University
ty of Engineering
puter Engineering Dept.

Pattern Classification & NNs
Fall 2021 - Midterm Exam
Duration: 1 Hour

Name.

1. What is machine learning? (4 points)

It's a subfield of AI that uses statistical methods to determine some outputs corresponding to some inputs like classification problems. (4)

2. State the difference between classification and regression problems. Give examples. (2)

- Classification: the desired outputs is some specific value from set of values like classify an object to be cat or dog. (2 points)

Regression: the desired output is a quantity like expect the pricing of house or temperature level or. - - -

3. What is the feature space? and how decision regions are constructed? (2)

- Feature space is the all possible values that a feature can take and the features are mapped on a 2-D or 3-D or N-D Planes to see the distributions. (2 points)
- Decision regions are constructed by finding a boundry between classes that can seperate those classes clearly. ex: linear boundry, non-linear.

4. Discuss the K-Nearest Neighbor Classifier. Mention its merits & demerits.

the K-nearest calculates the distances between the test Point and all the points in the space and takes the K min Points that are near to the testPoint (min distances) and take the majority vote from the K Points to determine which class should the testPoint belongs to (5 points) (5)

Advantages: As it takes the vote from multiple Points the outliers cannot affect the result in comparison with the nearest neighbour that takes only one Point so if the Point is outlier the test Point may be classified wrong

Disadvantages: It's sensitive to K if K is too large the decision may be wrong as it could take Points from another regions into considrations.

13.5

5. Why are Bayes classifiers considered as optimum classifiers? and what are the necessary conditions for that? (3 points)

2

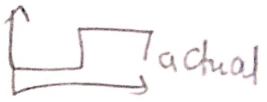
- AS bayes classifiers always take the class with highest probabili so it's always the optimum decision.
- But this consideration is just theoritical not practical as the bayes assumes that it has the distribution of the data but in real we have the distribution estimated from the training data only so it may be not accurate or biased.

6. Discuss kernel density estimation. (8 points)

7.5

the Kernel density estimation is used to get the denisty of data by using bump functions for each point the get the sum of the bumpfunctions to get estimated density distribution as it's continuous it's better than the histogram and naïve methods

we can take the bump function as $\phi(x) = \frac{1}{h} g(\frac{x}{h})$



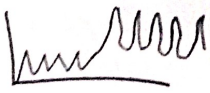
ex, bump function of gaussian distribution

h: is the width of the standard deviation in case of gaussian

7. How does the performance of the kernel density estimation rely on the kernel width? (4 points)

2

if the Kernel width (h) is too small the distribution will be too bumpy



if the(h) is too large it will be so smooth that may loose some information

as the actual dist ⇒

$$opt(h) = \sigma \left(\frac{4}{(N+2)H}\right)^{\frac{1}{N+4}}$$

Compare between wrapper and filter feature selection methods. (5 points)

Filter type : It doesn't Consider the classifier used just the relations between the features.

Wrapper type : It takes into Consideration the classifier used as sfs or sbs. - -

4

9. Write the pseudo code of Multi-Class AdaBoost classifier. (8 points)

① Put the weight $w_m = \frac{1}{M}$ equal

② for t : 1 to T

③ get the h_t that best fit the data

④ calculate the error for $h_t \rightarrow err = \frac{\sum w_m \text{ that } (h_t(x) \neq C(x))}{\sum w_m}$ ← error in estimation

⑤ Calculate the weight of the classifier $\alpha = \log\left(\frac{1-err}{err}\right) + \log(K-1)$

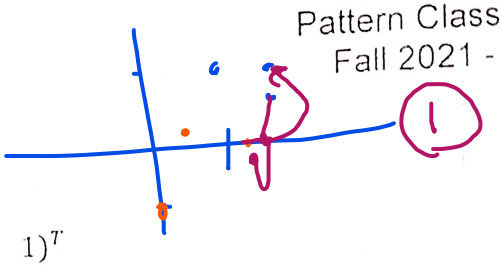
⑥ re calculate the weights $w_m = w_m e^{\alpha \text{ if } (h_t(x) \neq C(x))}$

⑦ re normalize the weights, return to ②

the output $\Rightarrow C(x) = \underset{K}{\text{argmax}} \sum \alpha_m \cdot \text{if}(h_t(x) = K)$

8

Cairo University
Faculty of Engineering
Computer Engineering Dept.



