Assessed Coursework 2

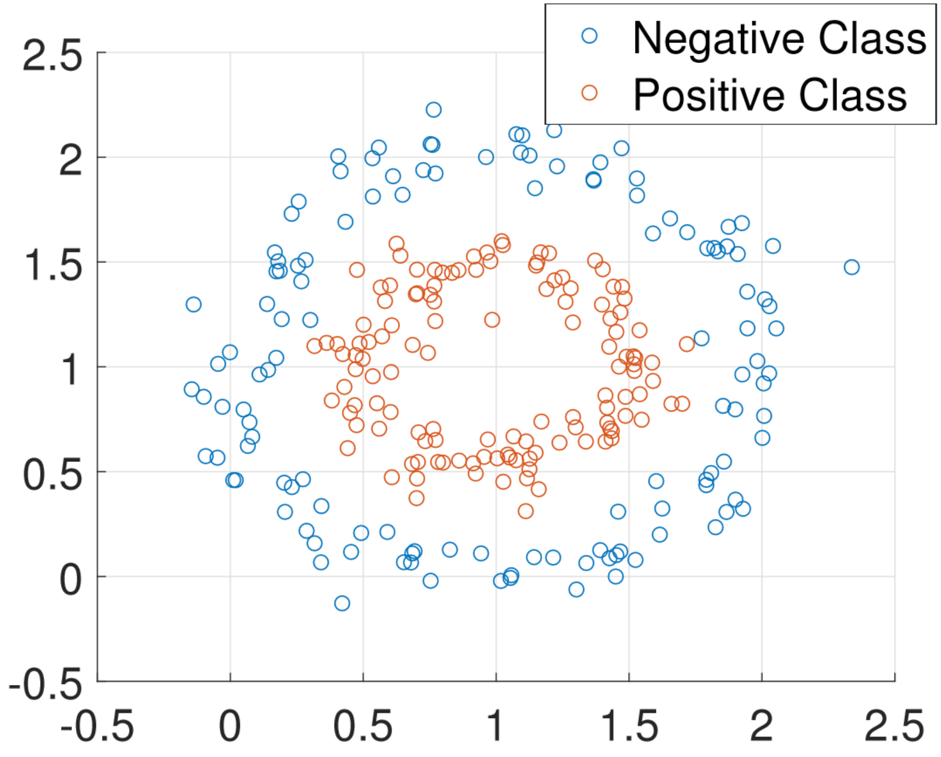
Please submit by 5pm, Monday

Description

- There are 4 questions, worth 20% in total.
- You can use whatever material you can find.
  - However, **do not** copy answers directly from internet.
  - Cite external sources properly.
- You are expected to complete these questions **independently**. Questions of the coursework should be directly addressed to the lecturer.
- You are recommended to use latex to typeset all the answers to the questions.

Q0: Least Square (LS) Classification (2 marks)

- The LS classifier using **which feature transform** function is likely to perform well on the dataset below and has the lowest computational cost?

