Course DD2427 2011 - Final Exam

You may use a calculator but you shouldn't need one.

In **Part I** of the exam your top 7 scoring answers will be used to compute your score S_{I}. Here each question is worth 10 points. To pass the exam you must get $S_{\mathrm{I}} \geq 45$. From **Part II** of the exam I will use you top 4 scoring answers to compute S_{II} (assuming you have passed the exam). Here each question is worth 20 points. Your final score will then be calculated as

$$S_F = (S_{\mathrm{I}} - 50) + S_{\mathrm{II}} + S_{\mathrm{P}}$$

where S_{P} are your bonus points from the *Poster Session*. The thresholds on S_F for achieving the higher grades:

Grade			
D	C	B	A
≥ 8	≥ 18	≥ 40	≥ 65

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

Part I

Question 1: *Bayes' rule*

Let X and Y be discrete binary random variables with joint pdf $P(X, Y)$.

a) Define $P(X = x \,|\, Y = y)$, in terms of $P(Y = y)$ and $P(X = x, Y = y)$.

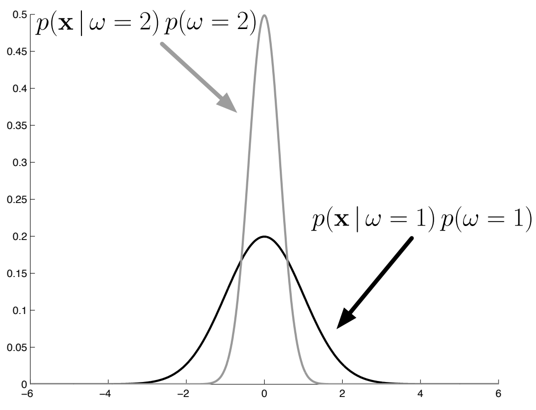
b) Express $P(X = x)$ in terms of $P(X = x \,|\, Y)$ and $P(Y)$.

c) Use these results to state and derive *Bayes' rule*.

Question 2: *Bayes' classifier*

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) Write down the mathematical expression for the *Probability of Error* 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: *Covariance matrices*

a) Let $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ follow a bivariate Gaussian distribution. Match the covariance matrices to the appropriate figure showing the iso-probability contours of the pdf.

$$\Sigma_1 = \begin{pmatrix} 10^2 & -12 \\ -12 & 3^2 \end{pmatrix}, \; \Sigma_2 = \begin{pmatrix} 10^2 & 0 \\ 0 & 3^2 \end{pmatrix}, \; \Sigma_3 = \begin{pmatrix} 10^2 & -24 \\ -24 & 3^2 \end{pmatrix}, \; \Sigma_4 = \begin{pmatrix} 10^2 & 18 \\ 18 & 3^2 \end{pmatrix}$$

(i) **(ii)** **(iii)** **(iv)**

b) Let $\mathbf{x}$ be a multi-variate Gaussian random variable with $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_x, \Sigma_x)$. Any linear transformation of $\mathbf{x}$, $\mathbf{y} = A\mathbf{x}$, also follows a Gaussian distribution that is $\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}_y, \Sigma_y)$. It is the case that $\boldsymbol{\mu}_y = A\boldsymbol{\mu}_x$. Show that $\Sigma_y = A\,\Sigma_x\,A^t$.

c) Use the previous result to compute the distribution of $z = x + y$ if $\mathbf{x} = (x, y)^t$ and

$$p(\mathbf{x}) = \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} 1, & .5 \\ .5 & 1 \end{pmatrix}\right)$$

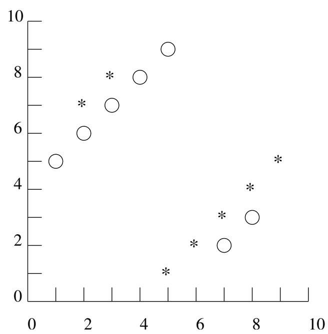
Question 4: *Nearest neighbour classifier*

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

b) State two advantages and disadvantages of a k-nearest neighbour classifier?

c) What are the two trade-offs between choosing a large or small value of k?

d) Draw the 1-nearest neighbour decision boundary in the following figure:



Question 5: *Cross Validation*

You are given training data $\{(\mathbf{x}_i, t_i)\}_{i=1}^n$ where each $\mathbf{x}_i \in \mathcal{R}^2$ is a feature vector and each $t_i \in \{-1, +1\}$ is its corresponding label. You decide to construct a Bayes' classifier using this data to help you classify a new feature vector $\mathbf{x}^*$. Answer the following:

a) You use a histogram to estimate each class conditional probability distributions. What is a histogram? Describe how you construct one and would use it to estimate the desired probability?

b) The histogram has one parameter, bin width b_w which has to be set. What happens if too large a value for b_w is chosen and if too small a value is chosen?

c) *k-fold cross-validation* can be used to estimate b_w. State the steps of this procedure and the criterion used to select a value for b_w.

