UNIVERSITY OF BRISTOL

JANUARY 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 *15* questions.
The maximum for this paper is *15 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) You have a gaussian likelihood and you are trying to infer the value of the covariance from data. You pick the conjugate prior to the covariance of a Gaussian which is an Inverse-Whishard distribution. When you derive the posterior over the covariance, what form will it have?

   (a) Gaussian

   (b) Beta

   (c) Normal Inverse-Wishard

   (d) Inverse-Whishard

   (e) Dirichlet

3. (1 point) A Kernel function is a function that computes an inner-product and induces a feature space as defined below,

$$k(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^{\mathrm{T}}\phi(\mathbf{x}_j) = k_{ij}$$

   Which of the following statements is true for kernel functions?

   (a) any function can be a kernel function

   (b) any function which is strictly positive can be a kernel function

   (c) a kernel function needs to be symmetric such that $k(\mathbf{x}_i, \mathbf{x}_j) = k(\mathbf{x}_j, \mathbf{x}_i)$

   (d) if d is the dimensionality of the output of the kernel functions, $k_{ij} \in \mathbb{R}^d$, and D is the dimensionality of the input data $\mathbf{x}_i \in \mathbb{R}^D$. The following relationship is always true: $d > D + 1$

   (e) it is not possible for $k(\mathbf{x}_i, \mathbf{x}_j) = 0$

4. (1 point) As the number of data points grows approaching infinity will a Maximum-a-posterior estimate always be the same as the Maximum Likelihood estimate independent of choice of Prior?

   a Yes

   b No

5. (1 point) To find the minima of a continous non-explicit function, we can use Bayesian optimisation. Which of the following statements is true for Bayesian optimisation?

   (a) We are guaranteed to find the global minima of the function in finite time

   (b) After performing N tests of the function we know how far away from the global minima our current estimate is

   (c) The number of tests it takes for us to reach a solution is independent of our prior assumption of the function that we minimise

   (d) It is impossible to know how close to the true minima our current estimate is

   (e) None of the above

6. (1 point) Given the conditional probability $p(y|\mu) = \mathcal{N}(\mu, 1.0)$ where $\mu \in \{0, 1, 2, 3\}$ and a probability mass function as follows:

μ	$p(\mu)$
0	$\frac{1}{6}$
1	0
2	$\frac{1}{3}$
3	$\frac{1}{2}$

   compute the marginal distribution $p(y)$ by taking the expectation $\mathbb{E}_{p(\mu)}[p(y|\mu)]$. Which of the following is the correct marginal?

   (a) $p(y) = \mathcal{N}(0, 1)$

   (b) $p(y) = \mathcal{N}(0 + 2 + 3, 1)$

   (c) $p(y) = \frac{1}{6}\mathcal{N}(0, 1) + \frac{1}{3}\mathcal{N}(2, 1) + \frac{1}{2}\mathcal{N}(3, 1)$

   (d) $p(y) = \frac{1}{6}\mathcal{N}(0, 1) + \frac{1}{3}\mathcal{N}(2, 1) + \frac{1}{2}\mathcal{N}(3, 1) = \left(\frac{1}{6} + \frac{1}{3} + \frac{1}{2}\right)\mathcal{N}(0 + 1 + 3, 1 + 1 + 1)$

   (e) $p(y) = \frac{1}{6}\mathcal{N}(\frac{1}{6}0 + \frac{1}{3}2 + \frac{1}{2}3, \frac{1}{6}1 + \frac{1}{3}1 + \frac{1}{2}1)$

7. (1 point) A neural network is composed of hierarchies of functions, we often refer to the number of the hierarchies as the "depth" of the model. If we increase the depth of the hierarchy by adding a new function what will happen to the dimensionality of Image and the Kernel of the function that the whole network represents? Remember that this refers to the Kernel of the function, not a Kernel as in a inner-product function.

