Course DD2427 - Final Exam

You may use a calculator but its use is not really necessary. A list of useful formulae is on the last page.

Please answer 10 questions from **Part I** of the exam. Each question carries equal weight. You must get a score of at least 70% to pass this part of the exam and get an E grade and also be eligible for a higher grade.

For a higher grade you must answer questions from **Part II** and/or **Part III**. Your top 4 scoring from these parts will be used to compute your grade. Each question in **Part II(III)** if answered completely correctly gives 17(30) marks.

The table below shows the points you have to score to obtain each grade.

Grade	**Score**
D	≥ 20
C	≥ 40
B	≥ 60
A	≥ 80

Up to 15 bonus points are available by **participating** in the poster session.

The bold face numbers in brackets in **Part II** and **Part III** indicate the percentage of the total score associated with each part of a question.

Part I

Question 1:

Let X and Y be discrete random variables which follow a joint probability distribution $P(X, Y)$

a) Define the conditional probability, $P(X \,|\, Y)$, in terms of $P(Y)$ and $P(X, Y)$.

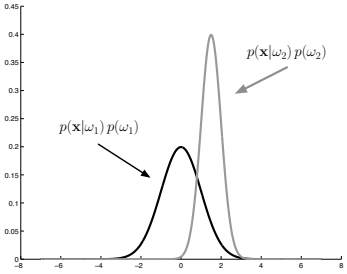
b) Express $P(X)$ in terms of $P(X \,|\, Y)$ and $P(Y)$.

c) Using these results, state and derive *Bayes' Rule*.

Question 2:

a) For the binary classification problem, how is the *Bayes' Classifier* defined given each class' prior probability and class conditional distribution?

b) Draw the decision boundaries defined by a Bayes' classifier in this figure:



c) Define the *Probability of Error* in words and write down its mathematical expression in terms of the $p(x\,|\,\omega_i)$'s and $p(\omega_i)$'s for the above figure.

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

Question 3:

Let $\mathbf{x}$ follow a bi-variate Gaussian distribution, that is $\mathbf{x} \sim \mathcal{N}\left((0,0)^T, \Sigma\right)$.

a) Sketch this probability distribution when $\Sigma = \begin{pmatrix} 10^2 & 0 \\ 0 & 3^2 \end{pmatrix}$.

b) Which of these two figures shows the distribution when $\Sigma = \begin{pmatrix} 10^2 & -24 \\ -24 & 3^2 \end{pmatrix}$?

(a) **(b)**

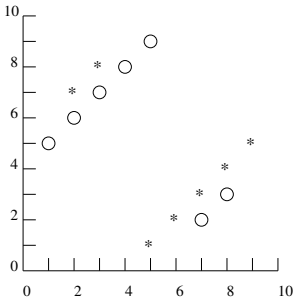
c) In the above figures draw the direction of the eigenvectors of Σ. Which quantities describe the spread in each of these directions?

Question 4:

a) State the steps of a k-nearest neighbour classifier.

b) What pre-processing should be applied to the feature data before applying the nearest neighbour classifier.

c) Draw the decision boundary defined by the 1-nearest neighbour classifier in the following figure.



Question 5:

You are given training data $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)$ where each feature vector $\mathbf{x}_i \in \mathbb{R}^2$ and its associated label $y_i \in \{-1, 1\}$. You want to build a *Bayes' Classifier* discriminating between these two classes and approximate each class conditional distribution $p(\mathbf{x} \,|\, \omega_i)$ with a two dimensional histogram.

a) What is the danger of achieving a training error of zero without any consideration of the complexity of your classifier?

b) Describe how cross-validation can be used to decide on the histogram's bin width, from a set of possible widths, given the classification task at hand.

Question 6:

A *Bayesian Classifier* requires an estimate of each class conditional probability $p(\mathbf{x} \,|\, \omega_i)$. During the course we have explored various ways to do this estimation.

a) What is the difference between a *parametric* and a *non-parametric* estimate?

b) Give one example of each approach.

c) State one advantage and one disadvantage of each approach.

d) *Maximum Likelihood Estimation* is a standard approach for what? Given training examples $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ write down the quantity MLE maximizes and how it performs this.

