

MACHINE LEARNING, SPEECH & LANGUAGE TECHNOLOGY MPhil

Wednesday 1st November 2017 11 to 12.45

MLSALT1

INTRODUCTION TO MACHINE LEARNING: CRIB

STATIONERY REQUIREMENTS

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 .

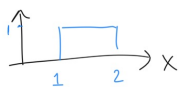
[50%]

(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 $N \rightarrow \infty$?

[50%]

a) $p(x)$ 

$$\mu = \int x p(x) dx = \frac{3}{2} \text{ by symmetry}$$

$$\mathbb{E}(x^2) = \int x^2 p(x) dx = \int_1^2 x^2 dx = \left[\frac{1}{3}x^3\right]_1^2$$

$$= \frac{8}{3} - \frac{1}{3} = \frac{7}{3}$$

$$\sigma^2 = \frac{7}{3} - \left(\frac{3}{2}\right)^2 = \frac{7}{3} - \frac{9}{4} = \frac{28 - 27}{12} = \frac{1}{12}$$

b) $q(x|\sigma^2) = N(x; 0, \sigma^2)$

$$\begin{aligned} \mathcal{L}(\sigma^2) &= \log q(\{x_{i=1:N}\}|\sigma^2) = \sum_{n=1}^N \log q(x_n|\sigma^2) \\ &= -\frac{N}{2} \log 2\pi\sigma^2 - \frac{1}{2\sigma^2} \sum_n x_n^2 \end{aligned}$$

$$\left. \frac{d}{d\sigma^2} \mathcal{L}(\sigma^2) \right|_{\sigma_{ML}^2} = \frac{-N}{2\sigma_{ML}^2} + \frac{1}{2\sigma_{ML}^4} \sum_n x_n^2 = 0$$

$$\therefore \sigma_{ML}^2 = \frac{1}{N} \sum_n x_n^2$$

$$N \rightarrow \infty \quad \sigma_{ML}^2 = \int p(x) x^2 dx = \frac{7}{3} \text{ from part (a)}$$

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 will be black? Explain your reasoning. [25%]

a) hypothesis $H=1$ hypothesis $H=2$

$\boxed{00} \mid \{WWB\}$ $\boxed{000} \mid \{UBB\}$

Observe $X=B$

Bayes' Rule

$$p(H=2 \mid X=B) = \frac{p(X=B \mid H=2) p(H=2)}{p(X=B)}$$

Assume uniform a priori distribution over $H=1$ & $H=2$

$$= \frac{p(X=B \mid H=2) p(H=2)}{p(X=B \mid H=1) p(H=1) + p(X=B \mid H=2) p(H=2)}$$

$$= \frac{2/3}{1/3 + 2/3} = 2/3$$

b) State after 1st ball removed

$H=1$ $1/3$ $H=2$ $2/3$

$\boxed{00} \mid \{WWB\}$ $\boxed{00} \mid \{WBB\}$

Sum rule

$$p(Y=B) = p(Y=B \mid H=1) p(H=1 \mid X=B) + p(Y=B \mid H=2) p(H=2 \mid X=B)$$

$$= 0 \times \frac{1}{3} + \frac{1}{2} \times \frac{2}{3}$$

$$= \frac{1}{3}$$

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%]

$$a) \quad p(y) = \sum_{s_1, s_2} p(y | s_1, s_2) p(s_1, s_2) = \sum_{s_1, s_2} \mathcal{N}(y; \mu_{s_1, s_2}, \sigma_{s_1, s_2}^2) \cdot \frac{1}{4}$$

$$p(s_1, s_2) = \begin{cases} 0 & \text{if } s_1 = s_2 = 0 \\ 1 & \text{if } s_1 = 1, s_2 = 0 \\ -1 & \text{if } s_1 = 0, s_2 = 1 \\ 0 & \text{if } s_1 = 1, s_2 = 1 \end{cases}$$

(means add for independent variables)

$$\sigma_{s_1, s_2}^2 = \begin{cases} 0 & \text{if } s_1 = s_2 = 0 \\ 1 & \text{if } s_1 = 1 \text{ \& } s_2 = 0 \\ 1 & \text{if } s_1 = 0 \text{ \& } s_2 = 1 \\ 2 & \text{if } s_1 = 1 \text{ \& } s_2 = 1 \end{cases}$$

(variances add for independent variables)

This specifies a mixture of Gaussians over y :

$$p(y) = \frac{1}{4} \mathcal{N}(y; 0, 0) + \frac{1}{4} \mathcal{N}(y; 1, 1) + \frac{1}{4} \mathcal{N}(y; -1, 1) + \frac{1}{4} \mathcal{N}(y; 0, 2)$$

— = $p(y)$
— = individual components

b) Lots of possible answers here, but models of the form

$$\underline{y} = \underline{X} \underline{s} + \underline{\varepsilon}$$

\swarrow binary elements
 \nwarrow Gaussian noise
 \uparrow Gaussian elements

are called "latent feature models" & can describe data points that might comprise one or more shared features eg.

