

$$\text{Q1: } p(x,y) = p(y|x)p(x) \quad E[l(f(x),y)] = \int_x \left\{ \int_y \{ l(f(x),y) p(y|x) \} p(x) dx \right\}$$

$$(a) \quad l(f(x),y) = (f(x) - y)^2$$

$$\Rightarrow E[l(f(x),y)] = \int_x \left\{ \int_y (f(x) - y)^2 p(y|x) dy \right\} p(x) dx.$$

SINCE FOR ANY VALUE OF  $x$ ,  $f(x)$  CAN BE CHOSEN INDEPENDENTLY

$$\Rightarrow E[l(f(x),y)] = \int (f(x) - y)^2 p(y|x) dy.$$

FOR OPTIMAL SOLUTION  $\frac{dE}{df} = 0$

$$\Rightarrow \frac{dE[l(f(x),y)]}{df(x)} = 2 \int (f(x) - y) p(y|x) dy = 0$$

$$\Rightarrow \int f(x) p(y|x) dy = \int y p(y|x) dy.$$

$$\Rightarrow \frac{\int f(x) p(x,y) dy}{p(x)} = \int y p(y|x) dy$$

$$f(x) = \int y p(y|x) dy$$

$$f(x) = E[y|x]$$

$$(b) \quad l(f(x),y) = |f(x) - y|$$

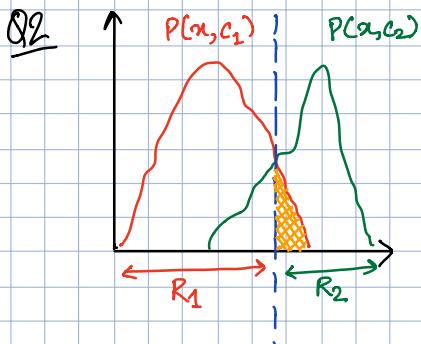
$$\Rightarrow E[l(f(x),y)] = \int_x \left\{ \int_y |f(x) - y| p(y|x) dy \right\} p(x) dx$$

$$\Rightarrow E[l(f(x),y)] = \int |f(x) - y| p(y|x) dy$$

FOR OPTIMAL SOLUTION  $\frac{dE}{df} = 0$

$$\Rightarrow \frac{d E[\ell(f(x), y)]}{df} = \int \text{sgn}(f(x) - y) p(y|x) dy$$

$$\Rightarrow \int_{f(x)}^{\infty} p(y|x) dy = \int_{-\infty}^{f(x)} p(y|x) dy$$



LET'S TAKE THE EXAMPLE OF TWO CLASSES  
 $R_1$  IS THE REGION WHERE  $P(c_1|x) > P(c_2|x)$   
 $R_2$  IS THE REGION WHERE  $P(c_2|x) > P(c_1|x)$

$\text{ERR}[y=c_1] = \text{SHADED REGION}$

$$\begin{aligned} &= \int_{R_2} P(x, c_1) dx \\ &= \int_{R_2} P(c_1|x) P(x) dx \\ &= \int_{R_1} P(c_1|x) P(x) dx + \int_{R_2} P(c_1|x) P(x) dx \\ &\quad - \int_{R_1} P(c_2|x) P(x) dx \\ &= \sum_{i=1,2} \int_{R_i} P(c_1|x) P(x) dx - \int_{R_1} P(c_1|x) P(x) dx \end{aligned}$$

GENERALIZING THE ABOVE TERM FOR N CLASSES, WE GET

$$\text{ERR}[y=c_j] = \sum_{i=1:N} \int_{R_i} P(c_j|x) P(x) dx - \int_{R_j} P(c_j|x) P(x) dx$$

Q3 (a) FISCHERS LDA FOR TWO CLASSES.

BETWEEN CLASS :  $S_B = (\mu_2 - \mu_1)(\mu_2 - \mu_1)^T$   
COVARIANCE

WITHIN CLASS :  $S_W = \sum_{m \in C_1} (x_m - \mu_1)(x_m - \mu_1)^T + \sum_{m \in C_2} (x_m - \mu_2)(x_m - \mu_2)^T$   
COVARIANCE

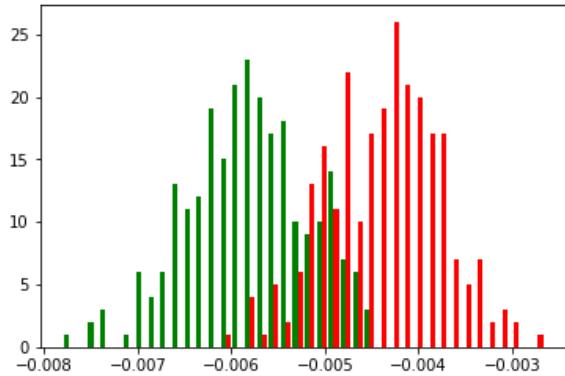
DESIABLE TO HAVE LOW WITHIN CLASS VARIANCE  
AND HIGH BETWEEN CLASS VARIANCE.

