UNIVERSITY OF EDINBURGH

COLLEGE OF SCIENCE AND ENGINEERING

SCHOOL OF INFORMATICS

INFR11073 MACHINE LEARNING AND PATTERN RECOGNITION (LEVEL 11)

Monday 25th April 2016

09:30 to 11:30

INSTRUCTIONS TO CANDIDATES

Answer QUESTION 1 and ONE other question.

Question 1 is COMPULSORY.

All questions carry equal weight.

CALCULATORS MAY NOT BE USED IN THIS EXAMINATION

Year 4 Courses

Convener: I. Stark
External Examiners: A. Burns, A. Cohn, P. Healey, T. Field, T. Norman

THIS EXAMINATION WILL BE MARKED ANONYMOUSLY

The following information may be of use in the questions on this paper.

A multivariate Gaussian random variable $\mathbf{y} \sim \mathcal{N}(\mu, \Sigma)$ in D dimensions has a probability density function

$$\mathcal{N}(\mathbf{y};\, \mu, \Sigma) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{y}-\mu)^\top \Sigma^{-1}(\mathbf{y}-\mu)\right)$$

1. **You MUST answer this question.**

(a) The Beta distribution

$$p(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

for $\theta \in [0, 1]$ is a conjugate prior distribution to the Bernoulli likelihood, where $\Gamma(\cdot)$ denotes the Gamma function.

i. Suppose you have counts of 10 and 5 for the number of 1s and the number of 0s respectively in a dataset. Assume a Beta prior for θ with $\alpha = \beta = 2$. Write down an expression for the posterior distribution for θ after observing the dataset. Sketch the posterior distribution. [*3 marks*]

ii. Explain what *conjugacy* means. [*1 mark*]

(b) Describe a simple gradient descent procedure for optimising the weights $\mathbf{w}$ (including the biases) of a feedforward neural network, given the network's negative log likelihood, denoted $E(\mathcal{D}|\mathbf{w})$, for data $\mathcal{D}$. Describe the problems associated with setting the learning rate. You do not need to say how to compute any derivatives you might need. [*4 marks*]

(c) An *ordinal regression* problem is one where we wish to predict an ordered, but non-numerical variable. For example we may wish to predict whether a student will obtain the grade distinction, merit, pass or fail on an exam.

Suppose that the ordinal variable y is mapped to K different values on a scale $1, \ldots, K$. We formulate a model

$$p(y \le k|\mathbf{x}, \mathbf{w}, \theta) = \sigma(\theta_k - \mathbf{w}^T\mathbf{x}) \qquad \text{for } k = 1, \ldots, K-1,$$

for input vector $\mathbf{x}$, parameters $\mathbf{w}$ and $\theta_1 < \theta_2 < \ldots < \theta_{K-1}$, and $\sigma(z) = (1 + e^{-z})^{-1}$.

i. Suppose that $\mathbf{x}$ is two-dimensional. Sketch the situation and show the decision boundaries defined by $p(y \le k|\mathbf{x}, \mathbf{w}, \theta) = 0.5$ for $k = 1, \ldots, K-1$. [*2 marks*]

ii. Write down the likelihood $p(y = k|\mathbf{x}, \mathbf{w}, \theta)$ and suggest how this model could be trained given a dataset of $(\mathbf{x}^n, y^n)$ pairs. [*2 marks*]

QUESTION CONTINUES ON NEXT PAGE

iii. A friend suggests that this model could be trained by carrying out logistic regression to separate examples with $y \leq k$ from those with $y > k$ for $k = 1, \ldots, K-1$. What is the problem with this suggestion? [*1 mark*]

(d) Briefly explain how the full Bayesian predictive distribution for a logistic regression classifier differs from predictions using a maximum a posteriori (MAP) estimate. Include an example of how the Bayesian predictions can be better. What is a reason to prefer MAP estimation? [*4 marks*]

(e) A regression task has N training datapoints $\mathcal{D} = \{\mathbf{x}^{(n)}, y^{(n)}\}_{n=1}^{N}$, where each $\mathbf{x}^{(n)}$ is a D-dimensional feature vector and $y^{(n)}$ is the corresponding real-valued label.