Question 6: *Linear discriminants*

A linear classifier classifies a points $\mathbf{x}$ with $\operatorname{sgn}(\mathbf{w}_1^t\mathbf{x} + w_0)$. Assume you have training data $\{(\mathbf{x}_i, t_i)\}_{i=1}^n$ where each $\mathbf{x}_i \in \mathbb{R}^d$ and label $t_i \in \{-1, +1\}$.

a) The parameters $\mathbf{w} = (\mathbf{w}_1, w_0)$ can be found by minimizing a cost function:

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \sum_{i=1}^{n} L(t_i, \mathbf{w}_1^t\mathbf{x}_i + w_0)$$

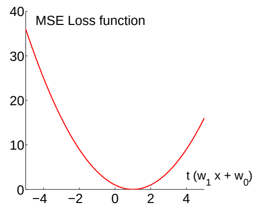
where $L(.,.)$ is a loss function penalizing differences between the prediction of the classifier for a training example and its true label. For *Perceptron Learning* write down this loss function?

b) Use the gradient descent algorithm to derive the perceptron learning rule.

c) Plot $t_i(\mathbf{w}^t\mathbf{x}_i + w_0)$ Vs $L(t_i, \mathbf{w}_1^t\mathbf{x}_i + w_0)$ for the loss function used in Perceptron Learning. How is this one better than the loss function used by the MSE criterion

$$L_{\text{MSE}}(t_i, \mathbf{w}_1^t\mathbf{x}_i - w_0) = (t_i - \mathbf{w}_1^t\mathbf{x}_i - w_0)^2$$

and shown here?



Question 7: *Dimensionality reduction*

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

b) PCA performs dimensionality reduction on d dimensional vectors by finding a set of $k \leq d$ basis vectors to represent the training vectors. Answer the following:

   i) How is this new basis found from examples $\mathbf{x}_1, \ldots, \mathbf{x}_n$?

   ii) How is the number of basis vectors k usually set?

   iii) Given a novel feature $\mathbf{x}^*$, how is it represented in this new basis?

c) What trick should I exploit when I compute the PCA basis if I have very high dimensional $\mathbf{x}$'s and a relatively small number of training examples? Why is it important to exploit this trick?

Question 8: *Integral image*

Let $I(x, y)$ denote the intensity of pixel (x, y) of an image of height H and width W.

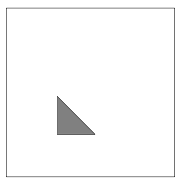
a) What is the integral image, $ii(x, y)$ and how is it calculated?

b) Look at this formula which is defined for each pixel (x, y)

$$ti(x, y) = \sum_{x'=1}^{x} \sum_{y'=y}^{y-x+x'} I(x', y')$$

The value $ti(x, y)$ corresponds to the sum of pixels in which type of region? Just consider the case where $y \geq x$. Draw a picture to help your explanation. (Note: $I(x, y)$ is assumed to be zero if $x \leq 0, y \leq 0, x > W$ or $y > H$.)

c) How can the images ii and ti be used to quickly compute the sum of the pixels in the right-angled equilateral triangle in this figure?



Question 9: *Boosting algorithm*

a) The boosting algorithm assumes one is given a set of weak classifiers and labelled training data. What does the boosting algorithm then output?

b) State the steps of the boosting algorithm - omit the exact mathematical details.

c) The boosting algorithm can be used to perform feature selection. Explain this statement and describe how it can be done.

d) State two weaknesses and two strengths of the boosting algorithm.

Question 10: *SVM I*

You have training data $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ from two classes which is linearly separable where each $\mathbf{x}_i \in \mathcal{R}^d$ and $y_i \in \{-1, 1\}$.

a) What does linearly separable mean?

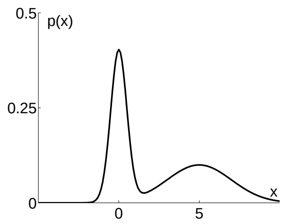
b) What is the criterion used by the SVM for choosing the optimal separating hyperplane. Why is maximizing this criterion a good idea?

c) Draw a two dimensional example to accompany your previous answer. Show in this diagram the optimal separating hyperplane and the *support vectors*.

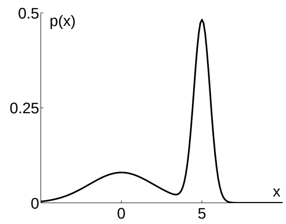
d) Write down the constrained optimization problem the SVM actually solves.

Question 11: *Gaussian mixture models*

a) Two GMM distributions are shown, match each one to one of the defined distributions



(a)



(b)

$$p_1(x) = .5\,\mathcal{N}(0, .5^2) + .5\,\mathcal{N}(5, 2^2)$$
$$p_2(x) = .5\,\mathcal{N}(0, 1) + .5\,\mathcal{N}(5, 1)$$
$$p_3(x) = .4\,\mathcal{N}(0, 2^2) + .6\,\mathcal{N}(5, .5^2)$$

b) You are given n independent samples, $x_1, \ldots, x_n$ from a distribution $p(x)$ and told that $p(x)$ is, in fact, either p_1, p_2 or p_3. Write down the score(s) you could use to indicate which of these distributions p is equal to.

c) Say instead you are told : $p(x) = \pi_1\,\mathcal{N}(\mu_1, \sigma_1^2) + (1 - \pi_1)\,\mathcal{N}(\mu_2, \sigma_2^2)$.
Describe how soft clustering of the samples $x_1, \ldots, x_n$ could be used to find the parameters $\pi, \mu_1, \mu_2, \sigma_1, \sigma_2$.

