ELE888 and EE8209:

Intelligent Systems, Midterm Exam

Name:_____

Student Number:_____

Program:_____

Problem 1 (20 points):	
Problem 2 (20 points):	
Problem 3 (20 points):	
Problem 4 (20 points):	
Problem 5 (20 points):	
Total (100 points):	

Instructions

This is a closed book exam. Everything you need in order to solve the problems is supplied in the pages attached to this.

The Exam starts at 2:10pm and ends at 4:10pm. It contains 5 problems. You have 120 minutes to earn a total of 100 points. Answer each question in the space provided. If you need more room, write to the back side of the paper and indicate that you have done so. If you detach the pages of this exam, write your name in all the pages detached.

Besides having the correct answer, being concise and clear is very important. For full credit you must show your work and explain your answers.

March 8, 2012

1. Suppose a bank classifies customers as either good or bad credit risks. On the basis of extensive historical data, the bank has observed that 1% of good credit risks and 10% of bad credit risks overdraw their account in any given month. A new customer opens a checking account at this bank. On the basis of a check with a credit bureau, the bank believe that there is a 70% chance the customer will turn out to be a good credit risk.

   (a) (5 pts) Suppose that this customer's account is overdrawn in the first month. How does this alter the bank's opinion of this customer's creditworthiness?

   (b) (5 pts) Given (a), what would be the bank's opinion of the customer's creditworthiness at the end of the second month if there was not an overdraft in the second month?

(10 pts) Chapter 1 of our textbook introduced an example of a pattern classification system to separate sea bass from salmon. Along the same lines, consider a four-class vegetable sorting problem, with tomato, cabbage, red apple, and lettuce as the four classes. Briefly describe the (i) sensor to be used, (ii) preprocessing and segmentation problem, and (iii) features that may separate these four classes.

Let G and O represent the following events = G : customer is considered to be a good credit risk
O : customer overdraws checking account.

From the bank's historical data, we have $P(O/G)=0.01$ and $P(O/\bar{G})=0.1$ and we also know that the bank's initial opinion about the customers creditworthiness is given by $P(G)=0.7$

Given the information that the customer has an overdraft in the first month, the bank's revised opinion about the customer's creditworthiness is given by the conditional probability $P(G/O)$. Using Baye's theorem and the law of total probability

$$P(G/O)=\frac{P(O/G)P(G)}{P(O)}=\frac{P(O/G)P(G)}{P(O/G)P(G)+P(O/\bar{G})P(\bar{G})}=\frac{0.01\times 0.7}{0.01\times 0.7+0.1\times 0.3}=0.189$$

(b) According to (a) at the begining of the second month, the prior probability of the customer being a good risk is 0.189. Since the customer does not have an overdraft in the second month, the bank's opinion on the customer's creditworthiness at the end of the month is given by

$$P(G/\bar{O})=\frac{P(\bar{O}/G)P(G)}{P(\bar{O})}=\frac{P(\bar{O}/G)P(G)}{P(\bar{O}/G)P(G)+P(\bar{O}/\bar{G})P(\bar{G})}$$

$$=\frac{0.99\times 0.189}{0.99\times 0.189+0.9\times 0.811}=0.204$$

Some suggestions:

(c) i sensor : colour sensitive video camera

ii preprocessing + segmentation : filtering, registration

iii features : colour, size, etc.

2. Consider a two-category classification problem with one-dimensional Gaussian distributions $p(x|w_i) \sim N(\mu_i, \sigma^2), i = 1, 2$ (i.e. they have same variance but different means).

(a) (3 pts) Sketch the two densities $p(x|w_1)$ and $p(x|w_2)$ in one figure.

(b) (7 pts) Sketch the two posterior probabilities $P(w_1|x)$ and $P(w_2|x)$ in one figure, assuming same prior probabilities.

(c) (10 pts) Suppose there is a two-class one-dimensional problem with the Gaussian distributions: $p(x|w_1) \sim N(-1, 1)$, and $p(x|w_2) \sim N(4, 1)$ and equal prior probabilities. In order to find a good classifier, you are given as much training data as you would like and you are free to pick any method. What is the best error on test data that any learning algorithm can attain and why? (Try to keep your reasoning to two sentences or fewer.)

