1. Should be 4 components for 4 parameters.

Not orthogonal, eg $0.5 \times 0.2 + 0.5 \times 0.1 + 0 \times -0.2$
$+ 0.2 \times 0.4$
$= 0.1 + 0.05 + 0.08$
$= 0.23 \neq 0.$

$0.5^2 + 0.5^2 + 0.4^2 = 0.25 + 0.25 + 0.16 = 0.66 \neq 1.$
So not a unit vector.

2. $P(D|\underline{\theta}) = \prod_{k=1}^{n} P(\underline{x}_k|\underline{\theta})$ ← Vary $\underline{\theta}$ to maximize this

$\underline{\theta} = \begin{bmatrix} \hat{\mu} \\ \hat{\sigma}^2 \end{bmatrix}$

take logs

$l(\underline{\theta}) = \sum_{k=1}^{n} \ln P(\underline{x}_k|\underline{\theta})$

$\frac{\partial l}{\partial \theta_p} = \sum_{k=1}^{n} \frac{\partial}{\partial \theta_p} \ln P(\underline{x}_k|\theta_p) = 0$ for maximum.

$$\ln p(x_k|\underline{\theta}) = -\frac{1}{2}\ln(2\pi\theta_2) - \frac{1}{2\theta_2}(x_k-\theta_1)^2$$

So
$$\frac{\partial l}{\partial \underline{\theta}} = \sum_{k=1}^{n}\begin{bmatrix} \frac{1}{\theta_2}(x_k-\theta_1) \\ -\frac{1}{2\theta_2} + \frac{(x_k-\theta_1)^2}{2\theta_2^2}\end{bmatrix}$$

First row:

$$\sum \frac{1}{\theta_2}(x_k-\theta_1) = 0 \qquad \text{so } \sum x_k = n\theta_1$$

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

ie estimate of mean = mean of samples

Second row:

$$0 = \sum -\frac{1}{2\theta_2} + \frac{(x_k-\theta_1)^2}{2\theta_2^2}$$

$$\sum -\theta_2 + (x_k-\theta_1)^2$$

$$\theta_2 = \frac{1}{n}\sum (x_k-\theta_1)^2$$

or $\hat{\sigma}^2 = \frac{1}{n}\sum_{k=1}^{n}(x_k-\mu)^2$ ie estimate of σ = σ of samples.

3 a)

$$\begin{pmatrix} 4 & -0.7 \\ -0.7 & 1 \end{pmatrix}$$

b)

$$\begin{vmatrix} 4-\lambda & -0.7 \\ -0.7 & 1-\lambda \end{vmatrix} = (4-\lambda)(1-\lambda) - 0.49$$

$$= \lambda^2 - 5\lambda + 3.51$$

$$\lambda = \frac{5 \pm \sqrt{25 - 4 \times 3.51 \times 1}}{2}$$

$$\lambda_1 = 4.155$$

$$\begin{pmatrix} 4 & -0.7 \\ -0.7 & 1 \end{pmatrix} \begin{pmatrix} P_1 \\ P_2 \end{pmatrix} = \lambda \begin{pmatrix} P_1 \\ P_2 \end{pmatrix}$$

set $P_2 = 1$

$$-0.7P_1 + 1 = \lambda$$

$$P_1 = \frac{\lambda - 1}{-0.7} = \frac{3.155}{-0.7} = -4.508$$

so $P = \begin{pmatrix} -4.508 \\ 1 \end{pmatrix}$ $\quad |P| = 4.617$

$$\hat{P} = \begin{pmatrix} -0.976 \\ 0.217 \end{pmatrix}$$

4 a)

$$\int_{-\infty}^{\infty} \phi \, dx = \int_{-1}^{1} \frac{3}{4}(1 - x^2) \, dx$$

$$= \frac{3}{4}\left[x - \frac{x^3}{3}\right]_{-1}^{1}$$

$$= \frac{3}{2}\left(1 - \frac{1}{3}\right) = 1$$

$\therefore$ integrates to 1.

$x^2 \leq 1$ in range, so $1 - x^2 \geq 0$.

