UNIVERSITY OF EDINBURGH

FACULTY OF SCIENCE AND ENGINEERING

LFD1

Date: ?? June 2001 DRAFT **Time: ??:??-??:??**

DRAFT

This will describe the degree (MSc/AI4,etc.)

Examiners:	Name of external examiner	(External)
	Name of exam board chair	(Chair)

INSTRUCTIONS TO CANDIDATES

Answer TWO questions.

If you attempt three questions, cross out one answer; if you do not, then the examiners will cross out the last one you answered.

Each complete question carries equal weight and is marked out of 100. The parts of a question may not all be worth the same amount; the marks at the side of the questions indicate how these will normally be apportioned.

Write as legibly as possible.

THIS EXAMINATION WILL BE MARKED ANONYMOUSLY

1. You are hired as the researcher in a startup company specializing in image processing. You are hired to provide the theory, algorithms and eventually to build a hand-written digit classifier to be used for Postcode recognition by the Post Office, giving a value from 0 to 9 for the class of a novel testpoint. The training data is of the form $\boldsymbol{x}^\mu, \mu = 1, \ldots, P$, where each image $\boldsymbol{x}^\mu$ has a corresponding class label, $c^\mu \in \{0, 1, \ldots, 9\}$.

   (a) Your line manager is very worried. She argues that for the case of binary valued images, with $100 \times 100 = 10000$ pixels, there are in total 2^{10000} such binary images. She argues that this number is astronomically larger than the number of atoms in the universe and that therefore the measly 25000 training examples they have will be woefully inadequate for training. Explain how you might persuade your line manager that the situation may not be so hopeless, giving compelling reasons for your optimism. [10%]

   (b) Your line manager is so impressed with your argument that she decides to discard all the training data currently in the database, and busies herself with making 1000 training examples of her own fair hand. Is this a reasonable thing to do? Justify your answer. [10%]

   (c) As a first step, you decide to use the K nearest neighbour method (KNN) to classify a novel test point $\boldsymbol{x}^*$.

   Describe fully how the KNN algorithm works, including how to determine the optimal number of neighbours K to use. [20%]

   (d) Your line manager is pleased with your algorithm but is concerned that it performs as well as you say it does. How can you persuade her that it really performs according to your claims? [10%]

   (e) One morning your line manager is delighted to tell you that she now has more training data than ever before, indeed the training data consists of $P = 100,000$ real valued images. You estimate that your current KNN method is going to be too slow to do real time classification of digits and you decide to use PCA to increase classification speed.

   Describe fully and mathematically how to use PCA to replace an N dimensional vector $\mathbf{x}$ with an M dimensional vector $\mathbf{y}$ where $M < N$. [15%]

   Derive a formula for approximating the distance $(\mathbf{x}^a - \mathbf{x}^b)^2$ between two vectors $\mathbf{x}^a$ and $\mathbf{x}^b$ using their corresponding PCA representations $\mathbf{y}^a$ and $\mathbf{y}^b$. [20%]

   (f) Your line manager is pleased with your faster algorithm, which you claim provides 95% classification accuracy. She tells you that it is important to make sure that 99% are classified correctly in the end, even if this means that 10% of test images need to be classified by a human. She asks you to adapt your algorithm accordingly. Suggest an amendment to your algorithm and explain how you would decide whether or not to leave a novel test point to be classified by a human. [15%]

Question 2 is on the next page.

2. Consider independently and identically distributed (iid) training data $D = \{(\boldsymbol{x}^\mu, c^\mu), \mu = 1, \ldots, P\}$, $c^\mu \in \{0, 1\}$, where each datapoint $\boldsymbol{x}^\mu$ has dimension N.

   (a) Define Bayes' rule, and show how this can be used to make a classifier $p(c|\boldsymbol{x})$ using a density model for $p(\boldsymbol{x}|c)$ and a model for $p(c)$. Explain the general procedure for training such models using maximum likelihood. [20%]

   (b) Define fully and mathematically the Naive Bayes classification method for binary data $x_i \in \{0, 1\}$ and for two classes $c \in \{0, 1\}$, and show that the maximum likelihood estimate of the parameter $p(x_i = 1|c = 1)$ is proportional to the number of times attribute i is 1 for class 1 data. Similarly, show that $p(c = 1)$ is proportional to the number of times class 1 occurs in the data. [40%]

   (c) A local supermarket specializing in breakfast cereals decides to analyze the buying patterns of its customers.

