Course DD2427 - Final Exam

Please answer 5 questions from **Part I** of the exam. You must get a score of at least 70% (with a minimum performance of 50% on each question) to pass this part of the question and get an E grade. If you get 85% or over on this part of the exam you will get at least a D grade. Each question in **Part I** and **Part II** carrys equal weight. To get a higher grade you must answer questions in **Part II**. To get an A grade you must answer 2 questions well in **Part II**.

Note if you attempt more than 5 or 2 questions in **Part I** and **Part II** respectively, I will use your best answers to compute your grade. Also you are allowed use a calculator,

The bold face number in brackets indicates the proportion of a question's total score accounted for by this sub-question.

Part I

Question 1: *Probability theory and Bayes' Rule*

a) **(.3)** Let x and y be random variables which follow a joint probability distribution $p(x, y)$. Say that $y \in \{0, 1\}$ is a binary variable, then answer the following questions:

   i) Define the conditional probability, $p(x|y)$, in terms of $p(y)$ and $p(x, y)$.

   ii) Write $p(x)$ in terms of $p(x|y)$ and $p(y)$.

   iii) State and derive Bayes' Rule.

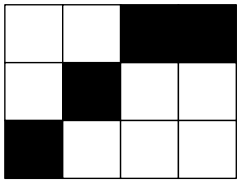
b) **(.7)** Consider a binary 3×4 image of a scene with a black horizontal line. In the noise free image all pixels are white except for one horizontal row with black pixels. Unfortunately, the camera used is far from perfect. Errors in different pixels are independent with

$$P(\text{observe white pixel} \mid \text{line}) = P(\text{observe black pixel} \mid \text{not line}) = \epsilon$$

Assume the prior probabilities for the location of the line are as follows:

$$P(\text{line on row 1}) = .3, \;\; P(\text{line on row 2}) = .4, \;\; P(\text{line on row 3}) = .3$$

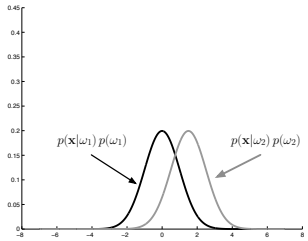
Given that you observe this image I



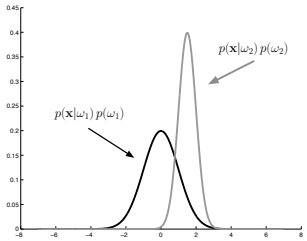
and $\epsilon = .3$, calculate $P(\text{line on row } i \mid I)$ for $i = 1, 2, 3$. Then what would be your decision as to which row the black line is on?

Question 2: *Bayes' Decision Theory*

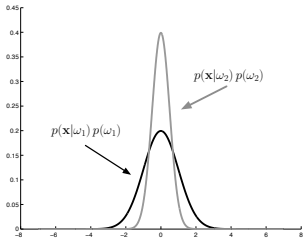
a) **(.7)** Assume you have a two class classification problem. Each class generates a one dimensional feature vector according to $p(x|\omega_i) = \mathcal{N}(\mu_i, \sigma_i^2)$ for $i = 1, 2$. The prior probabilities for each class are $p(\omega_1) = p(\omega_2) = .5$. In the graphs below $p(x|\omega_i)\, p(\omega_i)$ for $i = 1, 2$ are shown for different values of the μ's and σ's. For each example $\mu_1 = 0, \sigma_1 = 1$ and then **1**) $\mu_2 = 1.5, \sigma_2 = 1$, **2**) $\mu_2 = 1.5, \sigma_2 = .5$ and **3**) $\mu_2 = 0, \sigma_2 = .5$



(**1**) $\mu_1 \neq \mu_2, \sigma_1 = \sigma_2$



(**2**) $\mu_1 \neq \mu_2, \sigma_1 > \sigma_2$



(**3**) $\mu_1 = \mu_2, \sigma_1 > \sigma_2$

i) For the two-class problem how is the *Bayes' Classifier* defined?

ii) In the figure draw the decision boundaries/boundary defined by a Bayes' classifier.

iii) For case (**2**) explicitly calculate the decision boundaries.

iv) For case (**2**) write down the $P(\text{error})$ for the *Bayes' Classifier* and show in a diagram where the errors are being made.

v) What is optimal about the *Bayes' Classifier*?

b) **(.3)** Now consider d dimensional feature vectors. Let $p(\mathbf{x}|\omega_i) = N(\boldsymbol{\mu}_i, \Sigma)$ for $i = 1, 2$ in a two-class d-dimensional problem. The covariances are the same but the means and prior probabilities are arbitrary. Show that the decision boundary between the two classes is a hyper-plane.

