UNIVERSITY OF EDINBURGH

COLLEGE OF SCIENCE AND ENGINEERING

SCHOOL OF INFORMATICS

INFR11130 MACHINE LEARNING AND PATTERN RECOGNITION

Wednesday 3rd May 2017

14:30 to 16:30

INSTRUCTIONS TO CANDIDATES

Answer QUESTION 1 and ONE other question.

Question 1 is COMPULSORY. If both QUESTION 2 and QUESTION 3 are answered, only QUESTION 2 will be marked.

All questions carry equal weight.

CALCULATORS MAY NOT BE USED IN THIS EXAMINATION

MSc Courses

Convener: F. Keller
External Examiners: A. Burns, P. Healey, M. Niranjan

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. THIS QUESTION IS COMPULSORY

   (a) Explain the terms *training error* and *generalization error*. How can we estimate the generalization error of a model fitted to training data? [*4 marks*]

   (b) A logistic regression classifier is fitted to some training data. The training error is higher than required for the application, and so a team discusses how the classifier could be improved. One team member suggests that additional features could be created, based on the original features, so that the classifier has a more flexible decision boundary. They suggest using Principal Components Analysis (PCA) to create these features.

   Explain whether this suggestion could work. If the suggestion won't work, state an alternative that would give a more flexible decision boundary. [*4 marks*]

   (c) Two logistic regression classifiers are fitted, one with L1 regularization and the other with L2 regularization.

      i. State a difference that is usually observed between the weight vectors that result from these two regularizers.

      ii. State one advantage and one disadvantage of L1 regularization compared to L2 regularization. [*3 marks*]

   (d) Data from some medical records contains the following information for each individual in the dataset: name, address, phone number, age, sex, occupation, ethnicity, medical history: times and dates of all medical consultations with an anonymized summary of what was found.

   A researcher anonymizes the data by removing the name, address and phone number entries. Explain why it is still not ethically acceptable to release the data, including a specific example of what could go wrong. [*3 marks*]

   (e) You are given the compiled code for two functions; you can run these functions but don't have access to their source code. The first routine takes a matrix W as input and evaluates a scalar function $f(W)$. The second routine also takes a matrix W and claims to return a matrix of partial derivatives with elements $G_{ij} = \frac{\partial f}{\partial W_{ij}}$ evaluated at the input matrix. Describe a sensible test procedure that could reveal if the code for the derivatives is inconsistent with the code for the function. [*3 marks*]

QUESTION CONTINUES ON NEXT PAGE

(f) In a regression task, a user states that any prediction error within a tolerance t is no problem, and that any larger mistakes result in a constant cost c. The loss on one input-output example $(\mathbf{x}, y)$ on a model of a differentiable function f with parameters $\mathbf{w} = [w_1, w_2, \dots w_D]^\top$ can be written:

$$L(y, f(\mathbf{x}; \mathbf{w})) = \begin{cases} 0 & (y - f(\mathbf{x}; \mathbf{w}))^2 < t^2 \\ c & \text{otherwise.} \end{cases}$$

Explain the problem with attempting to fit a model to this loss function in a gradient-based optimizer using the derivatives $\frac{\partial L}{\partial w_d}$. *[3 marks]*

(g) *Stochastic variational inference* (SVI) is applied to a model with parameters $\mathbf{w}$ with prior $p(\mathbf{w})$, and likelihood $p(\mathcal{D} \mid \mathbf{w})$ given data $\mathcal{D}$. SVI fits a distribution over parameters, $q(\mathbf{w}) = \mathcal{N}(\mathbf{w}; \mathbf{m}, V)$, by minimizing a cost function $C(\mathbf{m}, V)$ based on the KL-divergence.

i. What distribution does the variational cost function compare $q(\mathbf{w})$ to?

ii. Often V is chosen to be diagonal. Given the partial derivatives $\frac{\partial C}{\partial V_{dd}}$, find an expression for $\frac{\partial C}{\partial r_d}$, where $r_d = \log V_{dd}$.

iii. Briefly explain why we would optimize $r_d = \log V_{dd}$ in a stochastic gradient based optimizer, rather than fitting V_{dd} directly. *[5 marks]*

