Foundations of Machine Learning
Classification: decisions and discriminants

Srinandan (“Sri”) Dasmahapatra

2-class classification: determine which class a given input $X=x$ belongs to

2-class classification: determine which class a given input $X=x$ belongs to

decision boundary at equal posterior probability for either class, which are:

$$P(C = +|X = x) = \frac{P(X = x|C = +)P(C = +)}{P(X = x|C = +)P(C = +) + P(X = x|C = -)P(C = -)}$$

$$P(C = -|X = x) = \frac{P(X = x|C = -)P(C = -)}{P(X = x|C = +)P(C = +) + P(X = x|C = -)P(C = -)}$$

2-class classification: determine which class a given input $X{=}x$ belongs to

decision boundary at equal posterior probability for either class, which are:

$$P(C=+|X=x) = \frac{P(X=x|C=+)P(C=+)}{P(X=x|C=+)P(C=+)+P(X=x|C=-)P(C=-)}$$

$$P(C=-|X=x) = \frac{P(X=x|C=-)P(C=-)}{P(X=x|C=+)P(C=+)+P(X=x|C=-)P(C=-)}$$

Simplify the notation: introduce $g_{\pm}(x)$

2-class classification: determine which class a given input $X=x$ belongs to

decision boundary at equal posterior probability for either class, which are:

$$P(C=+|X=x) = \frac{P(X=x|C=+)P(C=+)}{P(X=x|C=+)P(C=+)+P(X=x|C=-)P(C=-)}$$

$$P(C=-|X=x) = \frac{P(X=x|C=-)P(C=-)}{P(X=x|C=+)P(C=+)+P(X=x|C=-)P(C=-)}$$

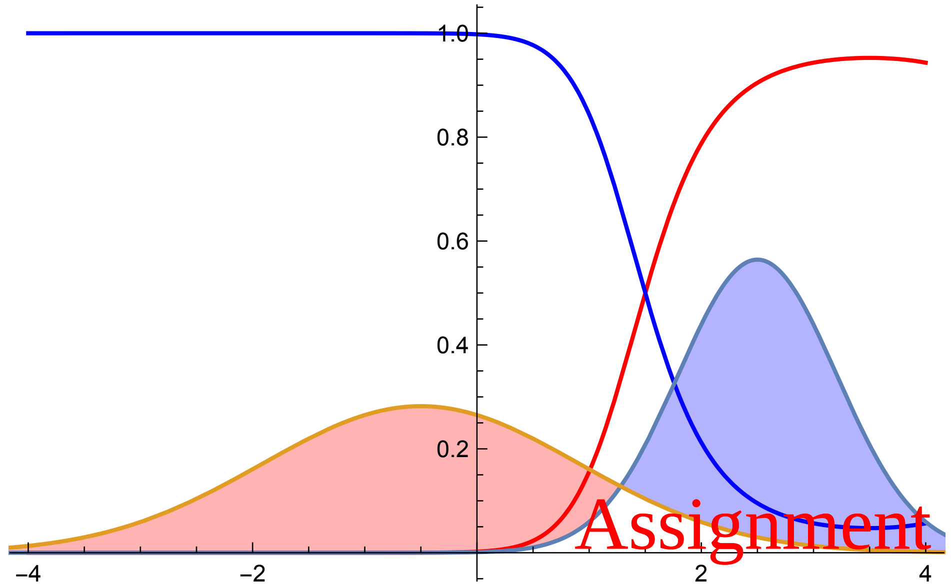
Simplify the notation: introduce $g_\pm(x)$

$$P(C=+|X=x) = \frac{e^{g_+(x)}}{e^{g_+(x)}+e^{g_-(x)}} = \frac{\exp\left[g_+(x)-g_-(x)\right]}{1+\exp\left[g_+(x)-g_-(x)\right]}$$

$$P(C=-|X=x) = \frac{e^{g_-(x)}}{e^{g_+(x)}+e^{g_-(x)}} = \frac{1}{1+\exp\left[g_+(x)-g_-(x)\right]}$$

