UNIVERSITY OF EDINBURGH

COLLEGE OF SCIENCE AND ENGINEERING

SCHOOL OF INFORMATICS

INFR11073 MACHINE LEARNING AND PATTERN RECOGNITION (LEVEL 11)

Monday 27 th April 2015

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. Cohn, T. Field

THIS EXAMINATION WILL BE MARKED ANONYMOUSLY

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

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

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

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

   (a) Consider a multivariate Gaussian $p(\mathbf{x} \,|\, \boldsymbol{\mu}, \Sigma)$ with mean $\boldsymbol{\mu}$ and covariance matrix Σ in $D=2$ dimensions. Assume that $\boldsymbol{\mu} = \mathbf{0}$. Sketch the contours of $p(\mathbf{x} \,|\, \boldsymbol{\mu}, \Sigma)$ if Σ is:

   i. A multiple of the identity matrix,

   ii. A general diagonal matrix,

   iii. A general positive definite matrix.

   iv. Also specify what type of covariance matrix would be appropriate if a Gaussian distribution is used as the class-conditional model in a Naive Bayes classifier. *[5 marks]*

   (b) Explain the terms *training error* and *generalization error*. Explain what it means if a hypothesis f is said to *overfit* the training data. *[4 marks]*

   (c) The Kullback–Leibler divergence between two univariate probability density functions $p(x)$ and $q(x)$ is given by

$$\mathrm{KL}(p \,||\, q) = \int p(x) \log_e \frac{p(x)}{q(x)}\, \mathrm{d}x.$$

   If $p(x) = \mathcal{N}(x;\, 0,\, \sigma^2)$ and $q(x) = \mathcal{N}(x;\, \mu,\, \sigma^2)$, calculate $\mathrm{KL}(p \,||\, q)$ and interpret the result. *[5 marks]*

   (d) The prior over a function f is a Gaussian process with zero mean and squared exponential covariance:

$$k(\mathbf{x}^i, \mathbf{x}^j) = \sigma_f^2 \exp\left(-\frac{1}{2}\sum_{d=1}^{D}(x_d^i - x_d^j)^2/\ell_d^2\right).$$

   N independent noisy observations of the function $\mathbf{y} = \{y^n\}$ are made at N corresponding inputs $\{\mathbf{x}^n\}$, such that $p(y^n \,|\, \mathbf{x}^n, f) = \mathcal{N}(y^n;\, f(\mathbf{x}^n),\, \sigma_n^2)$. The free parameters σ_f, $\{\ell_d\}_{d=1}^D$, and σ_n (also called hyperparameters) are assumed known.

QUESTION CONTINUES ON NEXT PAGE

i. Draw a sketch to illustrate the model for one-dimensional inputs. Include and label a typical function simulated from the prior, $N=5$ observations, and indicate what parts of the sketch are controlled by each of the free parameters of the model.

ii. Assuming the function is unknown but drawn from the prior, what is the model's distribution over outputs: $p(\mathbf{y} \mid \{\mathbf{x}^n\})$? *[6 marks]*

(e) We have an $N \times D$ data matrix X, containing N samples of D-dimensional vectors. One way to implement Principal Components Analysis (PCA), assuming each column has been centred to have zero mean, starts by finding the Singular Value Decomposition (SVD):

$$X = USV^\top,$$

where S is a diagonal matrix of singular values ordered from largest to smallest.

i. From the above matrices, how do we obtain the first two principal directions in the D-dimensional space, and the 2-dimensional PCA embedding of the data?

ii. Give a sense in which PCA gives the best low-dimensional linear projection of the data. *[5 marks]*

2. (a) Let $\mathcal{D} = \{x^1, \ldots, x^N\}$ be N independent samples drawn from a Gaussian with mean μ and variance σ^2. We have that

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

i. Write down the log likelihood of the data under the model. Hence show that $\hat{\mu} = \frac{1}{N}\sum_{n=1}^{N} x^n$, where $\hat{\mu}$ denotes the maximum likelihood estimator for μ. *[3 marks]*

ii. Derive the maximum likelihood estimator for σ^2. *[3 marks]*

iii. Now consider a Bayesian treatment of this problem for the inference of μ, assuming that σ^2 is known *a priori*. In this case a conjugate prior for μ is a Gaussian distribution $\mu \sim \mathcal{N}(\mu_0, \sigma_0^2)$. Explain what a conjugate prior is. *[1 mark]*

iv. Argue that the posterior distribution for μ is a Gaussian. You are *not* required to determine the specific values of the mean and variance of this posterior distribution. HINT: you can work with the quadratic forms in the exponent of the Gaussian distributions, and ignore details of the normalization factors. *[3 marks]*

(b) An internet startup company asks you to predict if an input image contains a cat or not. They provide you with a dataset with one million labelled images $\mathcal{D} = \{(\mathbf{x}^n, y^n)\}$ $n = 1, \ldots, N$, where $\mathbf{x}^n$ is an input image, and y^n is its corresponding 0/1 label.