Question 7:

a) Write down the mathematical definition of a linear classifier.

b) *Perceptron Learning* is a method to learn such a classifier. Given training data $\{(\mathbf{x}_i, t_i)\}_{i=1}^n$ where each $\mathbf{x}_i \in \mathbb{R}^d$ and each $t_i \in \{-1, +1\}$ is a scalar indicating the label of the training example it chooses the $\mathbf{w} = (\mathbf{w}_1, w_0)$ that minimizes

$$J_{\mathrm{p}}(\mathbf{w}) = -\sum_{i \in \mathcal{M}} (\mathbf{w}_1^T \, \mathbf{x}_i + w_0) \, t_i$$

Explain what the set $\mathcal{M}$ denotes? Use the gradient descent algorithm to derive the perceptron learning rule.

c) Name one advantage and one disadvantage of this method for finding $\mathbf{w}$?

Question 8:

a) What is the *curse of dimensionality*? Name one method affected by this.

b) PCA is a technique for performing dimensionality reduction. What is the criterion used by PCA to derive its basis?

c) 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 ?

d) Boosting can be seen as a way to perform dimensionality reduction. Explain this statement.

Question 9:

You want to discriminate between images taken in indoor environments and those taken in outdoors. Describe 3 different types of image features you could extract from these images which could be used by a suitable classifier.

Question 10:

a) State the steps in words of the boosting algorithm, you can omit the exact mathematical details.

b) What do you think is the most important step of the algorithm?

Question 11:

You have training data $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ from two classes which is linearly separable

a) What does linearly separable mean?

b) 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) Why is maximizing this criterion a good idea?

d) Write down the mathematical expression for the quantity maximized by the SVM.

Question 12:

Continuing the previous question

a) The hyperplane found by the SVM has the form

$$\mathbf{w}^* = \sum_{i=1}^{n} \alpha_i \, \mathbf{x}_i \, y_i$$

What values have the α_i's for the support vectors and non-support vectors?

b) Given a novel feature vector $\mathbf{x}$, what computations does the SVM perform to estimate $\mathbf{x}$'s class?

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

Question 13:

a) State the difference between *supervised* and *unsupervised* learning.

b) Describe an image recognition task where you would have to perform unsupervised learning.

c) State the steps of the k-means clustering algorithm.

Question 14:

a) Sketch this one dimensional probability distribution

$$p(x) = \pi_1 \mathcal{N}(0, .5) + (1 - \pi_1)\mathcal{N}(5, 1)$$

when $\pi_1 = \frac{1}{2}$; $\pi_1 = .1$; **and** $\pi_1 = .9$.

b) If you have n points generated from $p(x)$ when $\pi_1 = \frac{1}{2}$ and you fit a Gaussian distribution to this data. Sketch what this distribution will look like. What's the problem here?

c) This issue highlights a problem with parametric methods. What is it?

d) What method is used to find the parameters of a Gaussian mixture model from training examples generated from the distribution?

Question 15:

a) What are the *true positive* and *false positive* rates of a classifier?

b) If you use a sliding window approach to finding faces in an image of size 300×300 and your classifier has a false positive rate of .05 roughly how many faces will you find even if the image does not contain a face?

c) What is the name of the curve which plots the *false positive rate* Vs *true positive rate* as the parameter controlling classification is varied?

d) Draw this curve for a classifier which has

- random performance
- ideal performance
- performance similar to that of your final classifier in the face detector project.

Part II

Question 1:

a) **(.4)** Describe in mathematical terms and with the help of diagrams the initial constrained optimization problem which defines the SVM for non-separable data.

b) **(.05)** What's great about this constrained optimization problem?

c) **(.2)** Qualitatively, how does varying the value of the penalty term C in the objective function affect the optimal hyper-plane found.

d) **(.15)** What is the Lagrangian of this constrained optimization problem?

e) **(.2)** Use this Lagrangian to show that the optimal hyperplane has the form

$$\mathbf{w}^* = \sum_{i=1}^{n} \alpha_i \, y_i \, \mathbf{x}_i$$

Question 2:

This question exams the Adaboost algorithm. We have n training examples $(\mathbf{x}_1, y_1), \ldots (\mathbf{x}_n, y_n)$ where each $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \{-1, 1\}$. Let $h_t(\cdot)$ be the weak classifier obtained at step t. The final classifier has form

$$H(\mathbf{x}) = \operatorname{sgn}\left(f(\mathbf{x}) + \Theta\right) \quad \text{where} \quad f(\mathbf{x}) = \sum_t \alpha_t h_t(\mathbf{x})$$

where Θ is a threshold. Let $D_t(i)$ represent the weight associated with the i-th example at the t-th iteration. The training error on the weighted dataset is

$$\epsilon_t = \sum_{i=1}^{n} D_t(i) \operatorname{Ind}(H(\mathbf{x}_i) \neq y_i),$$

where $\operatorname{Ind}(a = b)$ is the indicator function and in the strong classifier

$$\alpha_t = \frac{1}{2} \log\left(\frac{1 - \epsilon_t}{\epsilon_t}\right)$$

Answer the following

a) **(.15)** Is it true that in each round of boosting there always exists a weak classifier $h_t(\cdot)$ such that its training error on the weighted dataset is $\epsilon_t \leq .5$? Explain your answer.