\underline{y} = phenotype eg height & width of individuals

\underline{s} = presence/absence of genes

\underline{X} = effect of gene on phenotype, if gene is active in an individual

$\underline{\varepsilon}$ = noise eg due to effect of environment

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 showing part of the input space is on the right.

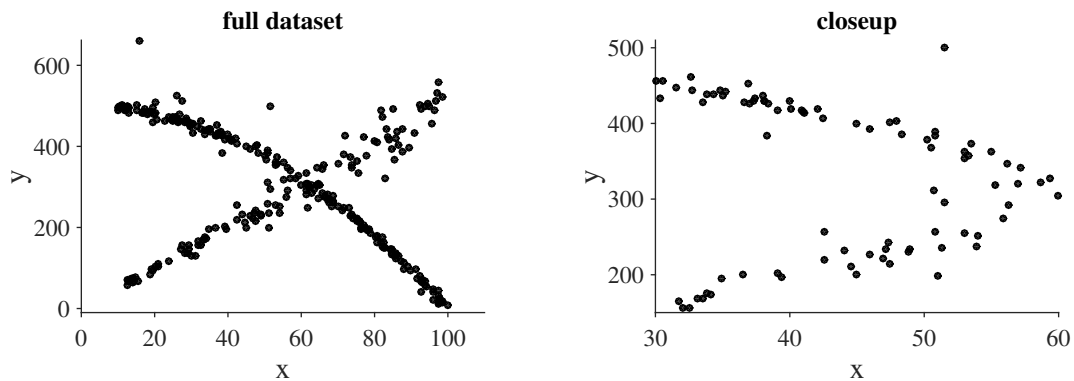


Fig. 1

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

- The model was for reference:
- $$p(y|x) = \pi \mathcal{N}(y; 5x, (\frac{1}{5}x)^2) + (1-\pi) \text{Student}(y; 500 - \frac{x^2}{20}, 16.2)$$
- Handwritten notes:*
- $\pi = 1/2$ (red arrow)
 - $\mathcal{N}(y; 5x, (\frac{1}{5}x)^2)$ (blue arrow)
 - $\text{Student}(y; 500 - \frac{x^2}{20}, 16.2)$ (red arrow)
 - $500 - \frac{x^2}{20}$ (red arrow) \rightarrow mean
 - 16.2 (red arrow) \rightarrow variance
 - $\pi = 1/2$ (blue arrow) \rightarrow I would not expect answer with precisely this form or level of detail. However I would expect the 5 observations below to be made to some level.
 - $\pi = 1/2$ (red arrow) \rightarrow or very near by
 - $\text{Student}(y; 500 - \frac{x^2}{20}, 16.2)$ (red arrow) \rightarrow degree of freedom (ie. 16.2)
- repeatedly probing @ the same input seems to reveal two underlying functions \Rightarrow mixture model at each input location (multimodal underlying function)
 - First component appears linear. Intercept looks close to zero & slope could be estimated to be near to 5.
 - Noise in first component looks Gaussian (not heavy tailed) but the standard deviation appears to grow possibly linearly with x .
 - The second component is non-linear & a polynomial or general basis function model is appropriate.
 - The noise in the second component looks to have a fixed distribution, but there are clear outliers \Rightarrow heavy tailed.

5 A friend has built a regression system that 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:

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

He is calculating posterior. Why in this case predictive distr. = posterior?

- (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%]

Difference between discriminative and generative?

- (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%]

$$\begin{aligned} \text{a) } p(y|x) &= \frac{p(y)p(x|y)}{p(x)} \propto e^{-\frac{1}{2\sigma_y^2}(y-a)^2} e^{-\frac{1}{2\sigma_x^2}(x-b-wy)^2} \\ &= e^{-\frac{1}{2}y^2\left(\frac{1}{\sigma_y^2} + \frac{w^2}{\sigma_x^2}\right) + y\left(\frac{a}{\sigma_y^2} + \frac{w}{\sigma_x^2}(x-b)\right)} \\ &= \mathcal{N}\left(y; \mu_{y|x}, \sigma_{y|x}^2\right) = e^{-\frac{1}{2}y^2\left(\frac{1}{\sigma_y^2} + \frac{w^2}{\sigma_x^2}\right) + y\left(\frac{a}{\sigma_y^2} + \frac{w}{\sigma_x^2}(x-b)\right)} \\ \therefore \sigma_{y|x}^2 &= \frac{\sigma_y^2 \sigma_x^2}{\sigma_x^2 + w^2 \sigma_y^2} \quad \& \quad \mu_{y|x} = \sigma_{y|x}^2 \left[\frac{a}{\sigma_y^2} + \frac{w}{\sigma_x^2}(x-b) \right] \\ \mu_{y|x} &= \frac{a \sigma_x^2}{\sigma_x^2 + w^2 \sigma_y^2} + \frac{\sigma_y^2}{\sigma_x^2 + w^2 \sigma_y^2} w(x-b) \end{aligned}$$