$J(W) = \frac{W^T S_B W}{W^T S_W W}$ . FISCHER'S CRITERION

MAXIMISING THIS GIVES

$$W \propto S_W^{-1}(\mu_2 - \mu_1)$$

PLOT OF  
BOSTON DATA  
ON BEING  
PROJECTED TO  
R1 DIMENSION



(b)  $S_B$  IS AN OUTER PRODUCT OF TWO VECTORS, THEREFORE IT HAS RANK 1. FOR MULTICLASS CASE,  $S_B$  IS WRITTEN AS

$$S_B = \sum_{k=1}^K N_k (\mu_k - \mu)(\mu_k - \mu)^T \quad (\text{WHERE } K \text{ IS THE # CLASSES})$$

$\Rightarrow S_B$  IS COMPOSED OF  $K$  MATRICES EACH OF WHICH ARE OF RANK 1  
MOREOVER, OUT OF THIS ONLY  $(K-1)$  ARE INDEPENDENT AND  
THEREFORE THERE ARE AT MOST  $(K-1)$  NONZERO EIGENVALUES.

THUS FOR TWO CLASS PROBLEM WE CANNOT PROJECT IP TO  
R2 DIMENSIONS.

(c) FISCHER'S LDA FOR MULTICLASS

$$S_B = \sum_{k=1}^K N_k (\mu_k - \mu)(\mu_k - \mu)^T$$

$$S_W = \sum_{k=1}^K S_k \quad \text{WHERE } S_k = \sum_{n \in C_k} (x_n - m_k)(x_n - m_k)^T$$

$$m_k = \frac{1}{N_k} \sum_{n \in C_k} x_n$$

$$J(W) = \text{Tr}\{S_W^{-1} S_B\} = \text{Tr}\{(W S_W W^T)^{-1} (W S_B W^T)\}$$

THE WEIGHT VALUES ARE THE EIGENVECTORS WITH  
THE D LARGEST EIGENVALUES.

GAUSSIAN GENERATIVE MODELLING:

$$p(y|x) = p(x|y)p(y) \leftarrow \begin{matrix} \text{CLASS PRIOR} \\ \uparrow \\ p(x) \end{matrix}$$

CLASS CONDITIONAL

$$p(x|C_k) = \frac{1}{(2\pi)^{d/2}} \frac{1}{|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2} (x - \mu_k)^T \Sigma^{-1} (x - \mu_k)\right\}$$

$\mu_k$  IS THE CLASS MEAN &  $\Sigma$  IS THE SAME FOR ALL

GAUSSIAN GENERATIVE MODELLING ON DIGITS DATASET PROJECTED TO 2 DIMENSIONS

Mean Accuracy: 0.7177777777777778

Standard Deviation in Accuracy: 0.017838438175558125

## Q4 (a) LOGISTIC REGRESSION (MULTI CLASS)

CLASS POSTERIOR  $P(c_k | x) = \pi_k(w_k^T x) = \frac{\exp(\alpha_k)}{\sum_j \exp(\alpha_j)}$

NEGATIVE LOG LIKELIHOOD =  $-\sum_{n=1}^N \sum_{k=1}^K y_{nk} \log \pi_{nk}$

GRADIENT =  $\sum_{n=1}^N (\pi_{nk} - y_{nk}) x_n$

✓

FOR TWO CLASS

$$P(1|x) = \frac{\exp(w^T x)}{1 + \exp(w^T x)} = \sigma(w^T x)$$

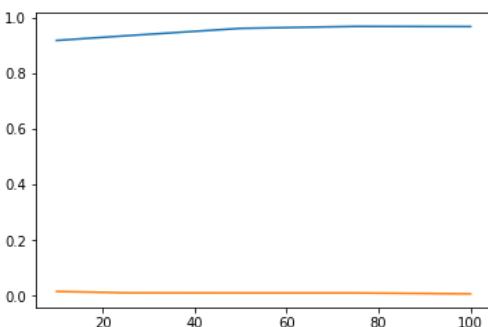
$$\text{LOG-LIKELIHOOD} = \sum_{n=1}^N y_n \log P(1|x_n) + (1-y_n) \log (1-P(1|x_n))$$

$$\text{GRADIENT} : \sum_{n=1}^N (\pi(w_t; x_n) - y_n) x_n = x^T (\pi(w_t; x) - y)$$



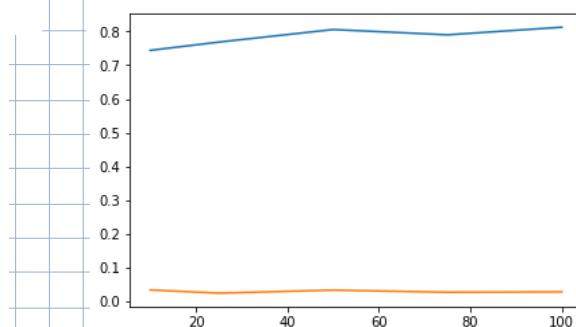
$$w_{t+1} = w_t - \alpha_t \nabla E(w_t)$$