(a) 0% error

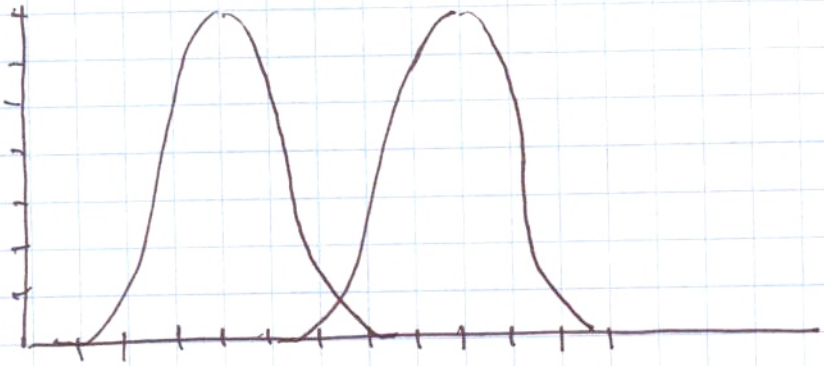
(b) more than 0% but less than 10%

(c) more than 20%

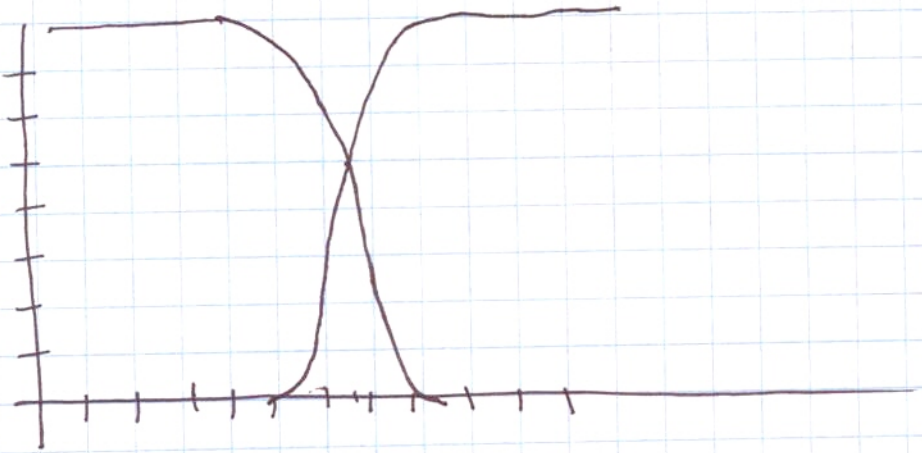
(d) can't tell from this information

2) A)

(a)



(b)



B) The best error can be achieved by using Bayes decision rule. The Bayes decision boundary is $x = 1.5$, which means that $|x - \mu| > 2\sigma$. Therefore the error rate should be less than 5%.

3. (8 pts) For a two class discrimination problem where each class is a Gaussian, the resulting decision boundary can sometimes be expressed as as a linear discriminant function. What conditions must be satisfied for this to be the case ?

(12 pts) In a two-class, two dimensional classification task the feature vectors are generated by two normal distributions sharing the same covariance matrix:

$$\begin{bmatrix} 1.1 & 0.3 \\ 0.3 & 1.9 \end{bmatrix}$$

and the mean vectors are $\mu_1 = [0, 0]^t$, $\mu_2 = [3, 3]^t$, respectively. Classify the vector $[1.0, 2.2]^t$ according to the Bayesian classifier.

Given that P(w)= P(w) = 0.5

3 (a)

case 1 and case 2 of the attached "cheat sheet" correspond to a linear discriminant function. case 3 is not.

