UNIVERSITY OF BRISTOL

August 2018

FACULTY OF ENGINEERING

Third year Examination for the Degrees of BSc, BEng and MEng

COMS30007J

Machine Learning

TIME ALLOWED:

2 Hours

This paper contains *14* questions.
The maximum for this paper is *14 marks*.
All answers will be used for assessment.
Each question has one correct answer.

1. Please make sure you read the instructions on the answer sheet.
2. Only the answer sheet will be marked, the empty pages at the back of the exam is only used for your calculations.
3. When selecting answers, make clear, horizontal marks within the two sets of brackets, making sure that the contained letter is struck through.
4. Avoid marking the answer sheet outside specified areas.
5. Do not crease, dog-ear or otherwise damage the answer sheet.

Other Instructions

Calculators must have the Engineering Faculty seal of approval.

TURN OVER ONLY WHEN TOLD TO START WRITING

1. (1 point) We have seen some data $\mathcal{D}$ which we try to represent using parameter θ. Which combination of "semantic" names for the distribution below is correct?

$$\underbrace{p(\theta|\mathcal{D})}_{A} = \frac{\overbrace{p(\mathcal{D}|\theta)}^{B}\overbrace{p(\theta)}^{C}}{\underbrace{p(\mathcal{D})}_{D}}$$

   (a) {A,B,C,D} = {Prior, Evidence, Posterior, Likelihood}

   (b) {A,B,C,D} = {Likelihood, Prior, Evidence, Posterior}

   (c) {A,B,C,D} = {Posterior, Likelihood, Prior, Evidence}

   (d) {A,B,C,D} = {Posterior, Joint, Posterior, Likelihood}

   (e) {A,B,C,D} = {Evidence, Posterior, Likelihood, Prior}

2. (1 point) Given the joint probability $p(x, y)$ which of the statements below is false,

   (a) The conditional probability of y given x: $p(y|x) = \frac{p(x,y)}{p(x)}$

   (b) The joint probability $p(x, y) = p(x|y)p(y)$

   (c) The joint probability $p(x, y) = p(y|x)p(x)$

   (d) The marginal probability over y: $p(y) = \int p(y|x)\mathrm{d}x$

   (e) The marginal probability over x: $p(x) = \int p(x, y)\mathrm{d}x$

3. (1 point) You have built a model that uses a Poisson likelihood and you are trying to infer the *rate* parameter of the distribution from data. You pick the conjugate prior which is a Gamma distribution. When you derive the posterior over the rate, what form will it have?

   (a) Gaussian

   (b) Gamma

   (c) Inverse-Wishard

   (d) Categorical

   (e) Bernoulli

4. (1 point) Which characteristics below are **false** for a non-parametric model

   (a) a non-parametric model cannot have any parameters

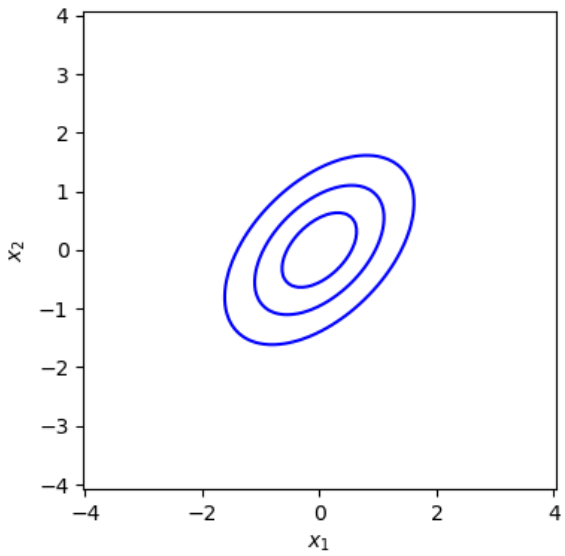
   (b) the model describes data by relating it to the training data

   (c) nearest neighbour classification is an example of a non-parametric method

   (d) a non-parametric method is always specified by a stochastic process

(e) Gaussian and Dirichlet processes are examples of stochastic process that can specify non-parametric models

5. (1 point) Match the Gaussian in the image below to the correct co-variance matrix



(a)
$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

(b)
$$\begin{bmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{bmatrix}$$

(c)
$$\begin{bmatrix} 1 & -\frac{1}{2} \\ -\frac{1}{2} & 1 \end{bmatrix}$$

(d)
$$\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$$

(e)
$$\begin{bmatrix} 0 & 2 \\ 2 & 0 \end{bmatrix}$$

6. (1 point) Given a set of associated input values $\mathbf{X}$ and target values $\mathbf{t}$ you derived the posterior distribution over regression weights $\mathbf{w}$ for a linear model,

$$p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \alpha, \beta),$$

where α and β are parameters of the likelihood and the prior respectively. In order to reach the *predictive* distribution of the model which random variable should you marginalise over

(a) t_* the output values at location $\mathbf{x}_*$

(b) $\mathbf{X}$ the input values of the training data

(c) α and β the parameters of the likelihood and the prior

(d) $\mathbf{w}$ the regression weights

(e) $\mathbf{t}$ the target values of the training data

7. (1 point) Which of the following statements are false for Gaussian processes,

   (a) It is a general function approximate, i.e. it places non-zero probability mass over each function

   (b) It is specified by a mean and a co-variance function specified over an infinite index set

   (c) The class of valid co-variance functions is the same as the class of valid kernels

   (d) It is an example of a stochastic process

   (e) Deriving the predictive posterior distribution of a Gaussian process prior with a Gaussian likelihood is analytically intractable

8. (1 point) Bayesian Optimisation is a way of finding the extreme point of a non-explicit function which of the following statements is **not** true for Bayesian Optimisation.

