white ball? Explain your reasoning and any assumptions that you make. [75%]

(b) What is the probability that the next ball drawn from the urn, without replacement of the first ball, will be black? Explain your reasoning. [25%]

3 Consider a model in which observed variables y are generated from two binary variables $\{s_1, s_2\}$ and two real valued variables $\{x_1, x_2\}$

$$y = s_1 x_1 + s_2 x_2.$$

The binary variables are independent and Bernoulli distributed with $p(s_1 = 1) = 1/2$ and $p(s_2 = 1) = 1/2$.

The real valued variables are independent and Gaussian distributed with $p(x_1) = \mathcal{N}(x_1; 1, 1)$ and $p(x_2) = \mathcal{N}(x_2; -1, 1)$ where

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

(a) Compute the marginal distribution, $p(y)$, and sketch the distribution as a function of y. [75%]

(b) Describe an application where a model like this would be useful. You may generalise the model if appropriate. [25%]

4 A regression dataset was collected from an industrial system. Each datapoint was collected by an experimenter who probed the system with a real valued input x_n and observed the real valued response y_n. The full dataset is shown on the lefthand side of Fig. 1 and a closeup, or zoom, showing part of the input space is on the right.

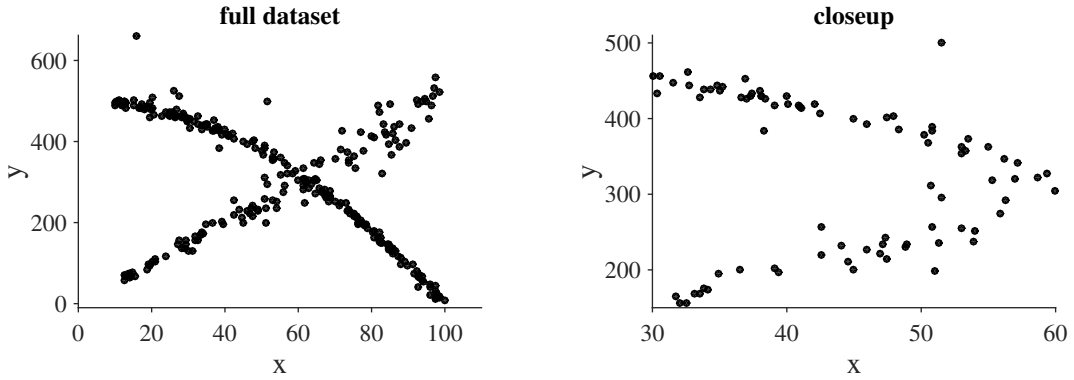

Fig. 1

Suggest a suitable probabilistic model for these data that could be used to predict an output from a new input. Explain your reasoning. [100%]

5 A friend has built a regression system than infers a person's age y (the output) from their height x (the input). They have used a generative model which assumes that the outputs are Gaussian distributed $p(y|\theta) = \mathcal{N}(y; a, \sigma_y^2)$ and that the inputs are a linear transformation of the outputs plus Gaussian noise $p(x|y,\theta) = \mathcal{N}(x; b+wy, \sigma_x^2)$.

Here the model parameters have been denoted $\theta = \{a, \sigma_y^2, b, w, \sigma_x^2\}$ and we have used the notation:

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

(a) Compute the form of the model's predictive distribution for an output y at an input x, that is $p(y|x,\theta)$. [80%]

(b) Compare and contrast the predictive distribution of your friend's regression model to a discriminative model $p(y|x,\theta') = \mathcal{N}(y; c+vx, \sigma^2)$. Will the two models give the same predictions when their parameters are both fit by maximum likelihood learning on a dataset $\{x_n, y_n\}_{n=1}^N$? [20%]

6 Consider a probabilistic model that includes binary latent variables s and binary observed variables x. In the generative model, for each data point, the latent variable is drawn first from a Bernoulli distribution with $p(s=1) = 1/2$. Second, the observed variable is drawn from a Bernoulli distribution whose parameters depend on s according to,

$$\begin{bmatrix} p(x=0|s=0) & p(x=0|s=1) \\ p(x=1|s=0) & p(x=1|s=1) \end{bmatrix} = \begin{bmatrix} T_{00} & T_{01} \\ T_{10} & T_{11} \end{bmatrix} = \mathrm{T}.$$

A dataset $\{x_n\}_{n=1}^N$ is observed and the model parameters are fit using the Expectation Maximisation (EM) algorithm which uses the free-energy

$$\mathcal{F}(\mathrm{T}, \{q_n(s_n)\}_{n=1}^N) = \sum_{n=1}^{N} \sum_{s_n=0}^{1} q_n(s_n) \log \frac{p(x_n, s_n)}{q_n(s_n)}.$$

(a) Compute the explicit form of the free-energy in terms of the model parameters T and variational parameters $q_n = q_n(s_n = 1)$ [50%]

(b) Using the answer to part (b), compute the M-step update equation for T in terms of q_n. [50%]

7 A real-valued biological time series dataset $y_{1:T} = \{y_t\}_{t=1}^{T}$ is shown in Fig. 2.

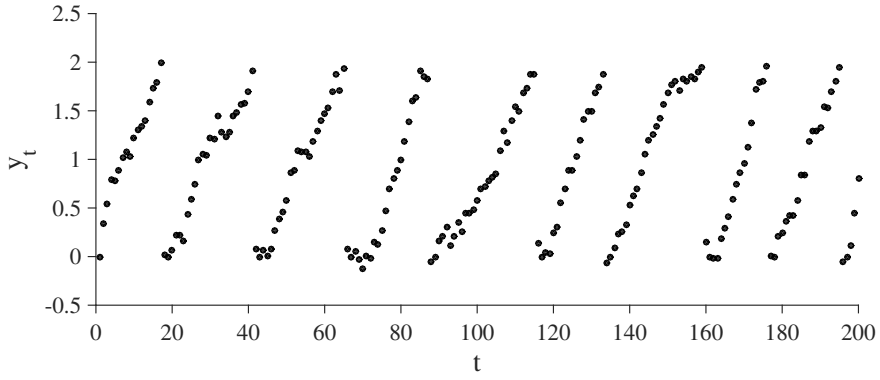

Fig. 2

(a) Suggest a first order Markov Model for these data. Explain your reasoning. [60%]

(b) A friend has code that fits non-linear functions to regression datasets using maximum likelihood estimation. The code takes in a dataset comprising input-output pairs $\{x_n, y_n\}_{n=1}^{N}$ and returns parameters of a very flexible function approximation. She suggests that this can be used to find the maximum likelihood fit of a suitable first order Markov Model for these data too. Explain whether she is correct. [40%]

8 Consider a first-order Gaussian autoregressive process, or AR(1) process for short, for a sequence of scalar variables $x_{1:T}$.

(a) Write down the probabilistic equations that define an AR(1) process. [40%]

(b) Show that the joint distribution over the variables $x_{1:T}$ induced by the AR(1) process is Gaussian, $p(x_{1:T}) = \mathcal{N}(x_{1:T}, \mu, \Sigma)$, and derive the T dimensional mean vector μ and $T \times T$ dimensional precision matrix (inverse covariance matrix) $P = \Sigma^{-1}$. [40%]

(c) Speculate on what property of a probabilistic model leads to a zero entry in an off-diagonal element of a precision matrix, $P_{i,j} = 0$. [20%]

END OF PAPER

THIS PAGE IS BLANK