   They make a small survey asking 6 randomly chosen people which of the breakfast cereals (Cornflakes, Frosties, Sugar Puffs, Branflakes) they like, and also asking for their age (older or younger than 60 years). Each respondent provides a vector with entries 1 or 0 corresponding to whether they like or dislike the cereal. Thus a respondent with (1101) would like Cornflakes, Frosties and Branflakes, but not Sugar Puffs.

   The older than 60 years respondents provide the following data :

$$(1000), (1001), (1111), (0001)$$

   For the younger than 60 years old respondents, the data is

$$(0110), (1110)$$

   A novel customer comes into the supermarket and says she only likes Frosties and Sugar Puffs. Using Naive Bayes trained with maximum likelihood, what is the probability that she is younger than 60? [40%]

Question 3 is on the next page.

3. As the leader of a team working on credit risk assessment for a large high street bank, you are assigned the task of building an algorithm to assess whether or not someone will be able to pay back a loan. The bank has historical information on many customers concerning their age (in years),profession (one of 30 jobs),income,marital status (married or single),amount borrowed (in pounds), and whether or not they paid back the loan successfully (1 successful, 0 unsuccessful). The bank needs a classifier that estimates the probability that a novel client will be able to repay the loan based on the client's age, profession, income, marital status and amount to be borrowed.

   (a) The data for each customer is an array, for example here are three datapoints from the database

   $$(46, lecturer, 25000, Single, 90000, 1)$$
   $$(19, student, 10000, Single, 30000, 0)$$
   $$(33, lawyer, 80000, Married, 200000, 1)$$

   To make a classifier you need to convert each entry into a numerical representation. You decide that the marital status is easy to encode, using a 1 for Married and 0 for Single. However, you recognize that there is a potential problem with how the 'profession' attribute is represented. Your colleague suggests that you use $lawyer = 1, student = 2, lecturer = 3, \ldots, accountant = 30$. Explain why this is not a good idea, and describe an alternative way to encode this attribute using $1 - of - M$ encoding. [15%]

   (b) Each original entry is now represented by a 33 dimensional vector $\mathbf{x}$ (representing the age, professsion, income, maritial status and amount borrowed) and a corresponding class label $c \in \{0, 1\}$ (representing whether or not the loan was repaid successfully). Your colleagues are concerned that this is very high dimensional data and urge you to consider using a dimension reduction technique. Explain why dimension reduction in this case would be inappropriate. [15%]

   (c) Your preprocessing step has now given you a dataset of the form $D = \{(\boldsymbol{x}^\mu, c^\mu), \mu = 1, \ldots, P\}$, $c^\mu \in \{0, 1\}$. You decide to use logistic regression to build a classifier. Describe mathematically the logistic regression classification method. (You do not need to describe in detail how to train the classifier).[20%]

   (d) For the logistic regression model with parameters θ and $\mathbf{w}$, show that the log likelihood of the training data is

   $$\sum_\mu c^\mu \log \sigma\left(\theta + \mathbf{w}^T \mathbf{x}^\mu\right) + (1 - c^\mu) \log\left(1 - \sigma\left(\theta + \mathbf{w}^T \mathbf{x}^\mu\right)\right)$$

   where $\sigma(x) = e^x/(1 + e^x)$.

   Calculate the derivative $\nabla_{\mathbf{w}} L$ and suggest a training algorithm to find the maximum likelihood solution for $\mathbf{w}$. [30%]

   (e) After careful training the logistic regression method performs fairly well. However, your bank is unhappy and requests that you design a more accurate method using the same training database. You realize that the logistic regression only models linear interaction of the attributes (that is, the classification depends only on summing the attributes x_i). Suggest a modification of logistic regression that can account for quadratic interaction of the attributes (the classifier depends on the sum of the attribute products $x_i x_j$). [20%]

LFD1

Brief notes on answers:

1. (a) Even though 25000 is a very small number compared to 2^10000, the point is that digits are not simply random point in a 10000 dimensional space. There is a great deal of regularity and constraint on the form that each digit can take, so that digits will occupy only a very small fraction of the space of all possible images. Indeed, humans are capable of learning digits based on only a small number of training examples, and there is therefore every reason to be optimistic that a machine could do the same.