These data are assumed to be noisy observations of a function,

$$y^{(n)} \sim \mathcal{N}\big(f(\mathbf{x}^{(n)}),\, \sigma_o^2\big),$$

where the function is modelled as a zero mean Gaussian Process (GP) with the kernel or covariance function

$$k(\mathbf{x}^{(n)}, \mathbf{x}^{(m)}) \;=\; \sigma_f^2 \exp\left(-\frac{1}{2}\sum_{d=1}^{D} \frac{\big(x_d^{(n)} - x_d^{(m)}\big)^2}{\ell^2}\right).$$

The posterior distribution over two function values $f_*^{(1)} = f(\mathbf{x}_*^{(1)})$ and $f_*^{(2)} = f(\mathbf{x}_*^{(2)})$ at test locations $\mathbf{x}_*^{(1)}$ and $\mathbf{x}_*^{(2)}$ is:

$$p(f_*^{(1)}, f_*^{(2)} \,|\, \mathbf{x}_*^{(1)}, \mathbf{x}_*^{(2)}, \mathcal{D}) \;=\; \mathcal{N}\left(\begin{bmatrix} f_*^{(1)} \\ f_*^{(2)} \end{bmatrix};\, \begin{bmatrix} m_1 \\ m_2 \end{bmatrix},\, \begin{bmatrix} s_{1,1} & s_{1,2} \\ s_{1,2} & s_{2,2} \end{bmatrix}\right),$$

where the parameters of the posterior m_1, m_2, $s_{1,1}$, $s_{1,2}$, and $s_{2,2}$ are easily computed.

i. Write down an expression for the predictive distribution $p(y_*^{(1)} \,|\, \mathbf{x}_*^{(1)}, \mathcal{D})$, the beliefs about a new unknown label $y_*^{(1)}$ sampled at test location $\mathbf{x}_*^{(1)}$, given the training data. Your answer should be in terms of the parameters of the model and posterior as given in the question. [*2 marks*]

ii. Briefly state what effect varying σ_f^2 has on the model, and what the specific assumption $\sigma_f^2 = 100$ would say about our prior on functions. [*2 marks*]

iii. Briefly explain what effect varying ℓ^2 has on the Gaussian process model, and what the model means in the limit $\ell^2 \rightarrow 0$. What are the values of m_1, m_2, $s_{1,1}$, $s_{1,2}$, and $s_{2,2}$ in the limit $\ell^2 \rightarrow 0$? You can assume that the test input locations are different from each other and from all of the training input locations. [*4 marks*]

2. (a) Assume that $\mathbf{x}$ consists of D 0/1 binary features, so that $\mathbf{x} \in \{0,1\}^D$, and that there are two classes $y \in \{0,1\}$. A naive Bayes model for this situation has the form

$$p(\mathbf{x}|y=i,\boldsymbol{\theta}) = \prod_{d=1}^{D} \theta_{id}^{x_d}(1-\theta_{id})^{(1-x_d)} \qquad \text{for } i=0,1.$$

Show that for this model the log odds ratio is linear in $\mathbf{x}$, i.e.

$$\log\left\{\frac{p(y=1|\mathbf{x},\boldsymbol{\theta})}{p(y=0|\mathbf{x},\boldsymbol{\theta})}\right\} = \mathbf{w}^T\mathbf{x} + w_0,$$

and relate the parameters $\mathbf{w}$ and w_0 to those of the naive Bayes model. *[6 marks]*

(b) A linear regression model using m basis functions $\boldsymbol{\phi}(\mathbf{x}) = (\phi_1(\mathbf{x}),\ldots,\phi_m(\mathbf{x}))^T$ has the form $f(\mathbf{x}) = \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x})$, where $\mathbf{w}$ is a vector of parameters. In a Bayesian treatment we place a prior $p(\mathbf{w})$ on the parameters, for example a $N(0,\tau^2)$ Gaussian prior so that

$$p(\mathbf{w}|\tau) = \frac{1}{(2\pi\tau^2)^{m/2}} \exp\left(-\frac{\mathbf{w}^T\mathbf{w}}{2\tau^2}\right).$$

You are provided with a dataset of input-output pairs $\mathcal{D} = \{(\mathbf{x}^n, y^n)\}$, $n = 1,\ldots,N$, and use a Gaussian noise model to define the likelihood term $p(\mathcal{D}|\mathbf{w})$. The *maximum a posteriori* or MAP solution for $\mathbf{w}$ is obtained as the one that maximizes $p(\mathbf{w}|\mathcal{D})$.