Question 3: *Linear discriminants*

A discriminant is a function that takes an input vector $\mathbf{x}$ and assigns it to one of K classes. In this question we will consider the case when $K = 2$ and we have a linear discriminant function that is

$$\text{Class}\,(\mathbf{x}) = \begin{cases} \omega_1 & \mathbf{w}^T\mathbf{x} + w_0 \geq 0 \\ \omega_2 & \text{otherwise} \end{cases}$$

Please answer the following questions

a) **(.4)** Given training data $\{(\mathbf{x}_i, t_i)\}_{i=1}^N$ where each $\mathbf{x}_i \in \mathbb{R}^d$ and each t_i is a scalar indicating the label of the training example, there are lots of different

ways to learn $\mathbf{w}^T$ and w_0. One method is to choose the $\tilde{\mathbf{w}} = (\mathbf{w}, w_0)$ that maximises

$$J_{\text{MSE}}(\tilde{\mathbf{w}}) = \frac{1}{2}\sum_{i=1}^{N}(\tilde{\mathbf{w}}^T\,\tilde{\mathbf{x}}_i - t_i)^2$$

where $\tilde{\mathbf{x}} = (\mathbf{x}, 1)$. Calculate the derivative of $J_{\text{MSE}}(\tilde{\mathbf{w}})$ with respect to $\tilde{\mathbf{w}}$ and set to zero. Convert the expression you have just computed to matrix notation where

$$X = \begin{pmatrix} \uparrow & \uparrow & & \uparrow \\ \tilde{\mathbf{x}}_1 & \tilde{\mathbf{x}}_2 & \cdots & \tilde{\mathbf{x}}_N \\ \downarrow & \downarrow & & \downarrow \end{pmatrix} \quad \text{and} \quad \mathbf{t} = \begin{pmatrix} t_1 \\ t_2 \\ \vdots \\ t_N \end{pmatrix}$$

and then show that the optimal $\tilde{\mathbf{w}}$ is given by

$$\tilde{\mathbf{w}} = (XX^T)^{-1}X\mathbf{t}$$

b) **(.4)** While another approach, perceptron learning, is to choose the $\tilde{\mathbf{w}} = (\mathbf{w}, w_0)$ that maximises

$$J_{\text{p}}(\mathbf{w}) = -\sum_{i\in\mathcal{M}} \tilde{\mathbf{w}}^T\,\tilde{\mathbf{x}}_i\,t_i$$

where each $t_i \in \{-1, +1\}$. Explain what the set $\mathcal{M}$ denotes? Show how the gradient descent algorithm can be then used to derive the perceptron learning rule.

c) **(.2)** What are the advantages and disadvantages of these methods for finding $\tilde{\mathbf{w}}$?

Question 4: *SVM*

Assume you have been given training data $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$ where each $\mathbf{x}_i \in \mathbb{R}^d$ is a feature vector and $y_i \in \{-1, 1\}$ is its corresponding label.

a) **(.1)** Assume these training examples are linearly separable. Explain what this means?

b) **(.3)** What is the criterion used by the SVM for choosing the optimal separating hyperplane. Please draw a diagram, with a 2 dimensional example to accompany your answer. Show in this diagram the optimal separating hyperplane and the *support vectors*.

c) **(.2)** Derive the optimization problem the SVM solves, refer back to your diagram to aid your explanation.

d) **(.2)** Write down the Lagrangian associated with the optimization problem you have just derived. Also write down the expression for the optimal separating hyperplane in terms of the Lagrange multipliers and the training features and their labels.

e) **(.1)** Given a novel feature vector $\mathbf{v}$, which computations does the SVM perform to estimate $\mathbf{v}$'s class.

f) **(.1)** You can transform your data to a higher dimensional space with a function $\psi : \mathbb{R}^d \to \mathbb{R}^n$ where $n > d$ and find the optimal separating hyperplane in this new space. Why would you do this ? For your novel feature vector $\mathbf{v}$, do you need to explicitly transform it to the higher dimensional space to estimate its class with this new SVM? Explain.