   (a) The Kernel and the Image will remain the same or increase

   (b) The Kernel and the Image will remain the same or decrease

   (c) The Kernel will increase or stay the same while the Image will decrease or stay the same

   (d) The Kernel will decrease or stay the same while the Image will increase or stay the same

   (e) It is not possible to say this without knowing the specific function added to the hierarchy

8. (1 point) We have a model of data $p(\mathcal{D}|\theta)$ and a prior over the parameters $p(\theta)$. To learn we wish to formulate the marginal distribution of the data $p(\mathcal{D})$. However this computation is intractable which is why we choose to approach the problem using a variational approximation. Which distribution is the variational distribution $q(\theta)$ trying to approximate?

   (a) $p(\mathcal{D})$

   (b) $p(\mathcal{D}|\theta)$

   (c) $p(\theta)$

   (d) $p(\theta|\mathcal{D})$

   (e) $p(\mathcal{D},\theta)$

9. (1 point) Which of the following statements is **not** true for a non-parametric model?

   (a) the process describing the model is parametrised using a set of hyper-parameters

   (b) a non-parametric model is often viewed as a parametric model with infinite number of parameters

   (c) the representative power of a non-parametric model adapts to the data

   (d) a Bayesian non-parametric model is defined by a stochastic process

   (e) a non-parametric model is always described by distribution

10. (1 point) Which of the following statements is **not** true for a Gaussian process prior?

    (a) it places non-zero probability mass over the whole hyperplane it is defined over

    (b) it is completely defined by a mean and co-variance function

    (c) the class of valid co-variance functions are the same as the class of kernel functions

    (d) it can only represent stationary functions, i.e. functions that have the same characteristics across the whole input domain

    (e) it is a general function approximator

11. (1 point) Which of the following statments is **not** true for a stochastic process?

    (a) the instantiation of a stochastic process is a distribution

    (b) the instantiation of an instantiation of a stochastic process is a value

    (c) a Gaussian process is an example of a stochastic process over an uncountable infinite index set while a Dirichlet process is an example of a stochastic process over a countably infinite index set.

    (d) a stochastic process is a collection of random variables

    (e) none of the above

12. (1 point) Which of the following is **not** true for a Type-II Maximum Likelihood estimate?

    (a) it is a point estimate

    (b) it will always be the same as the ML solution when the data grows towards infinite

    (c) it means that some parameters have been marginalised out while we optimise for the remaining ones

    (d) it maximises a marginal likelihood

    (e) None of the above

13. (1 point) The denominator in Baye's rule is often intractable to compute. If we have a set of data $\mathcal{D}$ parameterised by θ we wish to compute?

$$p(\mathcal{D}) = \int p(\mathcal{D}|\theta)p(\theta)\mathrm{d}\theta$$

If the above integral is intractable we use approximate computations. Which of the following statments is **not** true for approximate integration,

(a) When we use a sampling method the approximation will Generally be correct given infinite number of samples.

(b) Using Variational methods we approximate the integral with a simpler integral that we can compute

(c) Using a variational approach we will be able to recover the true distribution if we can solve an optimisation problem

(d) The Laplace approximation approximates an intractable posterior directly by mode matching a surrogate distribution with the true posterior

(e) none of the above

14. (1 point) Which of the following statements is **not** correct with respect to unsupervised learning?

(a) in unsupervised learning we are trying to infer a latent representation of some observed data

(b) unsupervised learning requires additional prior assumptions as the task is very ill-constrained compared to supervised learning

(c) in continous unsupervised learning we can find a lower dimensional representation of the data if it lies on, or close to a manifold in the observed space

(d) principal component analysis assumes the latent representation to be Gaussian distributed

(e) none of the above

15. (1 point) Which of the following statements are true for sampling?

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

(b) A sampling method can **never** recover the exact solution

(c) The more dependent the samples we use in the approximation the less samples we are likely to need

(d) a sampling method is an example of a deterministic approximation and will recover the same solution every time it is applied

(e) none of the above

End of Paper

this page is left blank for your rough working

this page is left blank for your rough working

this page is left blank for your rough working