2. ANSWER EITHER THIS QUESTION OR QUESTION 3

(a) A binary classifier outputs a prediction $f_n \in [0,1]$ for the nth training case. The training labels $\{y_n\}$ are each 0 or 1. State a major difference between minimizing the squared error E_s, rather than the negative log-probability loss E_l, where

$$E_s = \sum_n (y_n - f_n)^2, \qquad E_l = -\sum_n \left[y_n \log f_n + (1-y_n) \log(1-f_n) \right].$$

Briefly explain how the difference can be seen as an advantage or disadvantage for the squared loss. *[3 marks]*

(b) Sketch and label a plot showing five contours of the sigmoidal function

$$\phi(\mathbf{x}) = \sigma(\mathbf{w}^\top \mathbf{x} + b),$$

for $\mathbf{w} = [0\ 2]^\top$ and $b = 4$, where $\sigma(a) = 1/(1+\exp(-a))$ and $\sigma(4) \approx 0.98$.

Indicate the precise location of the $\phi = 0.5$ contour on your sketch. The identify of the other contours should be labelled, but their precise locations need not be. *[6 marks]*

(c) An approximate Bayesian approach to logistic regression makes predictions given training data $\mathcal{D}$ for an unknown label y at inputs $\mathbf{x}$ using:

$$P(y=1 \,|\, \mathbf{x}, \mathcal{D}) \;\approx\; \int \sigma(\mathbf{w}^\top \mathbf{x})\, \mathcal{N}(\mathbf{w};\, \mathbf{m}, V)\, \mathrm{d}\mathbf{w}.$$

i. Name two methods that can provide the Gaussian parameters $\mathbf{m}$ and V for the equation above.

ii. The above integral can't be computed in closed form. Write down a simple "Monte Carlo" or sampling-based approximation to this integral.

iii. A deterministic approximation to the integral is

$$P(y=1 \,|\, \mathbf{x}, \mathcal{D}) \;\approx\; \sigma(\kappa(\mathbf{x})\, \mathbf{m}^\top \mathbf{x}), \qquad \kappa(\mathbf{x}) = \frac{1}{\sqrt{1 + \frac{\pi}{8} \mathbf{x}^\top V \mathbf{x}}}\,.$$

Let $\mathbf{v}$ be a non-zero vector perpendicular to $\mathbf{m}$, and $\mathbf{x}^{(*)}$ be a test location. Consider the line of test locations $\mathbf{x} = \mathbf{x}^{(*)} + c\,\mathbf{v}$. What happens to the Bayesian prediction (at least as given by the deterministic approximation) as $c \to \infty$? What happens to the predictions as $c \to \infty$ if we instead assume the mean parameters are correct: $P(y=1 \,|\, \mathbf{x}, \mathbf{w}=\mathbf{m}) = \sigma(\mathbf{m}^\top \mathbf{x})$? *[7 marks]*

QUESTION CONTINUES ON NEXT PAGE

(d) The first feature in some data, x_1, is an integer taking values $\{1, 2, 3\}$ indicating whether an object is 'red', 'blue', or 'green'. Explain the problem with using this feature in a logistic regression classifier. How would you fix this problem? *[3 marks]*

(e) A one-dimensional curve fitting method transforms a scalar input $x \in [0, 1]$ into a vector:

$$\boldsymbol{\phi}(x) = [1 \; r(x; 0.0) \; r(x; 0.2) \; r(x; 0.4) \; r(x; 0.6) \; r(x; 0.8) \; r(x; 1.0)]^\top,$$

where $r(x; c)$ is a radial basis function centered at location c. The weights $\mathbf{w}$ are fitted by minimizing a regularized least squares cost over N training cases:

$$E(\mathbf{w}) = \sum_{n=1}^{N} (\mathbf{w}^\top \boldsymbol{\phi}(x_n) - y_n)^2 + \lambda \mathbf{w}^\top \mathbf{w},$$

where the y_n are the target outputs and λ is a regularization constant.

i. Describe a procedure for selecting λ, giving precise steps or pseudo-code.

ii. What problem does a user avoid by centering the outputs (subtracting $\frac{1}{N} \sum_n y_n$ from all outputs) before fitting this model? How could the cost function be modified to remove the need for centering? *[6 marks]*

3. ANSWER EITHER THIS QUESTION OR QUESTION 2

In the first part of this question we consider a dataset of I pairs $\{(t_i, z_i)\}_{i=1}^{I}$, where each pair gives a scalar observation z_i made at time t_i.

