Closed Book Midterm
Spring 2018

Department of Computer Science
Selected Topics II (Data Science II)

15 marks

1. Given the following datasets: $\omega_1: \ \{(1,1),(1,-1),(-1,1),(-1,-1)\}$, and $\omega_2: \ \{(2,2),(2,-2),(-2,2),(-2,-2)\}$.
   - draw the datasets.
   - design a QDA for general threshold. (**Hint:** find frist Σ and Σ^{-1})
   - what is the geometry of the decision surface, and how does the threshold value affects it?
   - draw the decision surface for the case of equal priors and costs, and indicate the point of intersection with the two coordinates.

15 marks

2. Consider the following decision function $f(X) = (1,2)(X - (2,4)')$.
   - Draw the decision boundary $f(X) = 0$.
   - If the decision regions are defined by $f(X) \underset{\omega_2}{\overset{\omega_1}{\gtrless}} 0$, what is the decision for $(0,0)$.
   - Consider the decision surface $f(X) = c, \ c > 0$. Draw this surface with the first surface above on the same figure.

15 marks

3. In a single dimensional two-class problem. Class ω_1 is non-diseased patients ("Negative"), and class ω_2 is diseased patients ("Positive"). The error e_{21} is defined as the error (or probability of misclassification) that results from classifying the diseased as non-diseased, called the False Negative Fraction (FNF). It is obvious that the True Positive Fraction (TPF) and the True Negative Fraction (TPF) are related to the other two quantities as: $TPF = 1 - FNF$ and $TNF = 1 - FPF$. Similarly, e_{12} is called the False Positive Fraction (FPF). The PDF of the two classes is given by:

$$f(X \mid \omega_1) = \begin{cases} ax^2 & 0 < x < 4 \\ 0 & \text{Otherwise} \end{cases},$$
$$f(X \mid \omega_2) = \begin{cases} -b(x-3) & 0 < x < 3 \\ 0 & \text{Otherwise} \end{cases},$$

   - Find a and b and draw the two density functions.
   - Sketch a dataset withdrawn from these distributions.
   - Derive the best decision function for this problem in terms of $\theta = \frac{c_{21}P(\omega_2)}{c_{12}P(\omega_1)}$ and draw it, along with the two decision regions.
   - Suppose that the decision point (not necessarily the Bayes') is taken as $X = c$. Derive an expression for each of FPF and FNF.
   - Derive an expression to express the TPF as a function of the FPF.
   - Is this a problem of "supervised" or "unsupervised" learning?
   - Suppose that the misclassification costs are λ_{12} and λ_{21} respectively; then write an expression for the risk.