You decide to train a multilayer neural network using sigmoid nonlinearities for this task.

i. Describe the form of the network you would use, paying particular attention to the activation function of the output unit. Explain how an input $\mathbf{x}$ is processed through the network. *[4 marks]*

ii. The network has parameters $\mathbf{w}$, and for input case $\mathbf{x}^n$ the output is $f(\mathbf{x}^n)$. The negative log likelihood is defined as

$$E(\mathbf{w}) = -\sum_{n=1}^{N} [y^n \log f(\mathbf{x}^n) + (1 - y^n) \log(1 - f(\mathbf{x}^n))].$$

Explain the difference between *batch* and *online* optimization methods. A function is available to compute the derivatives of E with respect to $\mathbf{w}$ by backpropagation. Describe a method that you would use to train the network. Explain why it is important to start training with small weights. *[5 marks]*

QUESTION CONTINUES ON NEXT PAGE

iii. There are many architectural choices that you could make in the network, e.g. the number of hidden layers, and the number of units in each layer. Describe how you would select a network architecture with the aim of obtaining good performance on an unseen test set. [*2 marks*]

iv. Compare and contrast the multilayer neural network with a model based on logistic regression, in terms of network architecture, the optimization problem, and the performance you would expect to obtain. [*4 marks*]

3. A linear model of a function states $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$. There is a prior over the weights:

$$P(\mathbf{w}) \;=\; \mathcal{N}\big(\mathbf{w};\, \mathbf{0},\, \tfrac{1}{\tau} I\big)\,.$$

Our data $\{\mathbf{x}^n, y^n\}$ consist of N independent noisy observations $\{y^n\}$ at input locations $\{\mathbf{x}^n\}$. An unobserved indicator variable is chosen independently for each datapoint:

$$P(z^n) \;=\; \text{Bernoulli}(z^n;\, g) \;=\; g^{z^n}\,(1-g)^{1-z^n}, \quad z^n \in \{0, 1\},$$

where the parameter g is known. For each data-point, the function is either (when $z^n = 0$) observed with Gaussian noise of known variance σ^2:

$$P(y^n \,|\, \mathbf{w}, \mathbf{x}^n, z^n = 0) \;=\; \mathcal{N}(y^n;\, \mathbf{w}^\top \mathbf{x}^n,\, \sigma^2),$$

or (when $z^n = 1$) the observation is drawn from a broad background distribution with known variance s^2, ignoring the input:

$$P(y^n \,|\, \mathbf{w}, \mathbf{x}^n, z^n = 1) \;=\; \mathcal{N}(y^n;\, 0,\, s^2).$$

Thus, on average, a fraction g of the observations don't relate to the function. The joint distribution of the unknowns and data is:

$$P(\{y^n, z^n\}, \mathbf{w} \,|\, \{\mathbf{x}^n\}) \;=\; P(\mathbf{w}) \prod_{n=1}^{N} P(z^n)\, P(y^n \,|\, \mathbf{w}, \mathbf{x}^n, z^n).$$

(a) Describe a Gibbs sampling procedure that could be used to approximately sample from the joint posterior $P(\mathbf{w}, \{z^n\} \,|\, \{\mathbf{x}^n, y^n\})$ for this model. Derive and include the details of the update for the indicator variables $\{z^n\}$, including how it is implemented given access to samples from Uniform$[0, 1]$. For the weights $\mathbf{w}$, describe what the Gibbs sampling updates should achieve, but you need not derive the details. *[6 marks]*

(b) What would happen to the samples of the weights $\mathbf{w}$ if the prior distribution is taken towards the limit $\tau \to \infty$? (Do not assume this limit is taken in other parts of this question.) *[1 mark]*

(c) Show how S posterior samples of the weights $\{\mathbf{w}^s\}_{s=1}^S$ from Gibbs sampling can be used to approximate the posterior probability $P(y^* \,|\, \mathbf{x}^*, \{\mathbf{x}^n, y^n\})$ of a test output y^* at a test input location $\mathbf{x}^*$.

HINT: if we knew the weights $\mathbf{w}$, the predictive probability is available in closed form, $P(y^* \,|\, \mathbf{x}^*, \mathbf{w}) = (1-g)\,\mathcal{N}(y^*;\, \mathbf{w}^\top \mathbf{x}^*,\, \sigma^2) \;+\; g\,\mathcal{N}(y^*;\, 0,\, s^2)$. *[3 marks]*

(d) Explain a way in which predictions made using the model will differ when i) using Gibbs sampling as in the previous part, and ii) assuming a point estimate of the weights $\widehat{\mathbf{w}}$, such as a MAP estimate. *[3 marks]*

QUESTION CONTINUES ON NEXT PAGE

In this model, the indicator variables can be summed out analytically, giving the joint probability of the weights and outcomes:

$$P(\{y^n\}, \mathbf{w} \,|\, \{\mathbf{x}^n\}) \;=\; P(\mathbf{w}) \prod_{n=1}^{N} \left[(1-g)\mathcal{N}(y^n;\, \mathbf{w}^\top \mathbf{x}^n,\, \sigma^2) \;+\; g\mathcal{N}(y^n;\, 0,\, s^2)\right].$$

(e) Describe how to use importance sampling to estimate expectations of a function $h(\mathbf{w})$ under the marginal posterior over the weights, $P(\mathbf{w} \,|\, \{\mathbf{x}^n, y^n\})$. What problems might there be if the weight vector is high-dimensional? [*5 marks*]

(f) Name a Monte Carlo method that could be used to draw samples from $P(\mathbf{w} \,|\, \{\mathbf{x}^n, y^n\})$ given the ability to evaluate $P(\{y^n\}, \mathbf{w} \,|\, \{\mathbf{x}^n\})$ as given above. State any parameters or settings of this algorithm that a user would need to specify. [*2 marks*]

(g) Name a method that does not involve Monte Carlo, that could be used to approximate the posterior $P(\mathbf{w} \,|\, \{\mathbf{x}^n, y^n\})$ with a Gaussian distribution. What principle does this method use to choose the Gaussian distribution? [*2 marks*]

(h) Discuss the relative merits of making predictions using the Gaussian approximation to the posterior of part (g), or using Monte Carlo samples, as in part (c). [*3 marks*]