   (a) we require a prior assumption of the function that we are optimising

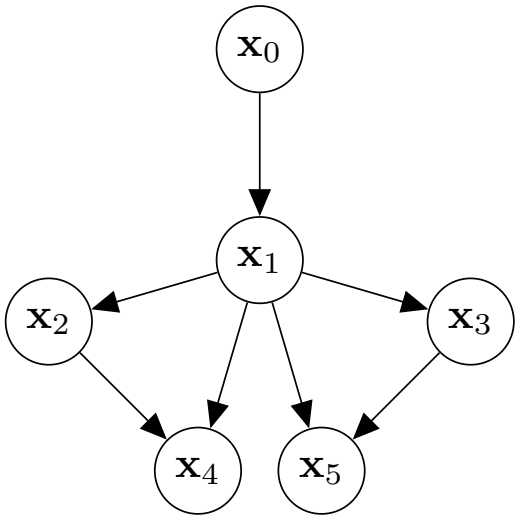
   (b) for efficiency it is important that the acquisition function is computationally cheap to evaluate

   (c) we are guaranteed to find the global optima

   (d) we need to choose an acquisition function that we can find the extreme points (maxima or minima) of

   (e) the uncertainty in our estimate of the function is essential as input to the acquisition function

9. (1 point) A Graphical model is a visual description of the joint distribution factorised into its components. Which factorisation does the following model encode?



(a) $p(x_4, x_5|x_1, x_2, x_3)p(x_2)p(x_3)p(x_1|x_0)$

(b) $p(x_0)p(x_1)p(x_2)p(x_3)p(x_4)p(x_5)$

(c) $p(x_4, x_5|x_1, x_2, x_3)p(x_1|x_0)$

(d) $p(x_5|x_1, x_3)p(x_4|x_1, x_2)p(x_2)p(x_3)p(x_1|x_0)p(x_0)$

(e) $p(x_5|x_1, x_2, x_3)p(x_4|x_1, x_2, x_3)p(x_2)p(x_3)p(x_1|x_0)p(x_0)$

10. (1 point) The Laplace approximation is a method to approximate an intractable posterior distribution. Which of the following statements is **true** for the Laplace approximation,

    (a) the Laplace approximation is exact

    (b) we need to be able to find the maximum of the posterior to apply the approximation

    (c) the optimisation problem the approximation leads to is non-convex

    (d) the more terms we keep in the Taylor expansion around the mode of the posterior the better the Gaussian approximation will fit the true posterior

    (e) The Laplace approximation can only be used for classification

11. (1 point) In variational inference we aim to try and find an explicit approximation $q(\theta)$ to an intractable posteriors $p(\theta|\mathcal{D}$ such that,

$$q(\theta) \approx p(\theta|\mathcal{D}),$$

    where θ is some parameter we want to infer from the data $\mathcal{D}$. Which of the following statements is **false** for variational inference,

    (a) by using Jensen's inequality we can derive a lower bound on $p(\mathcal{D})$

(b) Jensen's inequality induces a measure of similarity between the approximation and the true posterior referred to as the *Kullback-Leibler divergence*

(c) if the posterior is intractable the approximation can never be exact

(d) maximising the lower bound on the marginal likelihood is a convex function

(e) the *mean-field* approximation assumes that all latent variables factorises

12. (1 point) We have a model that creates an output f and a set of observed data y, we define a likelihood function $p(y|f)$. Which of the following statements is **false** for this function.

(a) for a specific value of f the function integrates to one.

(b) it describes how likely the data y is to have come from the output of the model f.

(c) if the likelihood can be factorised across each data point such as,

$$p(\mathbf{y}|\mathbf{f}) = \prod_{i=1}^{N} p(y_i|f_i),$$

this implies that the instances i of the observed data is conditionally independent given f.

(d) The likelihood function encodes our assumption of the distribution of the noise in the data.

(e) The likelihood function is a continuous function specified over the whole real line.

13. (1 point) Which of the following statements is **false** for sampling,

(a) Using sampling we try to approximate an intractable integral with a sum

(b) A sampling method will **always** recover the exact integral in the limit independent of the proposal distribution.

(c) A sampling method is guaranteed to recover the exact integral in the limit if the proposal distribution has infinite support

(d) If $f(y)$ is the cumulative distribution function of $p(y)$ we can use this function to transform samples from a uniform distribution to $p(y)$

(e) A Markov Chain sampler introduces a state such that samples are drawn in a chain.

14. (1 point) Which of the following statements is **true**

(a) The maximum likelihood estimate is always the same as the maximum-a-posteriori estimate

(b) When the number of data-points approaches infinity the maximum-likelihood estimate and the maximum-a-posteriori estimate will always be the same

(c) In a Type-II Maximum Likelihood estimate we have marginalised out some of the variables

(d) The maximum-a-posteriori estimate is not guaranteed to be the maxima of the posterior

(e) all statements above are **false**

15. (1 point) Which of the following statements is **false**

(c) a generative model parameterise the observed data using a set of latent variables that describes the process of how to create the data

(c) a good generative model can reconstruct the observed data

(c) a generative model always describes the exact casual structure of how the data came into being

(c) Latent Dirichlet Allocation is an example of a generative model for text

(c) all of the statements above are **false**

End of Paper