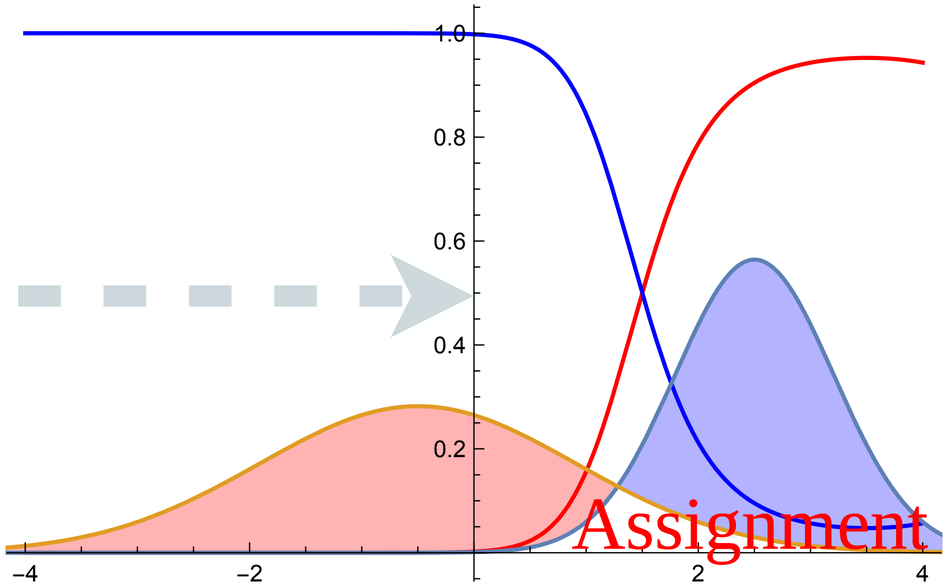
Decision boundary: determine which class a given input belongs to

Decision boundary: determine which class a given input belongs to



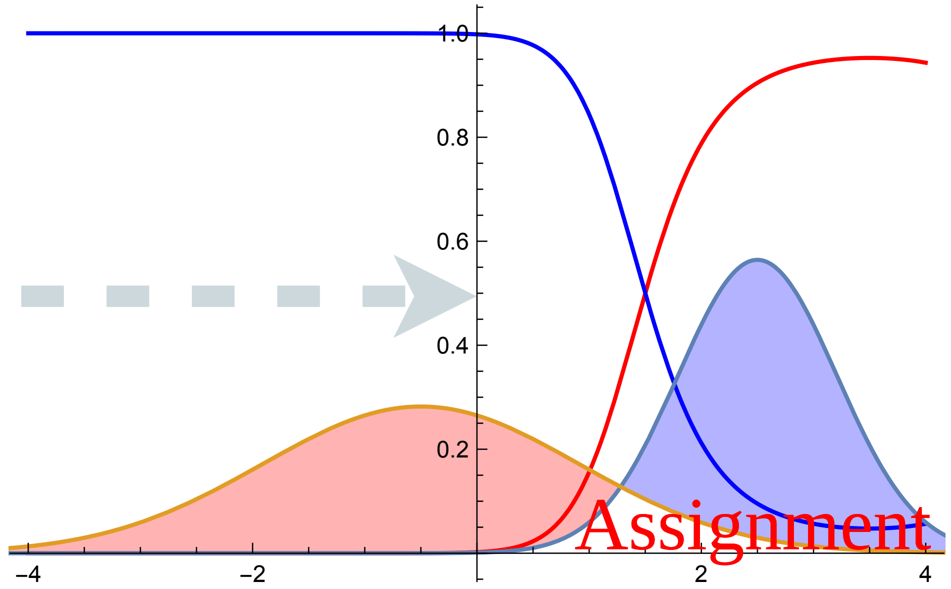
decision boundary at equal posterior probability (plotted) for either class

Decision boundary: determine which class a given input belongs to



decision boundary at equal posterior probability (plotted) for either class

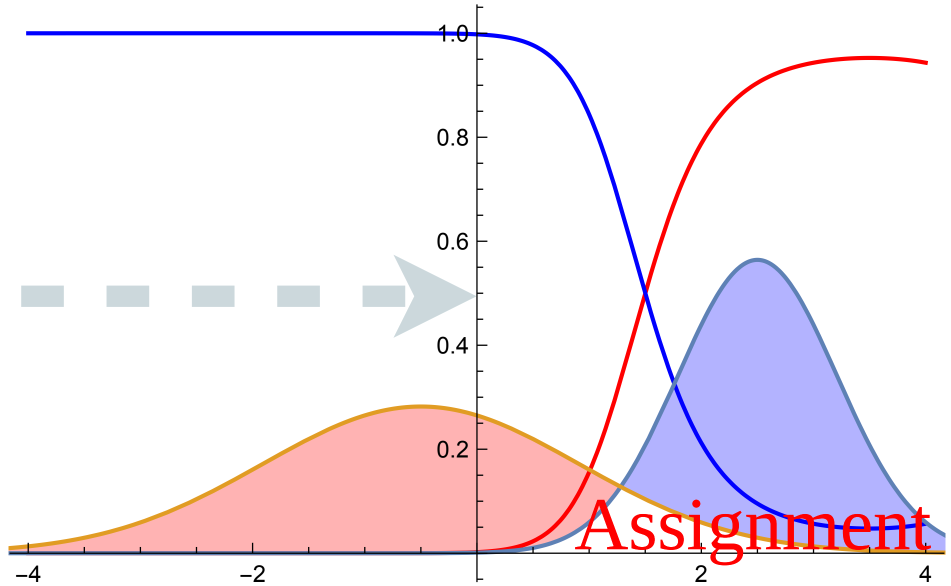
Decision boundary: determine which class a given input belongs to