(B) Calculate the Mahalanobis distance between the unknown vector and the two distributions.

$$r_i^2 = (x - \mu_i)^t \Sigma^{-1} (x - \mu)$$

$$\Sigma^{-1} = \begin{bmatrix} 1.1 & 0.3 \\ 0.3 & 1.9 \end{bmatrix}^{-1} = \frac{1}{2}\begin{bmatrix} 1.9 & -0.3 \\ -0.3 & 1.1 \end{bmatrix} = \begin{bmatrix} 0.95 & -0.15 \\ -0.15 & 0.55 \end{bmatrix}$$

$$r_1^2 = \begin{bmatrix} 1 \\ 2.2 \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \end{bmatrix}^t \Sigma^{-1} \begin{bmatrix} 1 \\ 2.2 \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 2.2 \end{bmatrix}^T \begin{bmatrix} 0.95 & -0.15 \\ -0.15 & 0.55 \end{bmatrix} \begin{bmatrix} 1 \\ 2.2 \end{bmatrix}$$

$$\begin{bmatrix} 1 \\ 2.2 \end{bmatrix}^T \begin{bmatrix} 0.95 - 0.15 \times 2.2 \\ -0.15 + 0.55 \times 2.2 \end{bmatrix} = \begin{bmatrix} 1 \\ 2.2 \end{bmatrix}^T \begin{bmatrix} 0.62 \\ 1.06 \end{bmatrix} = 0.62 + 2.2 \cdot 1.06 = 2.95$$

$$r_2^2 = \begin{bmatrix} 1 - 3 \\ 2.2 - 3 \end{bmatrix}^t \Sigma^{-1} \begin{bmatrix} 1 - 3 \\ 2.2 - 3 \end{bmatrix} = \begin{bmatrix} -2 \\ -0.8 \end{bmatrix}^t \begin{bmatrix} 0.95 & -0.15 \\ -0.15 & 0.55 \end{bmatrix} \begin{bmatrix} -2 \\ -0.8 \end{bmatrix} = \begin{bmatrix} -2 \\ -0.8 \end{bmatrix}^t \begin{bmatrix} -1.90 + 0.12 \\ 0.30 - 0.44 \end{bmatrix}$$

$$\begin{bmatrix} -2 \\ -0.8 \end{bmatrix}^t \begin{bmatrix} -1.78 \\ -0.14 \end{bmatrix} = 3.56 + 0.11 = 3.67$$

OR SOLVE IT BY LDFs.

$r_2^2 > r_1^2$. It belongs to class 1

4. (20 pts)

Let x have an exponential density

$$p(x|\theta) = \begin{cases} \theta e^{-\theta x} & x \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

(a) Plot $p(x|\theta)$ versus x for $\theta = 1$. Plot $p(x|\theta)$ versus θ, $(0 \leq \theta \leq 5)$, for $x = 2$.

(b) Suppose that n samples $x_1, \ldots, x_n$ are drawn independently according to $p(x|\theta)$. Show that the maximum-likelihood estimate for θ is given by

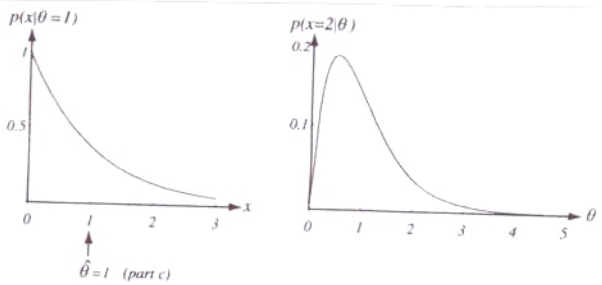
$$\hat{\theta} = \frac{1}{\frac{1}{n}\sum_{k=1}^{n} x_k}.$$

(c) On your graph generated with $\theta = 1$ in part (a), mark the maximum-likelihood estimate $\hat{\theta}$ for large n.

Our exponential function is:

$$p(x|\theta) = \begin{cases} \theta e^{-\theta x} & x \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

(a) SEE FIGURE. Note that $p(x = 2|\theta)$ is not maximized when $\theta = 2$ but instead for a value less than 1.0.



(b) The log-likelihood function is

$$l(\theta) = \sum_{k=1}^{n} \ln p(x_k|\theta) = \sum_{k=1}^{n} [\ln \theta - \theta x_k] = n \ln \theta - \theta \sum_{k=1}^{n} x_k.$$

We solve $\nabla_\theta l(\theta) = 0$ to find $\hat{\theta}$ as

$$\begin{aligned} \nabla_\theta l(\theta) &= \frac{\partial}{\partial \theta}\left[n \ln \theta - \theta \sum_{k=1}^{n} x_k \right] \\ &= \frac{n}{\theta} - \sum_{k=1}^{n} x_k = 0. \end{aligned}$$

Thus the maximum-likelihood solution is

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

(c) Here we approximate the mean

$$\frac{1}{n}\sum_{k=1}^{n} x_n$$

by the integral

$$\int_0^\infty x p(x) dx,$$

which is valid in the large n limit. Noting that

$$\int_0^\infty x e^{-x} dx = 1,$$

we put these results together and see that $\hat{\theta} = 1$, as shown on the figure in part (a).

