Version RET: final

MACHINE LEARNING, SPEECH & LANGUAGE TECHNOLOGY MPHIL

Wednesday 2nd November 2016 10.30 to 12.15

MLSALT1

INTRODUCTION TO MACHINE LEARNING, SPEECH AND LANGUAGE TECHNOLOGY

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 noisy depth sensor measures the distance to an object an unknown distance d metres away. The depth can be assumed, *a priori*, to be distributed according to a standard Gaussian distribution $p(d) = \mathcal{N}(d; 0, 1)$. The depth sensor returns y a noisy measurement of the depth, that is also assumed to be Gaussian $p(y|d, \sigma_y^2) = \mathcal{N}(y; d, \sigma_y^2)$.

(a) Compute the posterior distribution over the depth given the observation, $p(d|y, \sigma_y^2)$. [80%]

(b) What happens to the posterior distribution as the measurement noise becomes very large $\sigma_y^2 \to \infty$? Comment on this result. [20%]

The formula for the probability density of a Gaussian distribution of mean μ and variance σ^2 is given by

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

2 A sequence of coin tosses are observed from a biased coin $x_{1:N} = \{0, 1, 1, 0, 1, 1, 1, 1, 0\}$ where $x_n = 1$ indicates flip n was a head and $x_n = 0$ indicates that it was tails. An experimenter would like to estimate the coin's probability of landing heads, ρ, from these data.

The experimenter assumes that the coin flips are drawn independently from a Bernoulli distribution $p(x_n|\rho) = \rho^{x_n}(1-\rho)^{1-x_n}$ and uses a prior distribution of the form

$$p(\rho|n_0, N_0) = \frac{1}{Z(n_0, N_0)} \rho^{n_0}(1-\rho)^{N_0-n_0}.$$

Here n_0 and N_0 are parameters set by the experimenter to encapsulate their prior beliefs. $Z(n_0, N_0)$ returns the normalising constant of the distribution as a function of the parameters, n_0 and N_0.

(a) Compute the posterior distribution over the bias $p(\rho|x_{1:N}, n_0, N_0)$. [40%]

(b) Compute the *maximum a posteriori* (MAP) estimate for the bias. [40%]

(c) Provide an intuitive interpretation for the parameters of the prior distribution, n_0 and N_0. [20%]

3 A data-scientist has computed a complex posterior distribution over a variable of interest, x, given observed data y, that is $p(x|y)$. They would like to return a point estimate of x to their client. The client provides the data-scientist with a reward function $R(\hat{x},x)$ that indicates their satisfaction with a point estimate $\hat{x}$ when the true state of the variable is x.

(a) Explain how to use *Bayesian Decision Theory* to determine the optimal point estimate, $\hat{x}$. [40%]

(b) Compute the optimal point estimate $\hat{x}$ in the case when the reward function is the negative square error between the point estimate and the true value, $R(\hat{x},x) = -(\hat{x}-x)^2$. Comment on your result. [60%]

4 A data-scientist has collected a regression dataset comprising N scalar inputs $(\{x_n\}_{n=1}^N)$ and N scalar outputs $(\{y_n\}_{n=1}^N)$. Their goal is to predict y from x and they have assumed a very simple linear model, $y_n = ax_n + \varepsilon_n$.

The data-scientist also has access to a second set of outputs $(\{z_n\}_{n=1}^N)$ that are well described by the model $z_n = x_n + \varepsilon'_n$.

The noise variables ε_n and ε'_n are known to be zero mean correlated Gaussian variables

$$p\left(\begin{bmatrix}\varepsilon_n\\ \varepsilon'_n\end{bmatrix}\right) = \mathcal{N}\left(\begin{bmatrix}\varepsilon_n\\ \varepsilon'_n\end{bmatrix};\mathbf{0},\Sigma\right) \quad \text{where} \quad \Sigma^{-1} = \begin{bmatrix}1 & 0.5\\ 0.5 & 1\end{bmatrix}.$$

(a) Provide an expression for the log-likelihood of the parameter a. [20%]