b) **(.15)** What happens in the cases when $\epsilon_t = .5$; and $\epsilon_t = 1$? In each case what is the training error of the final strong classifier.

c) **(.7)** Consider the following toy problem. You will apply Adaboost to this dataset:

Class ω_1 points: $(0, -1), (0, 1)$
Class ω_2 points: $(1, 0), (-1, 0)$

with the set of vertical and horizontal lines as weak classifiers.

- Plot the points. Are they linearly separable?
- Sketch how 3 rounds Adaboost would qualitatively run on this dataset. For each timestep draw the weak classifier chosen and how the weights associated with each example changes. (Make the size of the data-point proportional its weight.) What is the training error of the final strong classifier?

Question 3:

Describe how you would build a person recognition system using a PCA representation of face images. In your description include, with appropriate mathematical detail,

- **(.4)** How to learn the PCA basis.
- **(.3)** How to represent an image with this basis and how to decide how many basis vectors one should include.
- **(.3)** How recognition could be performed using this learnt basis.

Question 4:

Imagine you have a face detector such that

$$P(\hat{f} = 1 \,|\, f = 1) = 1 - \epsilon \quad \textbf{and} \quad P(\hat{f} = 0 \,|\, f = 0) = 1 - \alpha$$

where $f \in \{0, 1\}$ indicates the ground truth of whether a patch is present or not while $\hat{f} \in \{0, 1\}$ is the prediction of the face detector.

Imagine you have K *independent* face detectors each having the same true and false positive rate of the detector just described. If these detectors are applied to an image patch we get

$$\hat{\mathbf{f}} = (\hat{f}_1, \hat{f}_2, \ldots, \hat{f}_K)$$

where $\hat{f}_i \in \{0, 1\}$ is the prediction of the ith detector. $k_{\hat{\mathbf{f}}} = \sum_i \hat{f}_i$ is equal to the number of detectors which predict a face while $K - k_{\hat{\mathbf{f}}}$ is the number which predict a non-face. Let γ be the prior probability that the patch contains a face. Given this information answer the following:

a) **(.2)** Write down the expression, remembering to exploit the independence, for

$$p(\hat{\mathbf{f}} \,|\, f = 1)$$

b) **(.2)** What is the posterior probability the patch contains a face given $\hat{\mathbf{f}}$?

c) **(.3)** Let $\gamma = \epsilon = \alpha = .01$ what is the constraint $k_{\hat{\mathbf{f}}}$ must fulfill such that

$$p(f = 1 \,|\, \hat{\mathbf{f}}) \geq .99$$

For $K = 4$, what is the minimal value of $k_{\hat{\mathbf{f}}}$ such that the above performance level is met? For $K = 10$? And as $K \to \infty$ what ratio of detections should correctly predict a face to ensure this level of performance.

d) **(.3)** Define a final classifier such that

$$F(\hat{\mathbf{f}}) = \begin{cases} 1 & \text{if } \sum_i \hat{f}_i \geq K_0 \\ 0 & \text{otherwise} \end{cases}$$

Continuing with the parameter settings just given, write down an expression for the $P(\text{error})$ of this classifier.

Question 5:

a) **(.5)** Show that $d+1$ points in R^d can be shattered by a hyper-plane.

b) **(.2)** Explain why the VC dimension of a 1 nearest neighbour classifier is infinity.

c) **(.3)** The type of weak binary classifier for data $\mathbf{x} \in \mathbb{R}^d$ you have been using in the face lab has a name. It is called a *decision stump*. As you know the classification rule has parameters $q \in \{-1, 1\}, j \in \{1, 2, \ldots, d\}$ and θ and takes the form:

$$h(\mathbf{x}; j, q, \theta) = \begin{cases} 1 & \text{if } q\, x_j \geq q\, \theta \\ 0 & \text{otherwise} \end{cases}$$

where $\mathbf{x} = (x_1, x_2, \ldots, x_d)$. For two dimensional $\mathbf{x}$ consider convex combinations of two vertical decision stumps

$$f(\mathbf{x}) = \alpha_1\, h(\mathbf{x}; 1, q_1, \theta_1) + (1 - \alpha_1)\, h(\mathbf{x}; 1, q_2, \theta_2)$$

with $0 \leq \alpha_1 \leq 1$. Classify a point according to some threshold T such that

$$\text{Class}\,(\mathbf{x}) = \begin{cases} \omega_1 & \text{if } f(\mathbf{x}) \geq T \\ \omega_2 & \text{if } f(\mathbf{x}) < T \end{cases}$$

Sketch the three qualitative ways $\mathbb{R}^2$ can be partitioned into regions classified as class ω_1 and class ω_2. What is largest number of points which can be shattered by a classifier of this type?