   (b) If we wish to make a classifier that works well on a wide variety of peoples handwriting, we need training data that is representative of a wide variety of styles. Otherwise, the trained classifier may be appropriate for recognizing the handwriting of the line manager, but not necessarily anyone else.

   (c) For the nearest neighbour method $K = 1$, we search through the dataset to find the training datapoint $\boldsymbol{x}^n$ that is closest (using the Euclidean distance) to $\boldsymbol{x}^*$. We assign the label of $\boldsymbol{x}^*$ to be c^n. For $K > 1$, we find the K nearest neighbours of $\boldsymbol{x}^*$, and get their associated class labels. The majority class of these neighbours is taken as the class of $\boldsymbol{x}^*$.
   We can use a validation set to determine the optimal value of K, in which the classification of different values of K can be independently assessed.

   (d) We can use an independent test set of data, not used during the training process to assess the performance of the method.

   (e) Find the sample mean and covariance matrix of the data. Then calculate the M largest eigenvalues of the covariance matrix, and their corresponding eigenvectors, $\boldsymbol{e}^i, i = 1, \ldots, 20$. The sample mean $\boldsymbol{m}$ is given by

   $$\boldsymbol{m} = \frac{1}{P} \sum_\mu \boldsymbol{x}^\mu$$

   The covariance matrix is defined as

   $$\mathbf{S} = \frac{1}{P} \sum_\mu \boldsymbol{x}^\mu (\boldsymbol{x}^\mu)^T - \boldsymbol{m}\boldsymbol{m}^T$$

   If they define the biased or unbiased version of the covariance, either is fine.

   The lower dimensional data is then given by the projection, $y_i^\mu = (\boldsymbol{x}^\mu - \boldsymbol{m})^T \boldsymbol{e}^i, i = 1, \ldots 20, \mu = 1, \ldots, P$.

   (f) Using the approximations, we have

   $$(\boldsymbol{x}^a - \boldsymbol{x}^b)^2 \approx (\sum_i y_i^a \boldsymbol{e}^i - \sum_i y_i^b \boldsymbol{e}^i)^T (\sum_j y_j^a \boldsymbol{e}^j - \sum_i y_j^b \boldsymbol{e}^j)$$

   Due to the orthonormality of the eigenvectors, this is $\sum_i (y_i^a)^2 - 2y_i^a y_i^b + (y_i^b)^2 = (\mathbf{y}^a - \mathbf{y}^b)^2$

(g) The classification of the KNN method is based on finding the K nearest neighbours. If none of the neighbours is very close, this will result is potentially inaccurate classification. A simple method is therefore to use an independent testset, and set a threshold value. Measure the distance to the nearest neighbour for each testpoint to be classified, and discard this point if it is greater than the threshold. For the remaining undiscarded points, determine the classification. If this is not 99%, increase the threshold and repeat the procedure until a just sufficient value of the threshold has been found.

2. (a) Bayes : $p(a|b) = p(b|a)p(a)/p(b)$, hence $p(c|\boldsymbol{x}) = p(\boldsymbol{x}|c)p(c)/p(\boldsymbol{x})$. Since $p(\boldsymbol{x})$ is a constant, the class probability is proportional to $p(\boldsymbol{x}|c)p(c)$, and the proportionality constant can be determined by normalisation : $p(\boldsymbol{x}) = \sum_c p(\boldsymbol{x}|c)p(c)$.

We can use the likelihood to fit each distribution. Assuming iid data, this is given by

$$\prod_{\boldsymbol{x} \in class(j)} p(\boldsymbol{x}|c = j, \theta^j)$$

Taking the log is numerically convenient. We can then find the parameters θ^j by numerically maximimising this function.

(b) In Naive Bayes, the inputs are assumed to be independent given the class:

$$p(\boldsymbol{x}|c = 1) = \prod_{j=1}^{m} p(x_j|c = 1).$$

Classification is then given by comparing

$$\log \frac{p(c = 1|\boldsymbol{x})}{p(c = 2|\boldsymbol{x})} > 0$$

Or

$$\sum_j \log \frac{p(x_j|c = 1)}{p(x_j|c = 2)} + \log \frac{p(c = 1)}{p(c = 2)} > 0$$

For convenience, define $\gamma_i \equiv p(x_i = 1|c = 1)$, then the likelihood is

$$p(\boldsymbol{x}|c = 1) = \prod_{j,\mu} \gamma_i^{x_j^\mu} (1 - \gamma_j^{x_j^\mu})$$

Take logs :

$$L = \sum_{\mu,j} x_j^\mu \log \gamma_j + (1 - x_j^\mu) \log(1 - \gamma_j)$$

Differentiate wrt γ_i and rearrange gives

$$\gamma_i = \frac{1}{P} \sum_\mu x_i^\mu$$

where P is the number of datapoints. γ_i is thus the fraction of times that x_i is 1 in the dataset.

