CS 559: Midterm
Duration: 2:00

Answer all problems. Show all calculations and provide sufficient explanation. If in doubt, explain more.

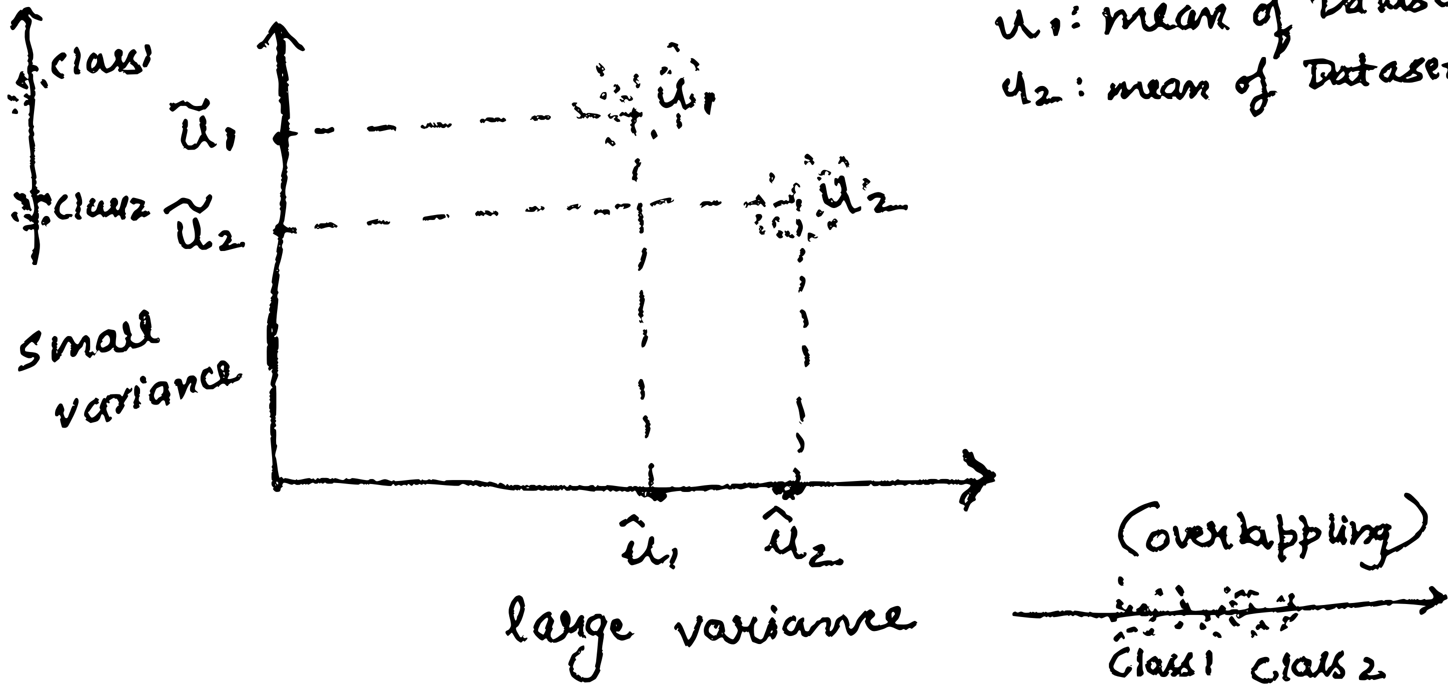
Write your name on this page and additional pages (if necessary)

NAME: Archana Kalburgi

CWID: 10469491

Problem 1. (10 Points) In the context of Fisher's Linear Discriminant Analysis, explain why maximizing the distance between the projected class means is not sufficient for obtaining well-separated data after the projection. A simple sketch may be helpful for your answer.

u_1: mean of Dataset 1
u_2: mean of Dataset 2.



For FDA.

1) Projected mean must be far apart.

2) Projected samples of class 1 & class 2 must be close to each other

Maximising distance betn the projected class means is not sufficient, we need to maximise the variance the sample, according to,

u_1 : mean of class c_1
u_2 : mean class c_2

$$J(v) = \frac{(u_1 - u_2)^2}{s_1^2 + s_2^2} \begin{matrix}\uparrow \\ \downarrow\end{matrix}$$

s_1^2 : scatter c_1
s_2^2 : scatter c_2

Problem 2. (15 Points)

(1) Maximum Likelihood Estimation (MLE) techniques assume a certain parametric form for the class-conditional probability density functions. This implies that (select one only) (5 pt):

✓ a. The form of the decision boundaries is also determined in some cases.

b. The form of the decision boundaries is always unpredictable.

(2) Linear Discriminant Analysis techniques assume a certain parametric form for the decision boundaries. This implies that (select one only) (5 pt):

a. The form of the class-conditional densities is also determined in some cases.

✓ b. The form of the class-conditional densities remains unknown in general.

(3) Briefly explain your above-two answers. (5 pts)

(1) covariance is identical but arbitrary.
decision boundary is : Hyperplane

when covariance is different.
decision boundary : Hyperquadric

(2) 2 class : hyperplane

multiclass : given by distance.

