UNIVERSITY OF EDINBURGH

COLLEGE OF SCIENCE AND ENGINEERING

SCHOOL OF INFORMATICS

INFR11130 MACHINE LEARNING AND PATTERN RECOGNITION

Tuesday 18th December 2018

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: M.Mistry
External Examiners: W. Knottenbelt, M. Dunlop, M. Niranjan, E. Vasilaki

THIS EXAMINATION WILL BE MARKED ANONYMOUSLY

The following information might 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) This part is about methods commonly-used to pre-process inputs:

i. State what type of input feature "one-hot encoding" is used for, and briefly describe what the encoding is.

ii. What is the "naive Bayes" assumption? Briefly explain why one-hot encoding an input feature to a classifier would break this assumption.

iii. State a classifier where you would use "one-hot encoding" for inputs of the appropriate type.

iv. A Gaussian classifier is used to predict if an employee is likely to leave a company. One of the features used is the salary of the employee. State a transformation/pre-processing you would you try for this feature and give a reason. *[7 marks]*

(b) State some advantages and disadvantages of using Principal Components Analysis (PCA) for dimensionality reduction compared to a non-linear auto-encoder. Your answer should contain at least 3 distinct points, including at least one advantage and one disadvantage of PCA. *[3 marks]*

(c) In this part, assume that $\mathbf{x}$ is a two-dimensional vector of input features.

i. Define a vector-valued function $\boldsymbol{\phi}(\mathbf{x})$, such that $f(\mathbf{x}) = \mathbf{w}^\top\boldsymbol{\phi}(\mathbf{x})$ is a general multivariate quadratic function of the original input features $\mathbf{x}$.

ii. We can fit the coefficients $\mathbf{w}$ of the general multivariate quadratic to N training examples $\{(\mathbf{x}^{(n)}, y^{(n)})\}_{n=1}^N$ using one of the following lines:

Matlab: `w = Phi \ y;  % size(y) is [N,1]`

Python: `w = np.linalg.lstsq(Phi, y)[0]  # y.shape is (N,)`

For *either* the Matlab *or* the Python, define what needs to be put in each of the elements of `Phi`. You don't have to write code.

iii. Write down an equation for the cost function that the above lines of code minimize.

iv. We are now told that the $\{y^{(n)}\}$ outputs that we need to fit are all positive, and that all future outputs will always be positive. Describe a way to modify the above approach so that the fitted function is guaranteed to be strictly positive. *[8 marks]*

QUESTION CONTINUES ON NEXT PAGE

(d) A colleague has fitted a classifier, which has similar average error on their training and validation sets. Your colleague believes it should be possible to get much better fits, and wants to try an ensemble method that uses their existing classification code and feature set. They are considering bagging or boosting. Explain which would be the better of these two options to try in this situation. [*2 marks*]

(e) Explain how by fitting a model to a dataset of users' ratings of movies, we can obtain a vector for each user and movie, which could be used for visualization or in other machine learning tasks.

You should include details of the model and a principle by which it can be fitted. As always, define any notation you introduce while explaining the ideas. There is no need to explain the details of a fitting algorithm. [*5 marks*]

2. ANSWER EITHER THIS QUESTION OR QUESTION 3

(a) Finite differences:

i. Given a loss function $L(W)$ for a weight matrix W, describe how to approximate the derivatives $\frac{\partial L}{\partial W_{ij}}$ by finite differences.

ii. Give two reasons that gradients for neural network training are usually computed using backpropagation (also known as reverse-mode differentiation) rather than by finite differences. [*5 marks*]

(b) Backpropagation: For any matrix A, we define $\overline{A}_{ij} = \frac{\partial L}{\partial A_{ij}}$, and for any vector $\mathbf{z}$, the vector $\bar{\mathbf{z}}$ has elements $\bar{z}_i = \frac{\partial L}{\partial z_i}$, where L is a loss function.

You may use the result that $C = AB$ implies $\overline{A} = \overline{C}B^\top$ and $\overline{B} = A^\top\overline{C}$.

i. A neural network contains a large $H \times D$ weight matrix W, used in an intermediate computation $\mathbf{h} = W\mathbf{x}$. Assume that the loss has been backpropagated to the output of this computation, so you know $\bar{\mathbf{h}}$.