Question 5: *Boosting*

a) **(.1)** What does the *AdaBoost* algorithm produce from a set of weak classifiers and labelled training data?

b) **(.4)** Describe the steps of the *AdaBoost* algorithm.

c) **(.1)** In your opinion what is the critical step in the *AdaBoost* algorithm?

d) **(.2)** What quality must the weak classifiers possess in order for *AdaBoost* to run successfully?

e) **(.2)** What are the strengths/weakness of using boosting to solve classification problems?

Question 6: *Unsupervised Learning*

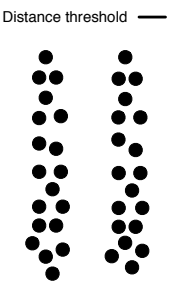
a) **(.2)** Describe an image recognition task where you would have to perform unsupervised learning.

b) **(.3)** Describe the steps of the k-means clustering algorithm.

c) **(.3)** Describe the steps of agglomerative clustering.

Then apply agglomerative clustering with $d_{\min}$ used as the measure of distance between clusters to the data-points in the figure to the right. Terminate the algorithm when the distance between the nearest clusters exceeds the threshold shown. Sketch the clusters at the time of termination. Note

$$d_{\min}(\mathcal{D}_i, \mathcal{D}_j) = \min_{\substack{\mathbf{x} \in \mathcal{D}_i \\ \mathbf{x}' \in \mathcal{D}_j}} \|\mathbf{x} - \mathbf{x}'\|$$



d) **(.2)** In k-means clustering you find a local minimum of the cost function

$$J_e = \sum_{i=1}^{k} \sum_{\mathbf{x} \in \mathcal{D}_i} \|\mathbf{x} - \mathbf{m}_i\|^2 \quad \text{where} \quad \mathbf{m}_i = \frac{1}{|\mathcal{D}_i|} \sum_{\mathbf{x} \in \mathcal{D}_i} \mathbf{x}$$

For the points examined in part c), if $k = 2$, sketch the clusters which would minimize J_e. Which weakness in k-means clustering does this highlight?

Question 7: *PCA/LDA*

a) **(.1)** PCA is a technique for performing dimensionality reduction. What is the criterion used by PCA to derive its basis.

b) **(.1)** Linear discriminant analysis (LDA) also represents a signal in a lower dimensional space. However, its criterion for choosing the lower dimensional space differs. What is this criterion ?

c) **(.3)** For the two class problem LDA projects the high dimensional data onto a line. This optimal line is found by maximizing the *Fisher* criterion. This criterion depends on two distinct measurements made from each class. What are they ?

d) **(.5)** Assume we have a two class problem. The feature vectors extracted from each class are two dimensional and the class conditional densities are:

$$p(\mathbf{x}|\omega_1) \sim N(\boldsymbol{\mu}_1, \Sigma) \quad \text{and} \quad p(\mathbf{x}|\omega_2) \sim N(\boldsymbol{\mu}_2, \Sigma)$$

where

$$\boldsymbol{\mu}_1 = (1,1)^T, \quad \boldsymbol{\mu}_2 = (3,2)^T \quad \text{and} \quad \Sigma = \begin{pmatrix} \sigma^2 & 0 \\ 0 & 0.2^2 \end{pmatrix}$$

In LDA you project each feature vector generated by these class conditionals onto a line via $\mathbf{w}^T\mathbf{x}$ to obtain a scalar value.

For $\sigma = 1$ sketch the class conditional densities and indicate why in this case it is better to project feature vectors from these two classes onto the y-axis as opposed to the x-axis for performing discrimination on the resulting scalar values.

For what values of σ will it be better to project onto the x-axis ?

For what values of σ will the y-axis tend towards the optimal projection line with respect to the Fisher criterion?