b) Same form as a discriminative model

$$p(y|x) = \mathcal{N}(y | c + vx, \sigma^2)$$

where

$$c = \frac{\sigma_y^2}{\sigma_x^2 + w^2 \sigma_y^2} [a \sigma_x^2 - wb]$$

$$v = \frac{\sigma_y^2}{\sigma_x^2 + w^2 \sigma_y^2} w$$

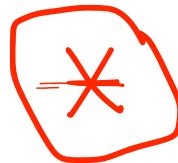
$$\sigma^2 = \frac{\sigma_x^2 \sigma_y^2}{\sigma_x^2 + w^2 \sigma_y^2}$$

But will give different solutions when fit since the generative model is fit to the joint distribution

$$\prod_n p(x_n, y_n | \theta)$$

Whilst the discriminative model is fit to :

$$\prod_n p(y_n | x_n, \theta)$$



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

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

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

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

- (a) Compute the explicit form of the free-energy in terms of the model parameters \mathbf{T} and variational parameters $q_n = q_n(s_n = 1)$ [50%]
- (b) Using the answer to part (a), compute the M-step update equation for \mathbf{T} in terms of q_n . [50%]

$$\begin{aligned} \text{a)} \quad \mathcal{F}(\mathbf{T}, \{q_n\}_{n=1}^N) &= \sum_n \sum_{s_n} q_n(s_n) \log \frac{p(x_n, s_n)}{q_n(s_n)} \\ &= \sum_{n=1}^N \sum_{s_n=0}^1 \left[q_n(s_n) \left(x_n \log p(x_n=1|s_n) + (1-x_n) \log p(x_n=0|s_n) \right) \right. \\ &\quad \left. - \sum_n \sum_{s_n} q_n(s_n) \log q_n(s_n) \right] + N \log 1/2 \\ &= \sum_n q_n(s_n) \left[s_n x_n \log T_{11} + (1-s_n) x_n \log T_{10} + (1-s_n) s_n \log T_{01} + (1-s_n)(1-s_n) \log T_{00} \right] \\ &\quad - \sum_n \left[q_n \log q_n + (1-q_n) \log (1-q_n) \right] + N \log 1/2 \\ &= \sum_n \left[q_n x_n \log T_{11} + (1-q_n) x_n \log T_{10} + q_n (1-x_n) \log T_{01} + (1-q_n)(1-x_n) \log T_{00} \right] \\ &\quad - \sum_n \left[q_n \log q_n + (1-q_n) \log (1-q_n) \right] + N \log 1/2 \\ &= \log T_{11} \left(\sum_n q_n x_n \right) + \log T_{10} \sum_n (1-q_n) x_n + \log T_{01} \sum_n q_n (1-x_n) + \log T_{00} \sum_n (1-q_n)(1-x_n) \\ &\quad - \sum_n \left[q_n \log q_n + (1-q_n) \log (1-q_n) \right] - N \log 2 \end{aligned}$$

$$b) \frac{\partial}{\partial T_{11}} \left(\frac{1}{T_{11}} - \lambda [T_{11} + T_{01}] \right) = 0 = \frac{1}{T_{11}^2} \sum_n g_n x_n - \lambda$$

by analogous method:

$$\Rightarrow T_{11} = \frac{\sum_n g_n x_n}{\sum_n g_n} \quad T_{01} = \frac{\sum_n g_n (1-x_n)}{\sum_n g_n}$$

$$T_{10} = \frac{\sum_n (1-g_n) x_n}{\sum_n (1-g_n)} \quad T_{00} = \frac{\sum_n (1-g_n) (1-x_n)}{\sum_n (1-g_n)}$$

- 7 A biological time series dataset $y_{1:T}$ is shown in Fig. 2.