Write down an expression for $\overline{W}$ in terms of $\bar{\mathbf{h}}$, $\mathbf{x}$ and/or W. Also write down the $O()$ cost of computing your expression.

ii. To save memory and computer time, the matrix W from the previous part is replaced with the product $W = UV$, where U is $H \times K$, V is $K \times D$, and the intermediate dimension is small: $K \ll H$ and $K \ll D$.

State how to compute $\mathbf{h}$ efficiently from $\mathbf{x}$, U, and V, and give the $O()$ cost for this computation.

iii. Find matrix-based expressions to compute $\overline{U}$ and $\overline{V}$ from $\bar{\mathbf{h}}$, $\mathbf{x}$, U, and V. For full marks, give computationally-efficient expressions. [*7 marks*]

(c) The KL-divergence between two density functions is a non-negative function:

$$D_{\mathrm{KL}}(r \,||\, s) = \int r(\mathbf{z}) \log \frac{r(\mathbf{z})}{s(\mathbf{z})}\, \mathrm{d}\mathbf{z}.$$

i. By comparing a reference distribution $q(\mathbf{w})$ to the posterior given data $p(\mathbf{w} \,|\, \mathcal{D})$, derive the following lower bound on the log marginal likelihood:

$$\log p(\mathcal{D}) \geq \mathbb{E}_{q(\mathbf{w})}[\log p(\mathbf{w}, \mathcal{D})] - \mathbb{E}_{q(\mathbf{w})}[\log q(\mathbf{w})],$$

where $\mathbb{E}_{q(\mathbf{w})}[\cdot]$ are expectations under the reference distribution.

ii. It is also possible to derive an upper-bound on the log marginal likelihood from the KL-divergence (you do not need to show this result):

$$\log p(\mathcal{D}) \leq \mathbb{E}_{p(\mathbf{w}\,|\,\mathcal{D})}[\log p(\mathbf{w}, \mathcal{D})] - \mathbb{E}_{p(\mathbf{w}\,|\,\mathcal{D})}[\log q(\mathbf{w})].$$

Why is this upper bound usually harder to estimate than the lower bound for models where we use variational inference? [*5 marks*]

QUESTION CONTINUES ON NEXT PAGE

(d) The distribution over a K-dimensional vector outcome $\mathbf{y}$, given a D-dimensional vector input feature vector $\mathbf{x}$, is modelled as a multivariate Gaussian:

$$p(\mathbf{y}\,|\,\mathbf{x},\theta) = \mathcal{N}(\mathbf{y};\,\mathbf{m}(\mathbf{x};\theta),\,V(\mathbf{x};\theta)),$$

where $\mathbf{m}$ and V are functions of the input $\mathbf{x}$, which can depend on some or all of the model parameters θ.

A function $\mathbf{h}(\mathbf{x};\theta)$ transforms the input into an H-dimensional 'hidden-layer' vector. All vectors in this question are column vectors.

i. We could choose to define the mean as $\mathbf{m}(\mathbf{x};\theta) = M\,\mathbf{h}(\mathbf{x};\theta)$, where the matrix M is one of the parameters in θ. Write down the dimensions of the matrix M.

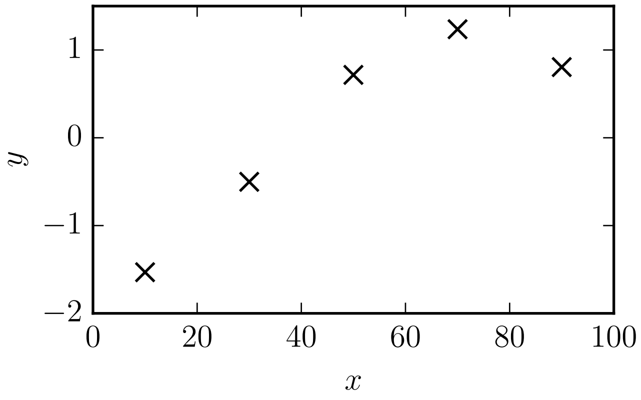
ii. Assuming we wish to model the elements of $\mathbf{y}$ as independent given the input $\mathbf{x}$, write down a suitable way we could choose to define $V(\mathbf{x};\theta)$ in terms of the hidden layer $\mathbf{h}$.

iii. For some flexible definitions of $\mathbf{m}$ and V, the maximum likelihood parameters θ will cause the model to predict one of the N training examples perfectly, and with very high confidence. That is, for some n, $\mathbf{m}(\mathbf{x}^{(n)}) = y^{(n)}$, and $|V(\mathbf{x}^{(n)};\theta)|$ is very small.