5. (a) (5pts) What is the Fisher linear discriminant method?

b) Given the 2-d data for two classes:
$\omega_1 = \{(1,1),(1,2),(1,4),(2,1),(3,1),(3,3)\}$ and
$\omega_2 = \{(2,2),(3,2),(3,4),(5,1),(5,4),(5,5)\}$ as shown in the figure:

i. (10 pts) Determine the optimal projection line in a single dimension.

ii. (5 pts) Show the mapping of the points to the line as well as the Bayes discriminant assuming a suitable distribution.

5) Let v be the direction of the projection line, then the Fisher linear discriminant method finds the best v, as follows

$$v = S_w^{-1}(\mu_1 - \mu_2) \qquad S_w = S_1 + S_2$$

$$S_i = \sum_{x \in D_i} (x - \mu_i)(x - \mu_i)^t \qquad i = 1, 2$$

$$\mu_1 = \begin{bmatrix} \frac{11}{6} \\ 2 \end{bmatrix} \qquad \mu_2 = \begin{bmatrix} \frac{23}{6} \\ 3 \end{bmatrix}$$

$$x - \mu_1 = \begin{bmatrix} -\frac{5}{6} & -\frac{5}{6} & -\frac{5}{6} & \frac{1}{6} & \frac{7}{6} & \frac{7}{6} \\ -1 & 0 & 2 & -1 & -1 & 1 \end{bmatrix} \qquad x - \mu_2 = \begin{bmatrix} -\frac{11}{6} & -\frac{5}{6} & -\frac{5}{6} & \frac{7}{6} & \frac{7}{6} & \frac{7}{6} \\ -1 & -1 & 1 & -2 & 1 & 2 \end{bmatrix}$$

$$S_1 = \begin{bmatrix} \frac{25 + 25 + 25 + 1 + 49 + 49}{36} & \frac{5 + 0 - 10 - 1 - 7 + 7}{1} \\ \frac{5 + 0 - 10 - 1 - 7 + 7}{6} & \frac{1 + 0 + 4 + 1 + 1 + 1}{1} \end{bmatrix} = \begin{bmatrix} \frac{29}{36} & -1 \\ -1 & 8 \end{bmatrix}$$

$$S_2 = \begin{bmatrix} \frac{121 + 25 + 25 + 49 + 49 + 49}{36} & \frac{11 + 5 - 5 - 14 + 7 + 14}{6} \\ \frac{11 + 5 - 5 - 14 + 7 + 14}{6} & 1 + 1 + 1 + 4 + 1 + 4 \end{bmatrix} = \begin{bmatrix} \frac{53}{6} & 3 \\ 3 & 12 \end{bmatrix}$$

then $$S_w = S_1 + S_2 = \begin{bmatrix} \frac{41}{3} & 2 \\ 2 & 20 \end{bmatrix}$$

$$S_w^{-1} = \frac{1}{|S_w|} \begin{bmatrix} 20 & -2 \\ -2 & \frac{41}{3} \end{bmatrix} = \frac{1}{\frac{808}{3}} \begin{bmatrix} 20 & -2 \\ -2 & \frac{41}{3} \end{bmatrix} = \begin{bmatrix} \frac{15}{202} & -\frac{3}{404} \\ -\frac{3}{204} & \frac{41}{808} \end{bmatrix}$$