A Gaussian process model for $\mathbf{z} = [z_1 \; z_2 \; \cdots \; z_I]^\top$ given $\mathbf{t} = [t_1 \; t_2 \; \cdots \; t_I]^\top$ is:

$$p(\mathbf{z} \,|\, \mathbf{t}, \theta) \;=\; \mathcal{N}(\mathbf{z};\, \mathbf{0},\, K + \sigma_n^2 \mathbb{I}), \qquad \text{with } K_{ij} = k(t_i, t_j;\, \theta),$$

where $\mathbb{I}$ is the identity matrix, the kernel function is

$$k(t_i, t_j;\, \theta) \;=\; \sigma_f^2 \exp\!\left(-\tfrac{1}{2\ell^2}(t_i - t_j)^2\right),$$

and $\theta = \{\sigma_f, \ell, \sigma_n\}$ are the parameters of the model.

The posterior predictive distribution for an output at time t_* given data and model is:

$$p(z_* \,|\, t_*, \mathbf{z}, \mathbf{t}, \theta) \;=\; \mathcal{N}(z_*;\, m, s^2),$$

where we can compute m and s^2 for any test time, training data, and parameters.

(a) Consider the posterior predictive distribution given above, evaluated with the test time set to the time of the first observed pair, $t_* = t_1$. Write down, with a brief reason, both the predictive mean and variance, m and s^2,

i. in the limit $\sigma_n^2 \to 0$,

ii. in the limit $\sigma_n^2 \to \infty$.

No derivations are required. *[6 marks]*

(b) If m_i is the posterior mean at the ith observed training time, $t_* = t_i$, we can write down a square error as:

$$E_s = \sum_{i=1}^{I} (m_i - z_i)^2.$$

Briefly explain why this error is not a good cost function for optimizing σ_n. *[2 marks]*

(c) What is a suitable cost function that we could minimize in a standard gradient-based optimizer to set the parameters θ? *[2 marks]*

(d) State another method that could be used to predict the output z_* at a new time t_*, and give an advantage and disadvantage compared to the Gaussian process approach. *[3 marks]*

QUESTION CONTINUES ON NEXT PAGE

In the remainder of this question we consider N labelled time series. The nth time series has I_n observations $\mathbf{x}^{(n)} = \{(t_i^{(n)}, z_i^{(n)})\}_{i=1}^{I_n}$. Each time series has a single binary label $y_n \in \{0, 1\}$. The time series are not all the same length; the number of observations I_n depends on the particular instance n. For a new unlabelled sequence $\mathbf{x} = (\mathbf{t}, \mathbf{z})$ we want to assign a probability to its label, y.

(e) Why is it not possible to fit a straightforward logistic regression model to address the classification task described above? [*2 marks*]

One way to model the training data is to assume that each sequence came from a Gaussian process as described in the first part. A parameter vector $\theta^{(0)}$ can be fitted to minimize a suitable cost function, as in (c), summed over the sequences where $y_n = 0$. A parameter vector $\theta^{(1)}$ can be fitted to the sequences with $y_n = 1$.

(f) Given the class-specific models with parameters $\theta^{(0)}$ and $\theta^{(1)}$ described above, we can build a Bayes classifier. The classifier has parameters α, and predicts the label for a new sequence with:

$$P(y \mid \mathbf{z}, \mathbf{t}, \alpha) \propto P(y, \mathbf{z} \mid \mathbf{t}, \alpha).$$

We assume that the times in the test sequence $\mathbf{t}$ are known and that the class-specific models only model the test observations $\mathbf{z}$ given the times.

i. What parameter(s) need to be included in α in addition to $\theta^{(0)}$ and $\theta^{(1)}$, and how can it/they be set?

ii. Describe how to calculate $P(y = 1 \mid \mathbf{z}, \mathbf{t}, \alpha)$, giving enough detail so that it is clear how to calculate each term in your equation(s). [*5 marks*]

(g) When the training data are inspected, it appears that the main difference between the two classes is that they have different trends. In class zero the observations tend to increase after time zero, and in class one the observations tend to decrease after time zero.

i. Briefly explain why a Bayes classifier with the Gaussian process model described above cannot capture this difference between the classes. Refer to the differences that can be captured by the model's parameters in your explanation.

ii. Briefly describe another classifier that could be applied to these time series, and could capture trends. [*5 marks*]