Part III

Question 1:

a) **(.1)** State one reason why you would use a discriminative model as opposed to a generative one and vice versa.

b) The class conditional distributions for a binary classification problem are

$$p(\mathbf{x} \,|\, \omega = i) = \mathcal{N}(\boldsymbol{\mu}_i, \Sigma_i)$$

- **(.25)** Given these condition distributions and assuming equal priors for the two classes, what is the form of the decision boundary for the Bayesian classifier?

- **(.1)** How many parameters have to be estimated if
  - $\Sigma_i = \sigma_i^2 I$ for $i = 1, 2$,
  - Σ_i is a full covariance for $i = 1, 2$.

c) **(.1)** *Logistic Regression* assumes for the binary classification problem the decision boundary is a hyperplane, that is

$$\log \frac{P(\omega = 1 \,|\, \mathbf{x})}{P(\omega = 0 \,|\, \mathbf{x})} = w_0 + \mathbf{w}_1^T \mathbf{x}$$

From this derive the $g(\cdot)$ such that

$$P(\omega = 1 \,|\, \mathbf{x}) = g(w_0 + \mathbf{w}_1^T \mathbf{x})$$

d) If the feature vector $\mathbf{x}$ is projected to a higher dimensional space via a function $\boldsymbol{\phi} : \mathbb{R}^d \to \mathbb{R}^m$ before finding the best hyperplane, it is possible to model decision boundaries more compliated than a hyperplane in the original space.

- **(.3)** How should $\boldsymbol{\phi}(\cdot)$ be defined if we want to replicate the decision boundary in part b) when
  - $\Sigma_i = \sigma_i^2 I$ for $i = 1, 2$,
  - Σ_i is a full covariance for $i = 1, 2$.
- **(.15)** In each case, how many parameters does logistic regression have to estimate? Comment on this with respect to part b) of the question.

Question 2:

a) **(.1)** For a binary classification problem LDA finds an *optimal* projection $\mathbf{w}$. In words which criterion does it use to find this optimal projection.

b) **(.2)** If we have $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ and then $\mathbf{x}$ is projected into 1 dimension via:

$$y = \mathbf{w}^T \mathbf{x}$$

Then y follows a 1 dimensional distribution, $y \sim \mathcal{N}(\mu, \sigma^2)$. What are the expressions for μ and σ^2 in terms of $\mathbf{w}, \boldsymbol{\mu}$ and Σ?

c) Given two classes where each is Normally distributed, $\mathcal{N}(\boldsymbol{\mu}_i, \Sigma_i)$ for $i = 1, 2$, then the optimal projection found by LDA is defined as

$$\mathbf{w}^* \propto (\Sigma_1 + \Sigma_2)^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$$

(.6) Consider the following example:

$$p(\mathbf{x} \,|\, \omega_1) = \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}10^2 & 0\\0 & 2^2\end{pmatrix}\right) \quad \text{and} \quad p(\mathbf{x} \,|\, \omega_2) = \mathcal{N}\left(\boldsymbol{\mu}, \begin{pmatrix}2^2 & 0\\0 & 10^2\end{pmatrix}\right)$$

For values of $\boldsymbol{\mu} = (0, 0)^T; (\frac{1}{10}, 0)^T; (0, \frac{1}{10})^T$; and $(\frac{1}{10}, \frac{1}{10})^T$ answer the following questions:

- What is the optimal projection according to LDA?

- Sketch the distribution of class 1 and class 2 points when they have been projected into 1d via the optimal projection.
- Sketch a classifier which discriminates between the projected points from the two classes. How is this classifier's performance relative to a Bayes classifier calculated on the original 2 dimensional data?

(.1) Which limitation of LDA does this example highlight?

Question 3:

a) **(.3)** Explain how using Expectation Maximization (EM) to find the parameters of a Gaussian mixture model can be viewed as a form of clustering.

b) **(.4)** 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) **(.3)** 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?

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

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