Problem 3. (15 points) Let D denote the data samples and H denote hypothesis. Provide the relationship between the following probability pairs, using one of the following operators: (1) $=$, (2) $\leq$, (3) $\geq$, and (4) (depends)
Explain your answers briefly.

$$P(H=h|D=d) = \frac{P(H \cap D)}{P(D)} \text{ — } 3$$

(a) $\sum_h P(H = h|D = d)$ and 1 (3 pts) $\leq$

(b) $\sum_h P(D = d|H = h)$ and 1 (3 pts) $\leq$

(c) $\sum_h P(D = d|H = h)P(H = h)$ and 1 (3 pts) $\leq$

(d) $P(H = h|D = d)$ and $P(H = h)$ (3 pts) $\leq$

(e) $P(H = h|D = d)$ and $P(D = d|H = h)P(H = h)$ (3 pts) $\leq$

a) $\Sigma_h P(H=h|D=d) \leq 1$ — 1

Hypothesis can be true given all observation, but some of them can be false.

b) Bayes rule:

$$P(D=d|H=h) = \frac{P(H=h|D=d).P(D=d)}{P(H=h)}$$

from 1

$P(H=h|D=d) \leq 1$

& $P(H=h) \leq 1$

$P(D=d) = 1$

$\therefore P(D=d|H=h) \leq 1$ — 2

c) $$\Sigma P(D=d|H=h) = \frac{P(H=h|D=d)\cdot P(D=d)}{P(H=h)}$$

$\Sigma P(D=d|H=h)\cdot P(H=h) = P(H=h|D=d)\cdot P(D=d)$

from 1 & 2

$P(H=h|D=d)\cdot P(D=d) \leq 1$

d) $P(H=h|D=d) \leq P(H=h)$

$\because P(H=h) = 1$

e) $P(H=h|D=d) \leq P(D=d|H=h)\cdot P(H=h)$

from 3

Problem 4. (20 points) Let x be a one-dimensional binary (0 or 1) variable following a Bernoulli distribution:

$$P(\mathbf{x}|\theta) = \theta^x(1-\theta)^{1-x},$$

where θ, the probability that $x = 1$, is the unknown parameter to be estimated. Show that the maximum-likelihood estimate for θ is

$$\hat{\theta} = \frac{1}{n}\sum_{k=1}^{n} x_k$$

Making estimation based on n points

$x_1, x_2, x_3 \cdots x_n$: IID random variables, samples from Bernoulli's

- for Bernoulli,

$$f(x) = p^x (1-p)^{1-x}$$

MLE Estim

$$L(\theta) = \prod_{K=1}^{n} p^{x_K} (1-p)^{1-x_K}$$

$$LL(\theta) = \sum_{K=1}^{n} \log p^{x_K} (1-p)^{1-x_K}$$

$$= \sum_{K=1}^{n} x_K (\log p) + (1-x_i) \log (1-p)$$

$$= Y \log p + (n-Y) \log (1-p)$$

where $Y = \sum_{i=1}^{n} x_K$

Chose a value of p that maximizes log likelihood, by taking derivative & setting it to 0

$$\delta LL(P) = Y \frac{1}{P} + (n-Y) \frac{-1}{1-P} = 0$$

$$\hat{p} = \frac{Y}{n} = \frac{1}{n} \cdot \sum_{K=1}^{n} x_K$$

Problem 5. True or False. No explanation needed. (20 points)

(1) MLE and MAP never produce the same result. False

(2) Posterior is always higher than prior. False

(3) One can find a closed-form solution for any optimziation problem. TRUE

(4) When applying generative models, we usually assume that the parameter estimation for each class is independent. TRUE

(5) K-NN is a parameteric approach. FALSE

(6) Histogram estimation is a parametric approach. FALSE

(6) A belief network is a directed acyclic graph. TRUE

(7) PCA can be sovled using SVD. TRUE

(8) PCA can be applied for face detection. TRUE

(9) Covariance matrix captures the shape of a distribution. TRUE

(10) A susbpace must pass through the origin. TRUE

Problem 6. (20 points) Assume we are given a set of D dimensional data samples.
PCA: Principal Component Analysis LDA: Fisher Linear Discriminant Analysis

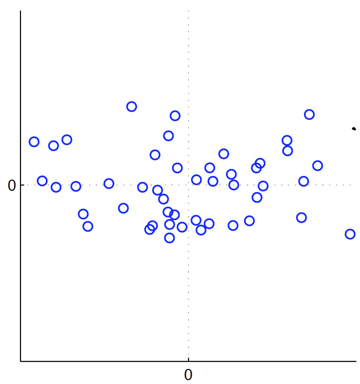
(a) Which quantity does PCA maximize in order to obtain the first projection direction? (3 pts)

(b) Which quantity does PCA minimize in order to obtain the first projection direction? (3 pts)

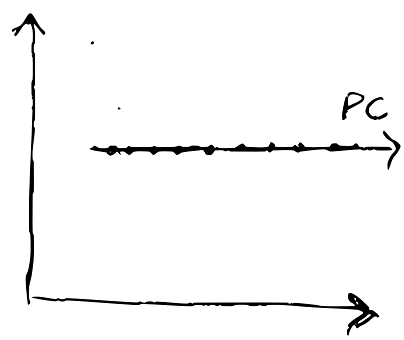
(c) Which quantity does LDA maximize in order to obtain the first projection direction? (3 pts)

(d) Consider a data set with two data points: $(2, 2), (-2, -2)$. Compute the covariance matrix (3 pts)

Suppose the covariance matrix of the two-dimensional data set plotted below is $\begin{bmatrix}\alpha, 0\\ 0, \beta\end{bmatrix}$. We assume the horizontal axis corresponds to the first dimension and the vertical one corresponds to the second.



e)



(e) Draw the first principal direction estimated from the data. (4 pts)

(f) How large is the variance of the data projected on the first principal component? (4 pts)

a) Projection variance

b) Reconstruction Error

c) Difference between the mean of two classes.

d) covariance matrix $= \dfrac{\Sigma(x_1-\bar{x})(y-\bar{y})}{N-1} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$

f) The line that has the largest length has the largest variance i.e, 2α, assuming α is horizontal