Question 8: *Non-parametric classification*

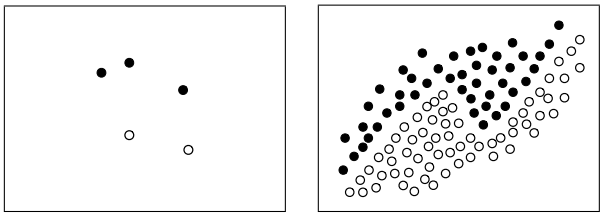
a) **(.2)** Describe the steps of the k-nearest neighbour classifier?

b) **(.2)** What are the advantages and disadvantages of a k-nearest neighbour classifier?

c) **(.2)** What are the trade-offs between choosing a large or small value of k? Draw a figure to illustrate your argument.

d) **(.2)** In two dimension draw the points at a distance 1.0 from the origin using the L_1, L_2 and L_∞ norms.

e) **(.2)** Draw the decision boundary formed by a 1-nearest neighbour classifier in the two different figures below. Not that the sparse set of points is a subset of the dense ones. What lesson should be learned from these examples?



(**a**) (**b**)

Part II

Question 9: *VC-dimension*

Remember that in order to prove that a class of functions $\mathcal{H}$ has VC-dimension d you need to show that

- There exists a set of d points which can be **shattered** by $\mathcal{H}$.
- There exists **no** set of $d+1$ points that can be shattered by $\mathcal{H}$

a) When does a class of functions $\mathcal{H}$ *shatter* a set of points $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$?

b) Show with appropriate diagrams that there exist 3 points in $\mathbb{R}^2$ that can be shattered by a line.

c) What is the VC-dimension of intervals in $\mathbb{R}$? In this case $\mathcal{H}$ is defined such that each $h \in \mathcal{H}$ is associated with an interval $[a, b]$ and $x \in \mathbb{R}$ has $h(x) = 1$ if and only if $x \in [a, b]$.

d) What is the VC-dimension of the union of k intervals on the real line? In other words each $h \in \mathcal{H}$ is associated with k closed intervals $[a_i, b_i]$, $i = 1, 2, \ldots, k$ and $h(x) = 1$ if and only if $x \in \cup_{i=1}^{n}[a_i, b_i]$.

e) What is the VC-dimension of axis parallel rectangles in $\mathbb{R}^2$? In other words $h \in \mathcal{H}$ is associated with 2 closed intervals $[a_i, b_i]$ for $i = 1, 2$ and then for any $\mathbf{x} = (x_1, x_2) \in \mathbb{R}^2$, $h(\mathbf{x}) = 1$ if and only if $x_i \in [a_i, b_i]$ for $i = 1, 2$.

f) Show that the VC-dimension of the class $\mathcal{H}$ of hyperplanes in $\mathbb{R}^2$ is 3?

g) Show that the VC-dimension of the class $\mathcal{H}$ of hyperplanes in $\mathbb{R}^d$ is $\geq d+1$?

Question 10: *Skin Detection*

You have been hired by a small start-up company that is developing an internet search tool for kids based on analysing text and images. Unfortunately, much of the content on the web contains adult material which is unsuitbale for children. Thus the company would like to remove this content automatically from their searches. They would like you to build a robust skin classifier which will be one component of their removal process. Your first assignment is to write a report suggesting

- plausible approaches to classifying skin pixels and representing skin colour that you will investigate commenting on their strengths and weakness **and**
- experiments that should be performed to compare and assess the performance for each classifier you investigate and how this comparison should be performed

To answer this question please write down what you would put in this report.

(You will score points for more detailed but succinct answers and also for motivating your suggested approaches with respect to efficiency and effectiveness. I

don't expect you to know the best approach, just ones that would seem plausible considering what you have learnt during the course.)

Question 11: *Face Detection & Recognition*

You have a friend who is living in shared accommodation and has the following problem. The yoghurt he stores in the shared fridge keeps on disappearing. All of his flatmates deny being the thief. But he has his suspicions. Then he hears that you have just completed a course in *Image recognition and classification* and has an idea. He wants to install a secret web-cam within the kitchen focused on the fridge and then have you help him write software to automatically analyse the images it grabs. He wants the software to be able to:

- detect when there is a face in the image,
- recognize the face given reference face images of the flatmates deviously obtained from *Facebook* and also
- spot if the very distinctive red yoghurt carton of his favourite brand is also present in the image

Describe the algorithms and methods you would propose to implement to achieve each of these goals. And also in which temporal order would you implement them for maximum efficiency. Explain what training you have to perform.