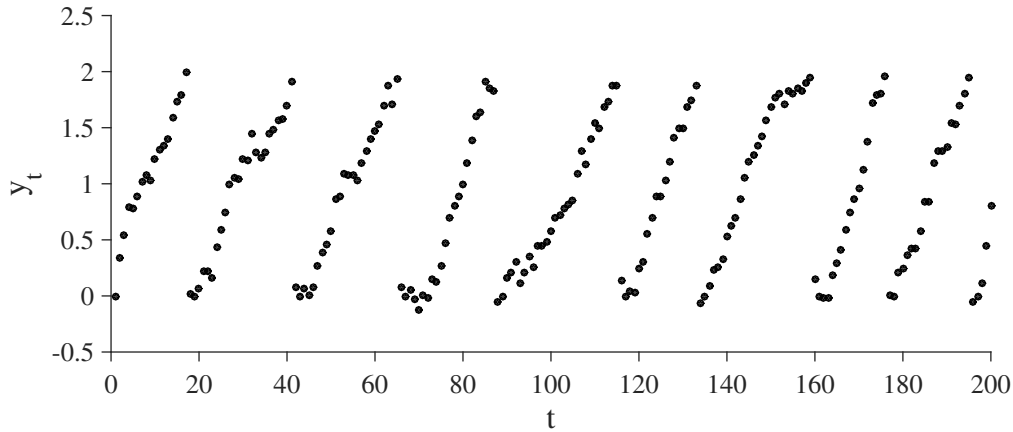


Fig. 2

- (a) Suggest a first order Markov Model for these data. Explain your reasoning. [60%]
- (b) A friend has code that fits flexible non-linear functions to regression datasets comprising input-output pairs using maximum likelihood estimation. 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%]

$$a) \quad p(y_{1:T}) = p(y_1) \prod_{t=2}^T p(y_t | y_{t-1}) \quad [\text{definition of a Markov Model}]$$

$$p(y_t | y_{t-1}) = N(y_t; f(y_{t-1}), \sigma^2)$$

$$\sigma^2 = \frac{1}{100}$$

$$f(y) = \begin{cases} y + 1/10 & \text{if } y \leq 2 \\ 0 & \text{if } y > 2 \end{cases}$$

b) She is correct

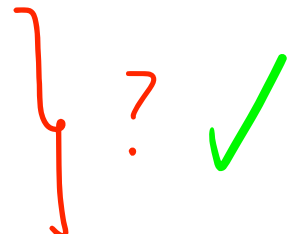
Her code: $\arg \max_{\theta} \sum_t \log p(y_t | x_t, \theta)$

We need: $\arg \max_{\theta} \sum_t \log p(y_t | y_{t-1}, \theta)$

• use her code on regression data set made from $\{y_t, y_{t-1}\}$ input output pairs

$$\{y_t, y_{t-1}\}_{t=2}^T$$

↑
new regression dataset we can use to fit the model.



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 mean vector μ and 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%]

a) AR(1) process

$$x_t = \lambda x_{t-1} + \sigma \varepsilon_t \quad \varepsilon_t \sim \mathcal{N}(0, 1)$$

$$x_1 \sim \mathcal{N}(0, \sigma^2)$$

everything is linear + gaussian

first variable 0 mean \Rightarrow all others will be

$$\begin{aligned} b) \quad p(x_{1:T}) &= p(x_1) \prod_{t=2}^T p(x_t | x_{t-1}) = \mathcal{N}(x_{1:T} | \underline{0}, \underline{P}^{-1}) \\ &= \exp \left[-\frac{1}{2} x_1^2 / \sigma^2 - \frac{1}{2\sigma^2} \sum_{t=2}^T (x_t - \lambda x_{t-1})^2 \right] \\ &= \exp \left(-\frac{1}{2} x_1^2 / \sigma^2 - \frac{1}{2\sigma^2} \sum_{t=2}^T (x_t^2 + \lambda^2 x_{t-1}^2 - 2\lambda x_t x_{t-1}) \right) \\ &= \exp \left(-\frac{1}{2} x_1^2 / \sigma^2 - \frac{\lambda^2}{\sigma^2} x_1^2 - \frac{1}{2\sigma^2} x_2^2 (1 + \lambda^2) + \frac{\lambda}{\sigma^2} x_2 x_1 \right. \\ &\quad \left. - \frac{1}{2\sigma^2} (1 + \lambda^2) x_3^2 + \frac{\lambda}{\sigma^2} x_2 x_3 + \dots \right) \end{aligned}$$

$$\therefore P = \Sigma^{-1} = \frac{1}{\sigma^2} \begin{bmatrix} \frac{\sigma^2}{\sigma^2} + \lambda^2 & -\lambda & 0 & 0 & \dots \\ -\lambda & 1 + \lambda^2 & -\lambda & 0 & \dots \\ 0 & -\lambda & 1 + \lambda^2 & -\lambda & \dots \\ 0 & 0 & -\lambda & 1 + \lambda^2 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

zeros emerge due to the Markov property

$$c) \quad P_{i,j} = 0 \Rightarrow x_i \perp x_j \mid x_{\neq i,j}$$

i.e. a zero in a precision matrix is a signature of conditional independence

END OF PAPER