15

10. Consider the following problem:

Class 1 patterns: $(1 \quad 1)^T, (1.5 \quad 0.7)^T, (1.5 \quad 1)^T$

Class 2 patterns: $(0 \quad -1)^T, (0.1 \quad 0.3)^T, (1.2 \quad 0)^T$

2

a) Assume that we would like to use the minimum distance classifier. What would be the classification of the following pattern $(1.3 \quad 0)^T$? What would be the classification is we use K-nearest neighbor classifier where K = 3.

10 (10 points)

① minimum distance:

$M_1 = \left(\frac{1+1.5+1.5}{3}, \frac{1+0.7+1}{3}\right) = \left(\frac{4}{3}, \frac{9}{10}\right)$ $\quad d_1^2 = \left(\frac{4}{3}-1.3\right)^2 + \left(\frac{9}{10}-0\right)^2 = 0.811$

$M_2 = \left(\frac{0+0.1+1.2}{3}, \frac{-1+0.3+0}{3}\right) = \left(\frac{13}{30}, \frac{-7}{30}\right)$ $\quad d_2^2 = \left(\frac{13}{30}-1.3\right)^2 + \left(\frac{-7}{30}-0\right)^2 = 0.805$

$d_2 < d_1 \rightarrow$ So (1.3, 0) is classified as class 2

② K-nn

	Point	distance d^2
P_1	(1, 1)	$(1-1.3)^2+(1-0)^2 = 1.09$
C_1 P_2	(1.5, 0.7)	$(1.5-1.3)^2+(0.7-0)^2 = 0.53$
P_3	(1.5, 1)	$(1.5-1.3)^2+(1-0)^2 = 1.04$
P_4	(0, -1)	$(0-1.3)^2+(-1-0)^2 = 2.69$
C_2 P_5	(0.1, 0.3)	$(0.1-1.3)^2+(0.3-0)^2 = 1.53$
P_6	(1.2, 0)	$(1.2-1.3)^2+(0-0)^2 = 0.01$

K nearest are P_2, P_3, P_6 (P_2, P_3: C_1; P_6: C_2) $\rightarrow$ majority Vote

Class 1

b) Compare between the results obtained in (a).

5 (5 points)

the result of the K-nn is more accurate than the min distance as the P_6 at class 2 is outlier so it affects the result. but the K-nn takes the majority Vote.

11. Consider a two-dimensional two-class classification problem, where the class-conditional densities are given by:

$$p(\underline{X}|C_1) = 0.5\,\frac{1}{2\pi}\,e^{\frac{-(x_1^2+x_2^2)}{2}} + 0.5\,\frac{1}{2\pi}\,e^{\frac{-((x_1-1)^2+(x_2-1)^2)}{2}}$$