LOGISTIC REGRESSION CLASSIFIER ON DIGITS DATASET  
 10 percent : Mean Accuracy: 0.9184065934065935  
     Standard Deviation: 0.015199534278706649  
 25 percent : Mean Accuracy: 0.9351648351648352  
     Standard Deviation: 0.010220371009746293  
 50 percent : Mean Accuracy: 0.9618131868131867  
     Standard Deviation: 0.009863364324734196  
 75 percent : Mean Accuracy: 0.968956043956044  
     Standard Deviation: 0.009832708883611678  
 100 percent : Mean Accuracy: 0.9684065934065934  
     Standard Deviation: 0.006413526115345445



Logistic Regression on Digits Dataset

LOGISTIC REGRESSION CLASSIFIER ON BOSTON DATASET  
 10 percent : Mean Accuracy: 0.8107142857142857  
     Standard Deviation: 0.021189494723377656  
 25 percent : Mean Accuracy: 0.8741758241758241  
     Standard Deviation: 0.02080669915248514  
 50 percent : Mean Accuracy: 0.8832417582417582  
     Standard Deviation: 0.01100616792528028  
 75 percent : Mean Accuracy: 0.9013736263736263  
     Standard Deviation: 0.012735524160988195  
 100 percent : Mean Accuracy: 0.9030219780219781  
     Standard Deviation: 0.012592490809193247



Logistic Regression on Boston Dataset

(b) NAIVE BAYES

$$P(x_i | C_k) = N(\mu_{ik}, \sigma_{ik}^2)$$

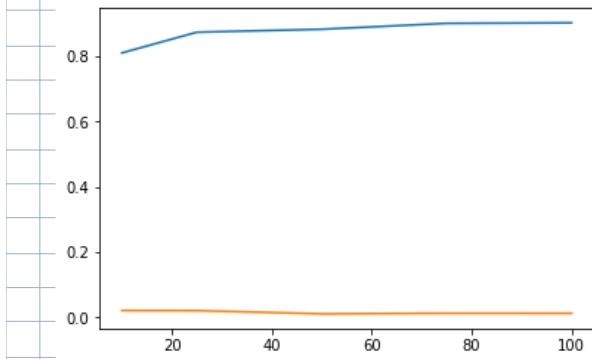
$$P(x | C_k) = \prod_{i=1}^D P(x_i | C_k) = \frac{1}{(2\pi)^{D/2} (\prod_{i=1}^D \sigma_{ik})} \exp \left\{ -\sum_{i=1}^D \frac{(x_i - \mu_{ik})^2}{2\sigma_{ik}^2} \right\}$$

NAIVE BAYES CLASSIFIER ON DIGITS DATASET

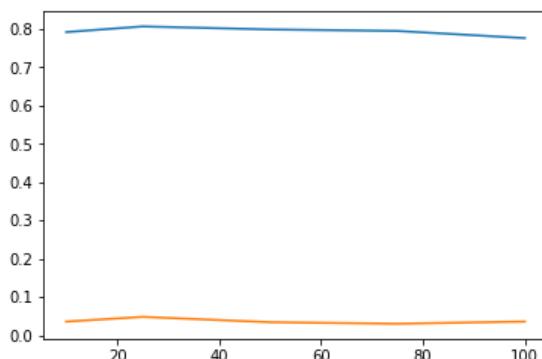
10 percent :	Mean Accuracy:	0.8107142857142857
	Standard Deviation:	0.021189494723377656
25 percent :	Mean Accuracy:	0.8741758241758241
	Standard Deviation:	0.02080669915248514
50 percent :	Mean Accuracy:	0.8832417582417582
	Standard Deviation:	0.01100616792528028
75 percent :	Mean Accuracy:	0.9013736263736263
	Standard Deviation:	0.012735524160988195
100 percent :	Mean Accuracy:	0.9030219780219781
	Standard Deviation:	0.012592490809193247

NAIVE BAYES CLASSIFIER ON BOSTON DATASET

10 percent :	Mean Accuracy:	0.7911764705882354
	Standard Deviation:	0.03563290557918874
25 percent :	Mean Accuracy:	0.8058823529411765
	Standard Deviation:	0.04798916960988398
50 percent :	Mean Accuracy:	0.7980392156862746
	Standard Deviation:	0.034018336417446005
75 percent :	Mean Accuracy:	0.7941176470588234
	Standard Deviation:	0.0300582543465802
100 percent :	Mean Accuracy:	0.7754901960784313
	Standard Deviation:	0.03574063947756507



Naive Bayes on Digits Dataset



Naive Bayes on Boston Dataset