So $\phi(x) \geq 0$

$\therefore$ Valid probability distribution

$\therefore$ Valid parzen window.

b)

$$p(x) = \begin{cases} \frac{3}{4}\left[1 - (x+1)^2\right] & \text{if } -2 < x < 0 \\ 0 & \text{otherwise} \end{cases}$$

c)

$$P(x \mid \omega_1) = 1 - \cos \pi x$$

$$P(x \mid \omega_2) = 1$$

$$P(\omega_1 | x) = \frac{P(\cancel{}x | \omega_1)\, P(\omega_1)}{P(x)}$$

$$P(x) = P(x | \omega_1) P(\omega_1) + P(x | \omega_2) P(\omega_2)$$

$$= (1 - \cos \pi x) 0.6 + 1 \times 0.4$$

$$= 1 - 0.6 \cos \pi x$$

$$P(\omega_1 | x) = \frac{(1 - \cos \pi x)\, 0.6}{1 - 0.6 \cos \pi x}$$

5. a) 2.1 → 0

1.8 → 1

1.2 → 1

Classify as 1.

b) More points means reduced overfitting
Reduced variance. Potentially increased bias.

c) None.

6. a) $w = \begin{pmatrix} 0 \\ -1 \end{pmatrix} - \begin{pmatrix} 2 \\ 1 \end{pmatrix} = \begin{pmatrix} -2 \\ -2 \end{pmatrix}$

$g(\underline{x}) = 0$ when $\underline{x} = \cancel{\begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}} \begin{pmatrix} 1 \\ 0 \end{pmatrix}$

so $0 = \begin{pmatrix} -2 & -2 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} + w_0$

$w_0 = 2$

b) $g\begin{pmatrix} 2 \\ 1 \end{pmatrix} = \begin{pmatrix} -2 & -2 \end{pmatrix} \begin{pmatrix} 2 \\ 1 \end{pmatrix} + 2$

$= -6 + 2 = -4$ so margins range from -4 to $+4$

$g\begin{pmatrix} 0.5 \\ 1 \end{pmatrix} = \begin{pmatrix} -2 & -2 \end{pmatrix} \begin{pmatrix} 0.5 \\ 1 \end{pmatrix} + 2$

$= -3 + 2 = -1$

$\therefore \epsilon_i = 1.25$ if classified as 1.

$\epsilon_i = 0.75$ if classified as 0.

7 a/ Type is a non-metric parameter, ie has no 'closeness' associated with it.

Decision tree-type methods work with non-metric problems.

b/ Other methods may be more suitable. The type parameter has effectively been removed so the data is metric.

8 a) Unsupervised learning

b) Supervised

c) Supervised

9 Impurity $= 1 - P_1^2 - P_2^2 - P_3^2$

$= 1 - 0.1^2 - 0.6^2 - 0.3^2$

$= 0.54$

10a) 0.106 0.066 0.288 0.609 0.548

diff: −0.04 0.162 0.381 −0.061

0.667 0.657 0.772

diff: 0.118 −0.010 0.115

So get step after 3rd ~~s~~ sample.

b) i) ~~M~~ Too sensitive to noise.

ii) ~~Free~~ Miss potential

11 a/ x_1^2, x_2^2, x_1x_2

b/ $C = \sum \left(y_i - \left(\beta_1 + \beta_2 x_{1i} + \beta_3 x_{2i}\right)\right)^2$

$r_i = y_i - (\beta_1 + \beta_2 x_{1i} + \beta_3 x_{2i})$

$$\frac{\partial C}{\partial \beta_1} = \sum 2r_i(-1) = -2\sum r_i$$

$$\frac{\partial C}{\partial \beta_2} = \sum 2r_i(-x_{1i}) = -2\sum r_i x_{1i}$$

$$\frac{\partial C}{\partial \beta_3} = -2\sum r_i x_{2i}$$

x_1	x_2	y	y'	r_i
5.3	8.3	6.2	-3.3	9.5
2.1	2.8	5.6	-0.1	5.7
3.9	5.5	7.2	-1.9	9.1
1.8	2.3	3.4	0.2	3.2

$\frac{\partial C}{\partial \beta_1} = -2 \times 27.5 = -55$

$\frac{\partial C}{\partial \beta_2} = -207.14$

$\frac{\partial C}{\partial \beta_3} = -304.44$

$\beta_1 \rightarrow 2.055$

$\beta_2 \rightarrow -0.792$

$\beta_3 \rightarrow 0.304$

12 a/ Reduce softness

b/ Reduce number of layers

13 $(-1x_1)0.1 + 0.8x_1 \times 0.8$

$+ 0.8x_2 \times 0.1 - 0.7x_2 \times 0.8$

$+ 1.5 \times 0.1 + 0.2 \times 0.8 + 0.3$

$= 0.54x_1 - 0.48x_2 + 0.61$

$g(x) = \underline{w}^t \underline{x} + w_0$

$w_0 = 0.61$

$$\underline{w} = \begin{pmatrix} 0.54 \\ -0.48 \end{pmatrix}$$