(You will score points for more detailed description and for motivating your suggested solution with respect to efficiency and effectiveness.)

Question 12: *Clustering*

a) Explain how using EM to find the parameters of a Gaussian mixture model can be viewed as a form of clustering.

b) Explain intuitively how EM proceeds for learning the parameters of a Gaussian mixture model. Explain how this procedure is more sophisticated than k-means clustering.

c) Explain how clustering can be used to perform image segmentation. Image segmentation refers to assigning pixels with similar properties to the same cluster. Which feature vectors could be extracted from the image to perform this segmentation?

Question 13: *Kernel magic*

Assume you are given m one dimensional training examples and their associated labels, that is $\{(x_i, y_i)\}_{i=1}^m$ where each $x_i \in \mathbb{R}^1$ and $y_i \in \{-1, +1\}$.

a) Draw a case where you have $m = 3$ training examples which are not linearly separable.

b) You know if you transform your one-dimensional data to a higher dimensional space then there is a higher likelihood that they will be linearly separable. Thus you define a feature transformation $\phi_n : \mathbb{R}^1 \to \mathbb{R}^n$ where

$$\phi_n(x) = \left\{ e^{-\frac{x^2}{2}}, x\, e^{-\frac{x^2}{2}}, \frac{x^2}{\sqrt{2}}\, e^{-\frac{x^2}{2}}, \ldots, \frac{x^n}{\sqrt{n!}}\, e^{-\frac{x^2}{2}} \right\}$$

Explain why any set of 3 points (with no duplicates) can be linearly separated when transformed via ϕ_2. Similarly explain why any set of $n+1$ points (with no duplicates) can be linearly separated when transformed by ϕ_n.

c) Consider the case when $n \to \infty$ and ϕ_n becomes

$$\phi_\infty(x) = \left\{ e^{-\frac{x^2}{2}}, x\, e^{-\frac{x^2}{2}}, \frac{x^2}{\sqrt{2}}\, e^{-\frac{x^2}{2}}, \ldots, \frac{x^j}{\sqrt{j!}}\, e^{-\frac{x^2}{2}}, \ldots \right\}$$

Can you explicitly construct $\phi_\infty(x)$? (Not a trick question)

d) Is there a finite set of points, containing no duplicates, that cannot be linearly separated after applying ϕ_∞?

e) A linear classifier can be expressed using only the inner products of support vectors in the transformed feature space. The Kernel trick, exploited by the SVM, is to define a function $K(\cdot, \cdot)$ such that

$$K(x, y) = \phi_\infty(x) \cdot \phi_\infty(y)$$

where the inner product between two infinte vectors $\mathbf{a} = (a_1, a_2, \ldots)$ and $\mathbf{c} = (c_1, c_2, \ldots)$ is defined as

$$\mathbf{a} \cdot \mathbf{c} = \sum_{i=1}^{\infty} a_i\, b_i$$

Given the definition of ϕ_∞ compute the form of $K(x, y)$. Hint you may want to use the Taylor series expansion of e^x:

$$e^x = \lim_{n \to \infty} \sum_{j=0}^{n} \frac{x^j}{j!}$$

f) With such a high dimensional feature space should we be concerned about overfitting?

List of Formulae

- If a one dimensional variable x follows a Gaussian distribution this is denoted by $\mathcal{N}(\mu, \sigma)$ and

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-.5\,\frac{(x-\mu)^2}{\sigma^2}\right)$$

- If $\mathbf{x}$ is a vector of dimension d and follows a multivariate normal/Gaussian distribution denoted by $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$, then

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-.5\,(\mathbf{x}-\boldsymbol{\mu})^T\,\Sigma^{-1}\,(\mathbf{x}-\boldsymbol{\mu})\right)$$

  where $|\Sigma|$ is the determinant of the matrix Σ.

- The L_2 norm (Euclidean distance)

$$L_2(\mathbf{x}, \mathbf{y}) = \left(\sum_{i=1}^{d} |x_i - y_i|^2\right)^{\frac{1}{2}}$$

- The L_1 norm (Manhattan distance)

$$L_1(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{d} |x_i - y_i|$$

- The L_∞ norm

$$L_\infty(\mathbf{x}, \mathbf{y}) = \max_i |x_i - y_i|$$