List three different possible ways to avoid obtaining parameters that fit only one or a few datapoints closely when training this model. [*8 marks*]

3. ANSWER EITHER THIS QUESTION OR QUESTION 2

Parts a) and b) relate to the plot below with $N = 5$ pairs $\{(x^{(n)}, y^{(n)})\}_{n=1}^{N}$.



Consider a linear regression model $f(x) = \mathbf{w}^{\top}\boldsymbol{\phi}(x)$, with $K = 100$ radial basis functions:

$$\phi_k(x) = \exp(-(x-k)^2/h^2), \quad \text{where } k = 1, 2, \ldots, K = 100,$$

where the "bandwidth" h is a free parameter.

(a) i. Explain why regularization is important for this model and data.

ii. Copy the plot above into your answer book (a rough copy is fine). Sketch what the functions $f(x)$ are likely to look like for fits, with a small L2 regularizer, of two models: one with $h = 1$, one with $h = 100$. [*5 marks*]

(b) i. Based on a Bayesian treatment of the model above, write down a conditional probability that we could maximize to optimize the bandwidth h. [Just specify which probability "$p(\ldots \,|\, \ldots)$", not its equation.]

ii. What other model assumptions would we need to specify (apart from those already given in the question) before the conditional probability you've given in i. is completely defined?

iii. We could also model the data with a Gaussian Process with kernel function: $k(x^{(1)}, x^{(2)}) = \exp(-(x^{(1)} - x^{(2)})^2/h^2)$. What would be the main qualitative difference between the probabilistic predictions of Bayesian linear regression using the basis functions above, and this Gaussian Process? [*7 marks*]

(c) A vector is sampled from a zero-mean multivariate Gaussian: $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \Sigma)$. This vector is transformed into $\mathbf{v} = B\mathbf{u}$, with a matrix B.

i. What is the distribution of the final outcome $\mathbf{v}$? Write down the covariance of $\mathbf{v}$ in terms of Σ and B.

ii. Given N vectors $\{\mathbf{u}^{(n)}\}_{n=1}^{N}$ sampled from the original distribution, write down how you would estimate one of the elements of the covariance, Σ_{ij}. [*4 marks*]

QUESTION CONTINUES ON NEXT PAGE

The remaining parts relate to a model of N experimental trials. In each independent trial a number of particles was recorded, these outcomes were $\{y^{(n)}\}_{n=1}^{N}$. The experimental conditions for each trial were summarized in corresponding vectors $\{\mathbf{x}^{(n)}\}_{n=1}^{N}$.

Each outcome is a count, modelled by a Poisson distribution, a standard probability distribution over non-negative integers, $y \in \{0, 1, 2, \ldots\}$:

$$\text{Poisson}(y; \lambda) = e^{-\lambda} \frac{\lambda^y}{y!}.$$

The parameter $\lambda > 0$ is known as the 'rate'.

In this context the rate is assumed to depend on the experimental conditions, according to $\lambda(\mathbf{x}; \mathbf{w}) = \exp(\mathbf{w}^\top \mathbf{x})$, where $\mathbf{w}$ are unknown parameters.

(d) i. Write down an equation for the negative log-likelihood of this model.

ii. Find the derivatives of the negative log-likelihood with respect to the parameters $\mathbf{w}$. [*6 marks*]

(e) It's common for count-based data to contain many more zeros relative to other count values than a Poisson model can explain. In this case, suppose we believe that the Poisson model predicts the number of particles well, but the device randomly fails to operate with probability f on each trial. When the device fails to operate correctly, it records a count of zero.

Write down a mathematical definition of $P(y \,|\, \mathbf{x}, \mathbf{w}, f)$, the distribution over the output for a given context $\mathbf{x}$, assuming the failure process described above. [*3 marks*]