Suppose that the input data dimensionality is $D = 1$, and that the basis functions used are polynomials up to a power of x^{m-1}, with $\phi_1(x) = 1$. The data is actually drawn from a a polynomial of power $p < m-1$, but each datapoint has noise added.

i. Explain the effect of the prior on the MAP solution in the case that (1) $\tau \to \infty$, (2) $\tau \to 0$, and (3) τ has an intermediate value in $(0,\infty)$. *[4 marks]*

ii. Define the term *generalization error*. Sketch the form of the generalization error for this model as a function of τ, and explain why it has this form. *[4 marks]*

iii. A full Bayesian treatment obtains a posterior distribution over the parameters, while the MAP solution selects one particular parameter vector. Do these two approaches make the same predictions, or is there at least some aspect in which they are different? Explain your answer. *[3 marks]*

QUESTION CONTINUES ON NEXT PAGE

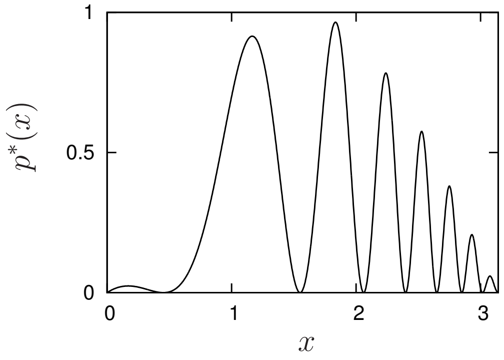
(c) You play a game with your friend Alice where you bet on the outcome of a coin toss. The coin has been provided by Alice. You think there is a 50% chance that she would have provided and unfair coin. If the coin is unfair then you believe that the probability that it will turn up heads is uniform in $[0, 1]$.

i. You toss the coin once and it comes up heads. What is the probability that the coin is fair? (You should compute this answer numerically). [*3 marks*]

ii. You toss the coin for a second time and it comes up heads again. Now what is the probability that the coin is fair? (You should compute this answer numerically). Interpret this result. [*5 marks*]

3. Consider a probability distribution with density function $p(x) = \frac{1}{Z}p^*(x)$, where Z is a constant, and

$$p^*(x) = \begin{cases} \sin(x)\cos^2(e^x) & x \in [0, \pi] \\ 0 & \text{otherwise.} \end{cases}$$



(a) Describe a practical procedure to draw independent samples from this distribution. You should not assume that you are able to evaluate Z. [*4 marks*]

(b) Write down a mathematical expression for the constant Z, then describe a Monte Carlo procedure to estimate it. [*4 marks*]

(c) Explain whether a standard application of the Laplace approximation would underestimate or overestimate Z in this case. [*4 marks*]

(d) The Kullback–Leibler divergence is defined as:

$$D_{\text{KL}}(r||s) = \int r(x)\log\frac{r(x)}{s(x)}\,\mathrm{d}x,$$

where the integrand is defined to be zero for any x where $r(x)$ is zero.

Sketch the Gaussian approximation $q(x) = \mathcal{N}(x;\, \mu, \sigma^2)$ to the distribution $p(x)$ defined above that minimizes $D_{\text{KL}}(p||q)$. Make a rough copy of the diagram above for your sketch. You should label how the position and scale of the Gaussian on your sketch relate to its parameters. You do not need to do any calculations; your sketch only needs to be qualitatively correct. [*3 marks*]

(e) Why can we not fit a Gaussian approximation q to the distribution p above by minimizing $D_{\text{KL}}(q||p)$? [*2 marks*]

The remainder of the question considers Markov chain Monte Carlo (MCMC) algorithms on discrete distributions on positive integers, rather than the distribution above.

(f) Consider the following Markov chain algorithm for transitioning from a current state x_s to a new state x_{s+1}:

A proposal is made,

$$x' \sim \text{Uniform}\{x_s - 1,\, x_s + 1\}.$$

QUESTION CONTINUES ON NEXT PAGE

If x' is outside the set $\{1, 2, 3, 4, 5\}$, a new state is proposed from the same distribution until an acceptable state $x' \in \{1, 2, 3, 4, 5\}$ is proposed. Then the new state is recorded:

$$x_{s+1} \leftarrow x'$$

i. Show that this procedure does *not* leave a uniform distribution on the integers $\{1, 2, 3, 4, 5\}$ invariant/stationary. *[3 marks]*

ii. Give a correct MCMC algorithm for a uniform target distribution on $\{1, 2, 3, 4, 5\}$, making use of the same core proposal mechanism: $x' \sim \text{Uniform}\{x_s - 1,\ x_s + 1\}$. *[3 marks]*

iii. Explain whether your algorithm would be valid MCMC method for a stationary distribution that is uniform on the integers $\{1, 2, 3, 4, 5, 10, 11, 12\}$. *[2 marks]*