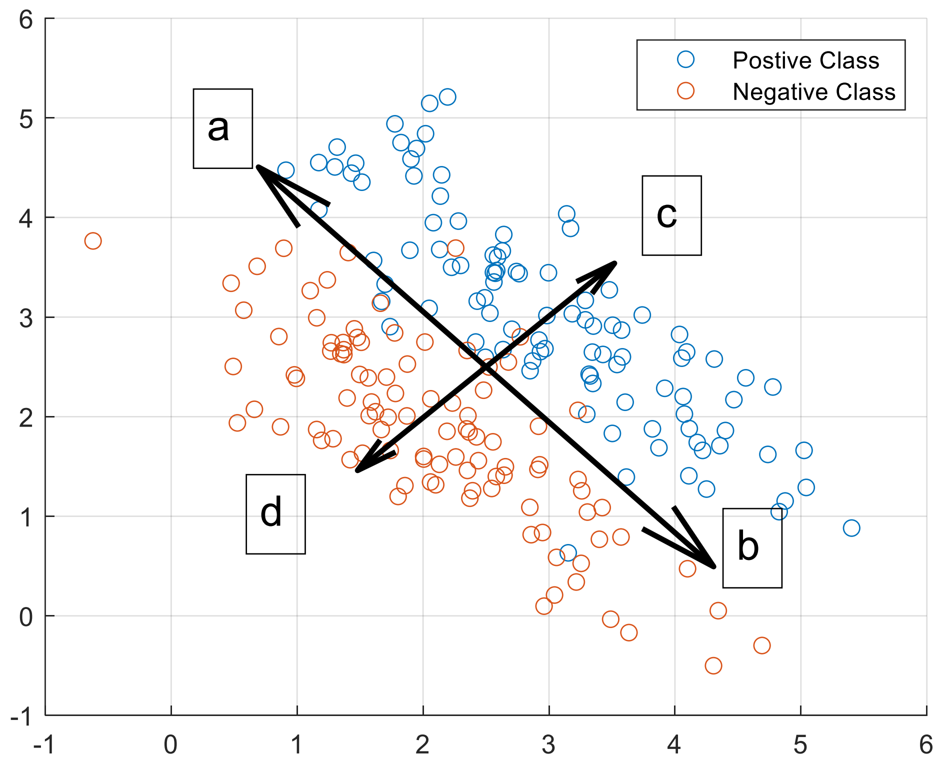


Q0: Least Square (LS) Classification

- A. No Feature Transform, $\phi(\boldsymbol{x}) = \boldsymbol{x}$.
- B. Polynomial feature transform, degree b = 1.
- C. Polynomial feature transform, degree b = 2.
- D. Polynomial feature transform, degree b = 3.
- E. Radius basis function feature transform, with number of basis, b = 50

Q1.1: Fisher Discriminant Analysis (FDA) (2 marks)

- Given the **classification dataset** visualized below, which of the following direction(s) are likely to be the direction of the FDA embedding vector $\boldsymbol{w}$? **Explain why In a few sentences**



Q1.2: Fisher Discriminant Analysis (FDA) (2 marks)

- Assume all datapoints in the previous figure are IID. Now construct a likelihood function **over the entire dataset** using a 2D Normal model $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.
- Denote the Maximum Likelihood Solution for $\boldsymbol{\Sigma}$ as $\boldsymbol{\Sigma}_{\mathrm{ML}}$. $\boldsymbol{\Sigma}_{\mathrm{ML}}$ being a symmetric positive definite matrix, can be decomposed as

$$\boldsymbol{\Sigma}_{\mathrm{ML}} = [\boldsymbol{u}_1, \boldsymbol{u}_2] \begin{bmatrix} D_1 & 0 \\ 0 & D_2 \end{bmatrix} [\boldsymbol{u}_1, \boldsymbol{u}_2]^{\top},$$

$$\boldsymbol{u}_1, \boldsymbol{u}_2 \in R^2, D_1 > D_2$$

- Which direction(s) are the possible direction for $\boldsymbol{u}_1$? Explain why in a few sentences.

Q3: Support Vector Machines (SVM) (2 marks)

- One drawback of classic SVM is that it does not take the costs of making wrong decisions of into account: false positive and false negative may have different weights in different applications.

- Suppose in our application, false negative is 1000 times more dangerous than false positive. Make **minor modifications** on soft-margin SVM objective function to reflect such a cost in this application.

Q3: Support Vector Machines (SVM)

- Recall, Soft-margin SVM:

- Minimize $||\boldsymbol{w}'||^2 + \sum_{i \in D} \epsilon_i$
- Subject to $\forall_{i \in D}, y_i f(\boldsymbol{x}_i; \boldsymbol{w}) + \epsilon_i \geq 1, \epsilon_i \geq 0$

- Hint: if it helps, you can convert the soft margin SVM to a formulation using the loss function. The lecture in Week 7 may be useful.

Q4.1 Markov Network (2 marks)

- Suppose P is a **sparse** Gaussian Markov Network of **5** random variables (here **sparse** means total number of edges is less than half of number of edges in a complete graph). Which of the following is likely to be the **covariance matrix** of P.