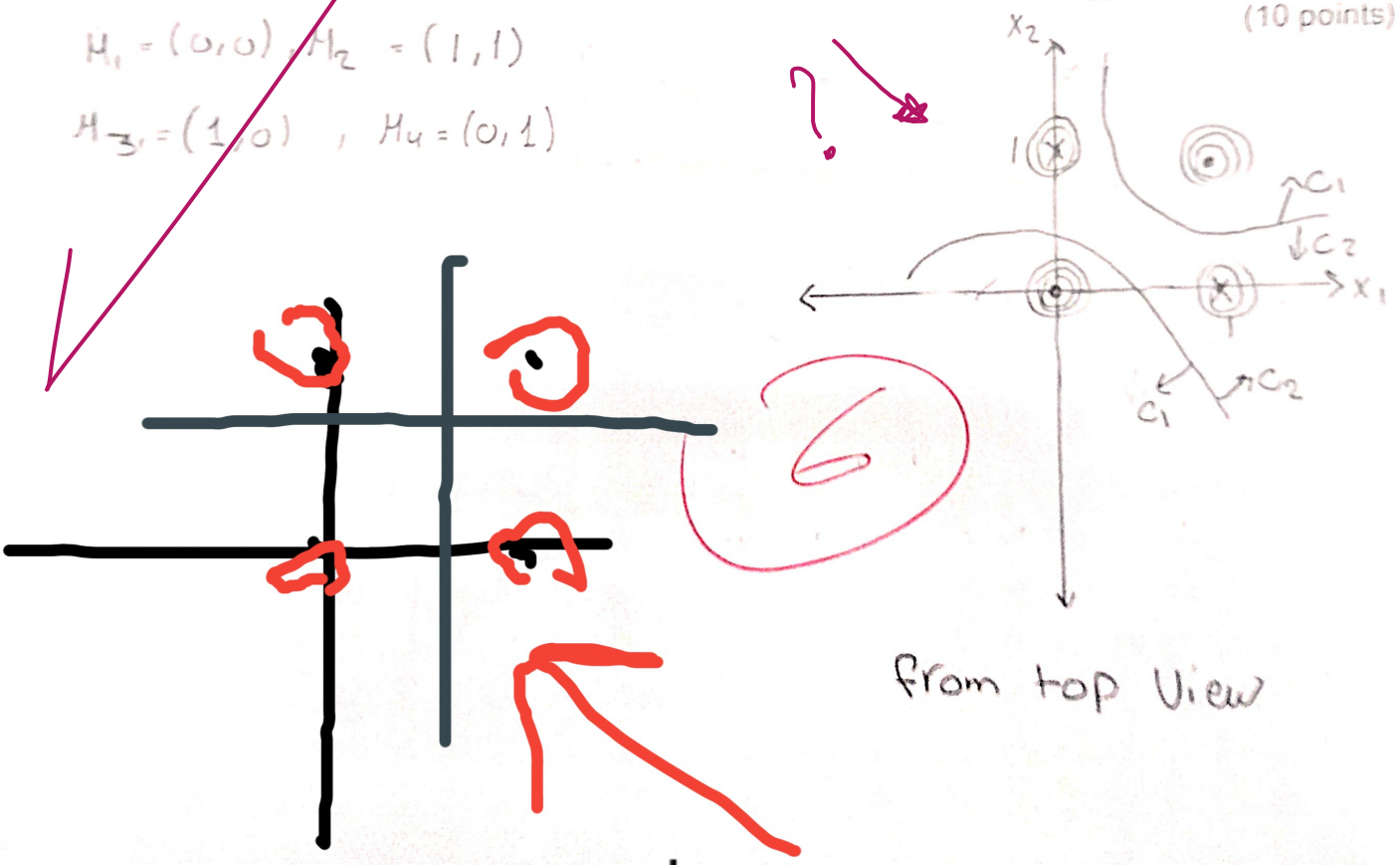
$$p(\underline{X}|C_2) = 0.5\,\frac{1}{2\pi}\,e^{\frac{-((x_1-1)^2+x_2^2)}{2}} + 0.5\,\frac{1}{2\pi}\,e^{\frac{-(x_1^2+(x_2-1)^2)}{2}}$$

Assume that $P(C_1) = P(C_2) = 0.5$.

Sketch the approximate decision boundary and point out the classification regions (10 points)

$\mu_1 = (0,0)$, $\mu_2 = (1,1)$

$\mu_3 = (1,0)$, $\mu_4 = (0,1)$



Right answer

12. Estimate the mean and covariance matrix of the following data points:
$(1 \quad 1)^T, (1.3 \quad 0.7)^T, (1.5 \quad -0.1)^T, (0 \quad -1)^T, (0.1 \quad 0.3)^T, (1.2 \quad 0.1)^T$

(10 poin

$$M = \left(\frac{1+1.3+1.5+0+0.1+1.2}{6}, \frac{1+0.7-0.1-1+0.3+0.1}{6} \right)$$

$$M = (0.85, 0.16667)$$

$$\Sigma = Cov(X,Y) = E[(X-M_1)(Y-M_2)] = \frac{1}{M}\sum (X-M_1)(Y-M_2)$$

$$Cov = \begin{bmatrix} 0.3425 & 0.1766 \\ 0.1766 & 0.1833 \end{bmatrix}_{2\times 2}$$

9

13. Discuss the main idea of the principle component analysis (PCA) and its key equations with derivations.

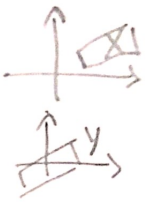
(10 points)

the PCA is a method for feature selection as we have a distribution of the data that is rotated and shifted so we need to transform this distribution to beat a direction of one of the most effective features to make the covariance matrix diagonal so then we can select the most effective features from the feature space by selecting the ones that have the highest variances (eigen values)

to estimate the PCA

① translate the data to the origin by subtracting the mean

$Y = X - M$



② estimate the covariance matrix of the Y then find the eign vectors $\underline{U}$ and eign values. $\underline{\Omega}$

③ estimate $Z = \underline{U}^T Y$

④ $Cov(Z) = U^T \Sigma U$

$= \underline{\Omega}$

10

⑤ select L features from the most eign values

then $Z = \begin{bmatrix} U_1^t \\ U_2^t \\ \vdots \\ U_L^t \end{bmatrix} Y$