Question 12: *ROC curves*

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

b) The *accuracy (acc)* of a classifier is the proportion of examples it classifies correctly. How is it defined in terms of number of true positives etc.. ?

c) Show that - $acc = tpr \times r_p + (1 - fpr) \times r_n$ - where $r_p(r_n)$ is the proportion of the test examples which are really positive(negative).

d) The ROC curve plots the *fpr* Vs *tpr* of a classifier as its classification threshold is varied. Sketch the ROC curve of your classifier from the face lab.

e) On your ROC plot draw the line joining the points $(0, 1)$ to $(1, 0)$. Write down the equation of this line and show that the ROC curve intersects this line at $(1 - acc, acc)$.

Part II

Question 1: *SVM II*

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

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

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

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

e) Use the dual formulation of the optimization problem to show that the optimal hyperplane has the form

$$\mathbf{w}^* = \sum_{i=1}^{n} \alpha_i \, y_i \, \mathbf{x}_i \qquad (.\mathbf{2})$$

Question 2: *VC-dimension*

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

b) Why is the VC-dimension of the 1 nearest neighbour classifier infinity. **(.2)**

c) What is the VC-dimension of the union of k intervals on the real line? **(.3)**

Question 3: *Integral histograms*

The concept of integral image can be transferred to histograms of an image.

a) Let $H'(x, y)$ denote the histogram of the pixel intensities in the rectangular region defined by $[(1, 1), (x, y)]$ (coordinates of the top left and bottom right corners). Explain how $H'(x, y)$ can be computed efficiently and stored for all possible values of x and y in the image. **(.3)**

b) Let $H(A)$ denote the histogram of the pixel intensities in the image patch A. For image patches A and B if $A \cap B = \emptyset$, show that $H(A \cup B) = H(A) + H(B)$. **(.1)**

c) How can one compute the histogram of the pixel intensities in the rectangular region $[(x_1, y_1), (x_2, y_2)]$ using the previous result and the $H'(x, y)$'s? **(.3)**

d) One could equivalently implement integral histograms by calculating B integral images, where B is the number of bins in the histogram. Explain. **(.3)**

Question 4: *Discriminative Vs Generative modelling*

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

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

   i) Given these conditional distributions and equal priors, what is the form of the decision boundary for the Bayesian classifier? **(.25)**

   ii) If $\mathbf{x} \in \mathbb{R}^d$, how many parameters have to be estimated to build the Bayesian classifier when each covariance matrix is of the form

      1) $\Sigma_i = \sigma_i^2 I$ **(.05)** **2)** Σ_i is a full covariance **(.05)**

b) *Logistic Regression* assumes the decision boundary - the $\mathbf{x}$ such that $P(\omega = 0 \,|\, \mathbf{x}) = P(\omega = 1 \,|\, \mathbf{x})$ - for the binary classification problem is a hyperplane

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

This assumption implies: $P(\omega = 1 \,|\, \mathbf{x}) = g(w_0 + \mathbf{w}_1^t \mathbf{x})$. What is the expression for $g(\cdot)$? **(.15)**

c) Projecting $\mathbf{x}$ to a higher dimensional space via $\boldsymbol{\phi} : \mathbb{R}^d \to \mathbb{R}^m$ and then finding the best hyperplane allows more complicated decision boundaries in the original space to be modelled. How should $\boldsymbol{\phi}(\cdot)$ be defined if we want to replicate the decision boundary in part a) when the covariance matrices are

1) $\Sigma_i = \sigma_i^2 I$ **(.15)**   **2)** Σ_i is a full covariance **(.15)**

In each case, how many parameters does logistic regression have to estimate? Comment on this with respect to part a) of the question. **(.2)**

Question 5: *LDA*

a) Given two classes LDA finds an *optimal* projection $\mathbf{w}$, where $y = \mathbf{w}^t\,\mathbf{x}$. In words which criterion does it use to find this optimal projection. **(.2)**

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

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

Consider the following example:

$$p(\mathbf{x} \,|\, \omega = 1) = \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}10^2 & 0\\ 0 & 2^2\end{pmatrix}\right) \quad \textbf{and} \quad p(\mathbf{x} \,|\, \omega = 2) = \mathcal{N}\left(\boldsymbol{\mu}, \begin{pmatrix}2^2 & 0\\ 0 & 10^2\end{pmatrix}\right)$$

Answer the following questions:

i) What is the optimal projection according to LDA? **(.05 × 4)**

ii) Sketch the distribution of class 1 and class 2 points when they have been projected into 1d via the optimal projection. **(.05 × 4)**

iii) 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? **(.05 × 4)**

for $\boldsymbol{\mu} = (0, 0)^t, (.1, 0)^t, (0, .1)^t$ and $(.1, .1)^t$.

c) The above example highlights which limitations of LDA? **(.2)**

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

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