A
$$\begin{pmatrix} 1 & 0.5 & 0 & 0 & 0 \\ 0.5 & 1 & 0.5 & 0 & 0 \\ 0 & 0.5 & 1 & 0.5 & 0 \\ 0 & 0 & 0.5 & 1 & 0.5 \\ 0 & 0 & 0 & 0.5 & 1 \end{pmatrix}$$

B
$$\begin{pmatrix} 1 & 0.5 & 0 & 0 & 0 \\ 0.5 & 1 & 0.5 & 0 & 0 \\ 0 & 0.5 & 1 & 0.5 & 0 \\ 0 & 0 & 0.5 & 1 & 0.5 \\ 0 & 0 & 0 & 0.5 & 1 \end{pmatrix}^{-1}$$

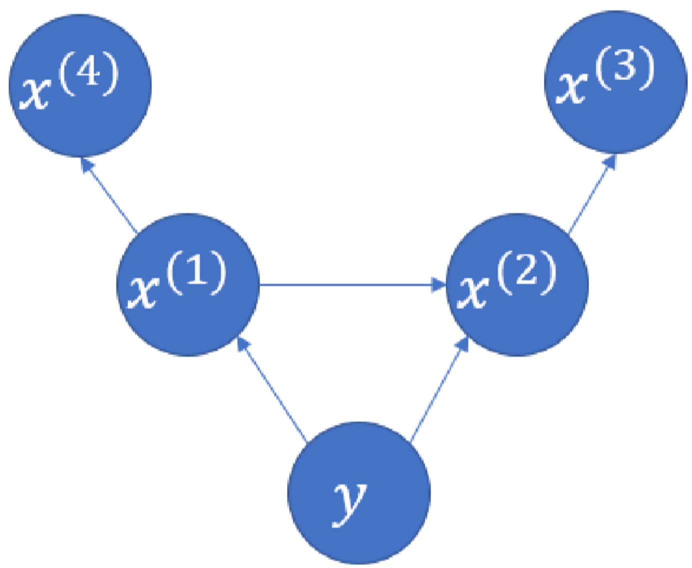
C
$$\begin{pmatrix} 1 & 0.5 & 0 & 0 & 0 \\ 0.5 & 1 & 0.25 & 0 & 0 \\ 0 & 0.5 & 1 & 0.5 & 0 \\ 0 & 0 & 0.5 & 1 & 0.5 \\ 0 & 0 & 0 & 0.5 & 1 \end{pmatrix}^{-1}$$

D
$$\begin{pmatrix} 1 & 0.1 & 0.25 & 0.1 & 0.1 \\ 0.1 & 1 & 0.1 & 0.1 & 0.25 \\ 0.25 & 0.1 & 1 & 0.1 & 0.25 \\ 0.1 & 0.1 & 0.1 & 1 & 0.1 \\ 0.1 & 0.25 & 0.25 & 0.1 & 0 \end{pmatrix}$$

E
$$\begin{pmatrix} 1 & 0.1 & 0.25 & 0.1 & 0.1 \\ 0.1 & 1 & 0.1 & 0.1 & 0.25 \\ 0.25 & 0.1 & 1 & 0.1 & 0.25 \\ 0.1 & 0.1 & 0.1 & 1 & 0.1 \\ 0.1 & 0.25 & 0.3 & 0.1 & 1 \end{pmatrix}$$

Q4.2 Bayesian Network (6 marks)

- Given a dataset $D := \left\{\left(y, x^{(1)}, x^{(2)}, x^{(3)}, x^{(4)}\right)_i\right\}_{i=1}^{n}, y \in \{-1,1\}$ whose joint distribution is a **Bayesian network** described by the following graph.



- Write down the factorization of the joint probability $p\left(y, x^{(1)}, x^{(2)}, x^{(3)}, x^{(4)}\right)$ according to the graph. [2 marks]
- Write down all the conditional independence encoded by this graph. [2 marks]
- Knowing such a graphical model, should I use all input features $x^{(1)}, x^{(2)}, x^{(3)}, x^{(4)}$ to predict y? Why? [2 marks]