A completely analogous argument gives $p(c = 1) = \frac{1}{P} \sum_\mu c^\mu$, which is the fraction of times that the class is 1 in the dataset.

(c) The decision boundary is given when $p(c = 1|x) = 1/2$. Using this gives

$$\frac{1}{2} = \frac{p(x|1)p(c = 1)}{p(x|1)p(c = 1) + p(x|2)p(c = 2)}$$

Rearranging this expression and taking the logarithm gives the desired result.

(d) Looking at the data, the estimates using maximum likelihood are

$$p(C = 1|Young) = 0.5, p(F = 1|Young) = 1, p(SP = 1|Young) = 1, p(B = 1|Young) = 0$$

and

$$p(C = 1|Old) = 0.75, p(F = 1|Old) = 0.25, p(Sp = 1|Old) = 0.25, p(B = 1|Old) = 0.75$$

and $p(Young) = 2/6$ and $p(Old) = 4/6$. Plugging this into Bayes formula, we get

$$p(Young|C = 0, F = 1, SP = 1, B = 0) \propto 0.5 * 1 * 1 * 1/6$$

$$p(Old|C = 0, F = 1, SP = 1, B = 0) \propto 0.25 * 0.25 * 0.25 * 0.25 * 4/6$$

Using the fact that these probabilities sum to 1, this gives $p(Young|C = 0, F = 1, SP = 1, B = 0) = 64/65$

3. (a) Using the suggested encoding gives an explicit ordering of the professions so that, lawyer is close to student but not to accountant. This is inappropriate since we do not wish to make such a priori value judgements about the professions themselves, rather we wish the data simply to assign a posteriori how profession is related to the loan risk. $1 - of - M$ encoding would expand each training vector so that the profession attribute is represented by 30 dimensions, with a single 1 in the dimension for that profession. For example : $(1, 0, 0, \ldots, 0)$ represents lawyer, $(0, 1, 0, \ldots, 0)$ represents student. This does not assign any ordering amongst the professions and is a more suitable numeric encoding of the data.

(b) Dimension reduction is inappropriate since the vector is very sparse due to the 1-of-M encoding. If we have an even spread amongst the professions, then each dimension representing the 1-of-M encoded profession will be important. Furthermore, these directions are independent and orthogonal and therefore cannot be compressed in the sense that they are describable by a lower dimensional subspace.

(c) Logistic regession is such that

$$p(c = 1|\mathbf{x}) = \sigma\left(\theta + \mathbf{w}^T\mathbf{x}\right)$$

The decision boundary is linear. The parameters θ and $\mathbf{w}$ can be found by maximum likelihood training.

(d) Let $L(\boldsymbol{w})$ be a function that we wish to maximise. Gradient ascent performs the update

$$\boldsymbol{w}^{new} = \boldsymbol{w} + \eta\nabla_{\boldsymbol{w}}L$$

(e) Follows straight from the iid assumption that the likelihood is

$$\prod_{\mu} \sigma\left(\theta + \mathbf{w}^T \mathbf{x}^{\mu}\right)^{c^{\mu}} \left(1 - \sigma\left(\theta + \mathbf{w}^T \mathbf{x}^{\mu}\right)\right)^{1-c^{\mu}}$$

Just take logs to get the result. We could maximise the L by gradient ascent, $\mathbf{w}^{new} = \mathbf{w} + \eta \nabla_{\mathbf{w}} L$.

(f) We could model non-linear interactions by using a feature vector that expands the data into a higher dimensional space, $\boldsymbol{\phi}(\mathbf{x})$. This vector contains non-linear interaction between the components x_i, in this case all quadratic products of the components of $\boldsymbol{x}$. We can use then a logistic regression classifier as above simply be replacing $\mathbf{x}^{\mu}$ by $\boldsymbol{\phi}(\mathbf{x}^{\mu})$. Training proceeds as before.