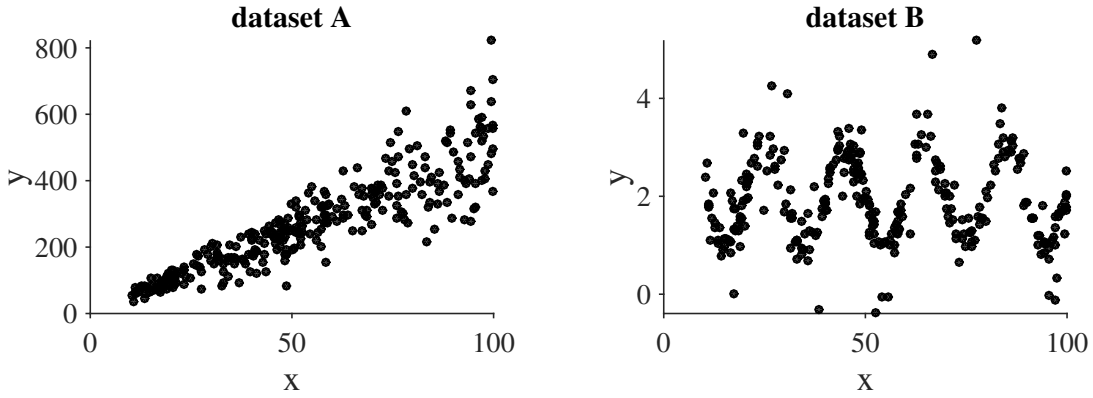
(b) Compute the maximum likelihood estimate for a. [60%]

(c) Do the additional outputs $\{z_n\}_{n=1}^N$ provide useful additional information for estimating a? Explain your reasoning. [20%]

The formula for the probability density of a multivariate Gaussian distribution of mean μ and covariance Σ is given by

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

5 A machine learner observes two separate regression datasets comprising scalar inputs and outputs $\{x_n, y_n\}_{n=1}^N$ shown below.



(a) Suggest a suitable regression model, $p(y_n|x_n)$ for the dataset A. Indicate sensible settings for the parameters in your proposed model where possible. Explain your modelling choices. [50%]

(b) Suggest a suitable regression model, $p(y_n|x_n)$ for the dataset B. Indicate sensible settings for the parameters in your proposed model where possible. Explain your modelling choices. [50%]

6 A data-scientist would like to summarise high dimensional data points $\mathbf{y}_n$ in terms of a single scalar variable x_n. They use an encoding weight $\mathbf{w}$ to produce the summary, $x_n = \mathbf{w}^\top \mathbf{y}_n$, and a decoding weight $\mathbf{r}$ to reconstruct the data point from the summary, $\hat{\mathbf{y}}_n = \mathbf{r}x_n$. The data-scientist would like to learn the encoding and decoding weights by optimising the squared error of the reconstruction,

$$\mathscr{C} = \sum_n ||\mathbf{y}_n - \hat{\mathbf{y}}_n||^2.$$

(a) Minimise the cost $\mathscr{C}$ with respect to the decoding weights $\mathbf{r}$, returning an expression for them in terms of x_n and $\mathbf{y}_n$. [50%]

(b) Substitute your expression for the optimised decoding weights $\mathbf{r}$ into $\mathscr{C}$ to obtain the cost purely in terms of the encoding weights $\mathbf{w}$. [20%]

(c) Now consider minimising the cost derived in part (b) with respect to the encoding weights. What is the solution? Is it unique? [30%]

It may be useful to know that the solution to the optimisation problem $\mathbf{z}^* = \arg\max_{\mathbf{z}} \frac{\mathbf{z}^\top H \mathbf{z}}{\mathbf{z}^\top \mathbf{z}}$ is the largest eigenvector of matrix H (arbitrarily scaled), $\mathbf{z}^* \propto \mathbf{e}_1$.