Finally we have: $v = S_w^{-1}(\mu_1 - \mu_2)$

$$V = \begin{bmatrix} \frac{15}{202} & -\frac{3}{404} \\ -\frac{3}{404} & \frac{41}{808} \end{bmatrix} \begin{bmatrix} -2 \\ -1 \end{bmatrix} = \begin{bmatrix} -\frac{57}{404} \\ -\frac{29}{808} \end{bmatrix} \simeq \begin{bmatrix} -0.1411 \\ -0.0359 \end{bmatrix}$$

ii) The samples are mapped by $x' = V^t x$ and we get

$$V_1^t = [-0.1770, -0.2129, -0.2847, -0.3181, -0.4592, -0.5309]$$

$$V_2^t = [-0.3540, -0.4950, -0.5668, -0.7413, -0.8490, -0.8849]$$

and we compute the mean and the standard deviation as

$$\mu_1 = 0.3304 \qquad \sigma_1 = 0.1388$$

$$\mu_2 = 0.6485 \qquad \sigma_2 = 0.2106$$

If we assume both $p(x/\omega_1)$ and $p(x/\omega_2)$ have a Gaussian distribution, then the Bayes decision rule will be

Decide ω_1 if $p(x/\omega_i) P(\omega_1) > p(x/\omega_2) P(\omega_2)$, otherwise decide ω_2

where $$p(x/\omega_i) = \frac{1}{\sqrt{2\pi}\,\sigma_i} \exp\left[-\frac{1}{2}\left(\frac{x-\mu_i}{\sigma_i}\right)^2\right]$$

If we assume the prior probabilities are equal, i.e. $P(\omega_1') = P(\omega_2') = 0.5$, then the threshold will be about -0.4933. That is we decide ω_1 if $V^t x > -0.4933$ otherwise decide ω_2.

Midterm Exam Cheat Sheet

Expected Loss

$$R(a_i / \mathbf{x}) = \sum_{j=1}^{c} \lambda(a_i / \omega_j) P(\omega_j / \mathbf{x})$$

Multivariate Gaussian

$$N(\mu,\Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} exp[-\frac{1}{2}(\mathbf{x}-\mu)^t \Sigma^{-1}(\mathbf{x}-\mu)]$$

Discriminant and Decision Boundary (Case 1: $\Sigma_i=\sigma^2$)

$$g_i(\mathbf{x}) = \mathbf{w}_i^t \mathbf{x} + w_{i0}$$

where $\mathbf{w}_1 = \frac{1}{\sigma^2}\mu_i$, and $w_{i0} = -\frac{1}{2\sigma^2}\mu_i^t \mu_i + \ln P(\omega_i)$

- Decision boundary is determined by hyperplanes; setting $g_i(\mathbf{x}) = g_j(\mathbf{x})$:

$$\mathbf{w}^t(\mathbf{x}-\mathbf{x_0}) = 0$$

where $\mathbf{w} = \mu_i - \mu_j$, and $\mathbf{x}_0 = \frac{1}{2}(\mu_i + \mu_j) - \frac{\sigma^2}{\|\mu_i - \mu_j\|^2} \ln \frac{P(\omega_i}{P(\omega_j}(\mu_i - \mu_j)$

Discriminant and Decision Boundary (Case 2: $\Sigma_i=\Sigma$)

$$g_i(\mathbf{x}) = \mathbf{w}_i^t\mathbf{x} + w_{i0}$$
(linear discriminant)

where $\mathbf{w}_i = \Sigma^{-1}\mu_i$, and $w_{i0} = -\frac{1}{2}\mu_i^t\Sigma^{-1}\mu_i + \ln P(\omega_i)$

- Decision boundary is determined by hyperplanes; setting $g_i(\mathbf{x}) = g_j(\mathbf{x})$:

$$\mathbf{w}^t(\mathbf{x} - \mathbf{x_0}) = 0$$

where $\mathbf{w} = \Sigma^{-1}(\mu_i - \mu_j)$ and $\mathbf{x}_0 = \frac{1}{2}(\mu_i + \mu_j) - \frac{ln[P(\omega_i)/P(\omega_j)]}{(\mathbf{x}-\mu_i)^t\Sigma^{-1}(\mathbf{x}-\mu_i)}(\mu_i - \mu_j)$

Discriminant and Decision Boundary (Case 3: Σ_i)

$$g_i(\mathbf{x}) = \mathbf{x}^t\mathbf{W}_i\mathbf{x} + \mathbf{w}_i\mathbf{x} + w_{i0}$$
(quadratic discriminant)

where $\mathbf{W}_i = -\frac{1}{2}\Sigma_i^{-1}$, $\mathbf{w}_i = \Sigma_i^{-1}\mu_i$, and $w_{i0} = -\frac{1}{2}\mu_i^t\Sigma^{-1}\mu_i - \frac{1}{2}ln|\Sigma_i| + \ln P(\omega_i)$