7 A set of N scalar data points $\{y_n\}_{n=1}^N$ are modelled using a mixture of Gaussians containing two equiprobable components with unknown means (μ_0 and μ_1) and unit variances,

$$p(s_n = 1) = \frac{1}{2}, \quad p(y_n|s_n = 0) = \mathcal{N}(y_n; \mu_0, 1), \quad p(y_n|s_n = 1) = \mathcal{N}(y_n; \mu_1, 1). \qquad (1)$$

(a) Compute the posterior distribution over the components, $p(s_n = 1|y_n)$ and sketch how this varies as a function of the observed data y_n. [40%]

(b) Explain how your solution to (a) can be used in the EM algorithm to estimate the component means. Your answer should include a expression for the M-step update. [40%]

(c) Do you expect the EM algorithm to overfit when used to train this model? [20%]

8 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. [40%]

(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. [50%]

(c) What happens to $p(y_{T+\tau}|y_{1:T})$ as $\tau\to\infty$? [10%]

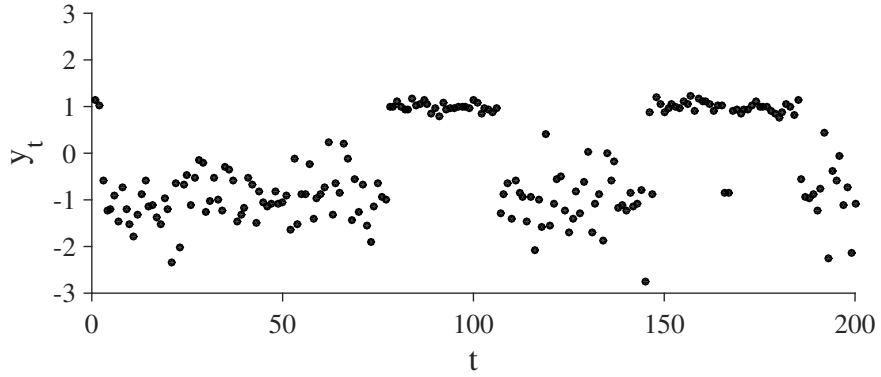
9 (a) Provide the probabilistic equations that define a Hidden Markov Model (HMM) for observed data that takes discrete values. Indicate what aspects of the model the following terms refer to: *initial state probabilities*, *transition matrix* and *emission matrix*. [20%]

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

AABBBACABBBACAAAAAAAAAABBBACAAAAABACAAAAAABBBBACAAAAAAAA
AAAABACABACAABBACAAABBBBACAAABACAAAABACAABACAAABBACAAAA
BBBBACABBACAAAAAABACABACAAABACAABBBACAAAABACABBACA

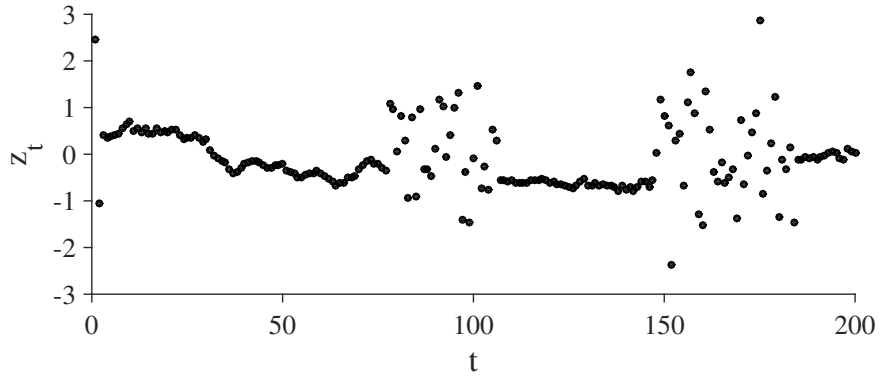
Carefully analyse the string paying close attention to repeated patterns. Describe an HMM model for the string. Your description should include the number of states in the HMM, the transition matrix including the values of the elements of the matrix, the emission matrix including the values of its elements, and the initial state probabilities. Explain your reasoning. [80%]

10 (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. [50%]

(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. [50%]

END OF PAPER

THIS PAGE IS BLANK