- Decision boundary is determined by superquadrics; setting $g_i(\mathbf{x}) = g_j(\mathbf{x})$

Maximum Likelihood Estimation

The Gaussian Case: Unknown μ

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

The Gaussian Case: Unknown μ and Σ

$$\hat{\mu} = \frac{1}{n}\sum_{k=1}^{n} x_k \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{k=1}^{n} (x_k - \hat{\mu})^2$$

$$\hat{\Sigma} = \frac{1}{n}\sum_{k=1}^{n} (x_k - \hat{\mu})(x_k - \hat{\mu})^t$$

Linear Discriminant Functions

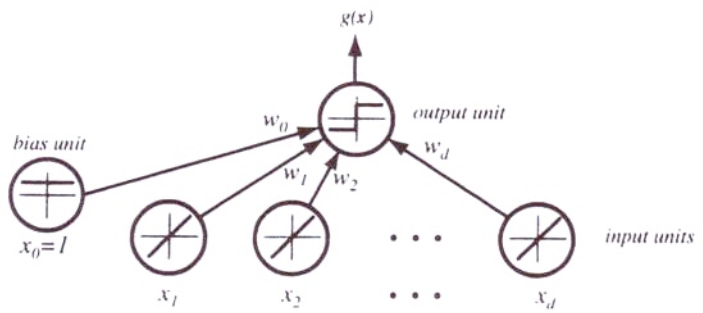


FIGURE 5.1. A simple linear classifier having d input units, each corresponding to the values of the components of an input vector. Each input feature value x_i is multiplied by its corresponding weight w_i; the effective input at the output unit is the sum all these products, $\sum w_i x_i$. We show in each unit its effective input-output function. Thus each of the d input units is linear, emitting exactly the value of its corresponding feature value. The single bias unit unit always emits the constant value 1.0.

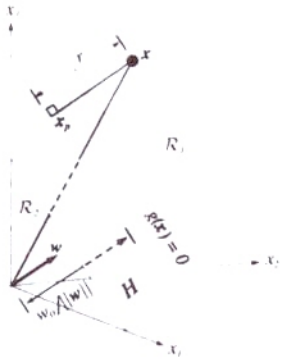


FIGURE 5.2. The linear decision boundary H, where $g(\mathbf{x}) = \mathbf{w}^t\mathbf{x}+w_0 = 0$, separates the feature space into two half-spaces $\mathcal{R}_1$ (where $g(\mathbf{x}) > 0$) and $\mathcal{R}_2$ (where $g(\mathbf{x}) < 0$). From: Richard O. Duda, Peter E. Hart, and David G. Stork, Pattern Classification. Copyright © 2001 by John Wiley & Sons, Inc.

Minimum Squared Error

$$Ya \simeq b \qquad \begin{bmatrix} y_1^{(0)} & y_1^{(1)} & \cdots & y_1^{(d)} \\ y_2^{(0)} & y_2^{(1)} & \cdots & y_2^{(d)} \\ \vdots & & & \\ y_n^{(0)} & y_n^{(1)} & \cdots & y_n^{(d)} \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_d \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} \qquad a = (Y^tY)^{-1}Y^tb$$

LDF : "Widrow-Hoff" Procedure $\quad a^{(k+1)} = a^{(k)} - \eta^{(k)} y_i \left(y_i^t a^{(k)} - b_i\right)$

Fisher Linear Discriminant

Maximize : $J(v) = \dfrac{(\tilde{\mu}_1 - \tilde{\mu}_2)^2}{\tilde{s}_1^2 + \tilde{s}_2^2}$

$v = S_w^{-1}(\mu_1 - \mu_2)$

$S_w = S_1 + S_2$

$$S_1 = \sum_{x_i \in \text{class 1}} (x_i - \mu_1)(x_i - \mu_1)^t$$

$$S_2 = \sum_{x_i \in \text{class 2}} (x_i - \mu_2)(x_i - \mu_2)^t$$

Matrix Inversion (2x2)

- For a 2×2 matrix,

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

the matrix inverse is

$$A^{-1} = \frac{1}{|A|}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix} = \frac{1}{ad - bc}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$