decision boundary at equal posterior probability (plotted) for either class

$$\text{odds} = 1, \ \log(\text{odds}) = 0$$

Decision boundary: determine which class a given input belongs to

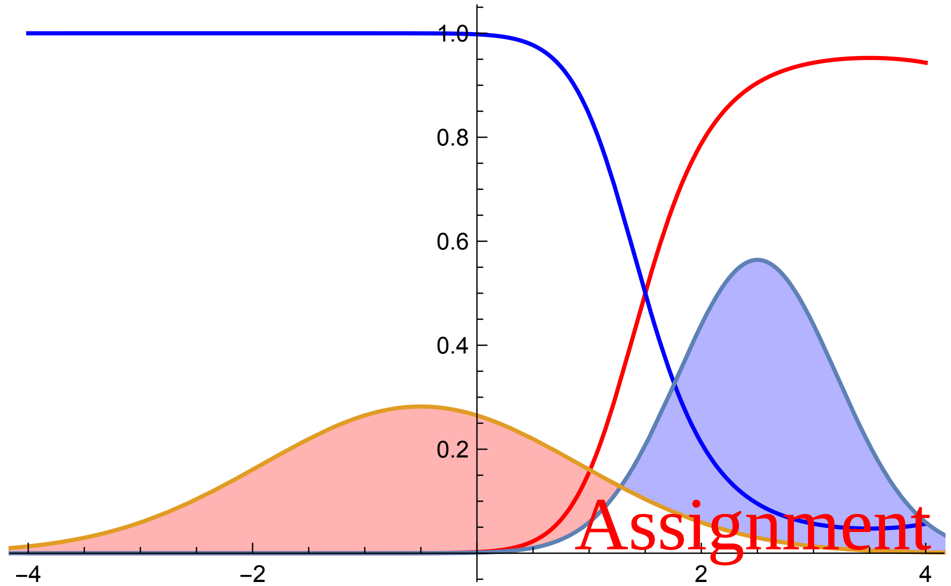


decision boundary at equal posterior probability (plotted) for either class

$$\text{odds} = 1, \ \log(\text{odds}) = 0$$

$$\text{odds} = \frac{P(C=+|X=x)}{P(C=-|X=x)} = \exp\left[g_{+}(x) - g_{-}(x)\right]$$

Decision boundary: determine which class a given input belongs to

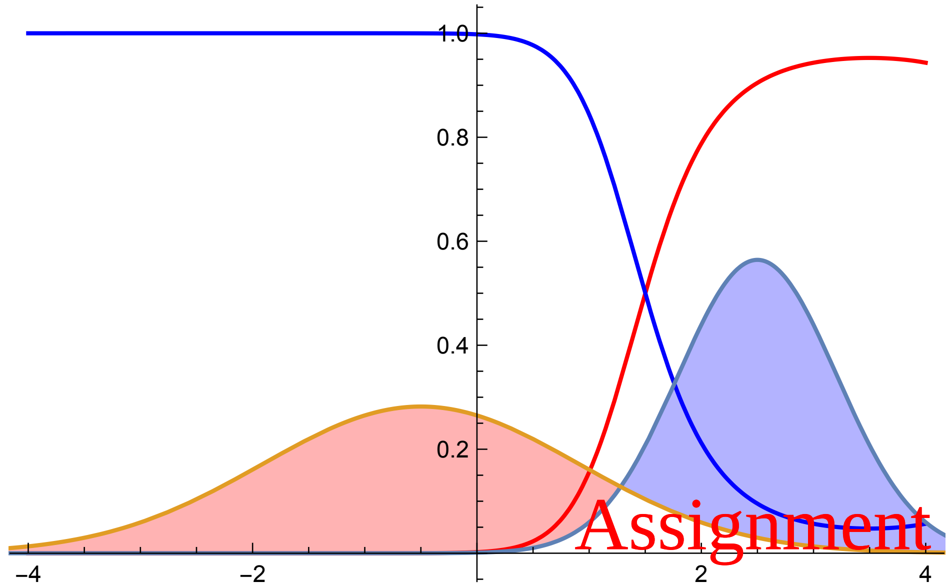


decision boundary at equal posterior probability (plotted) for either class

$$\text{odds} = 1, \ \log(\text{odds}) = 0$$

$$\text{odds} = \frac{P(C = +|X = x)}{P(C = -|X = x)} = \exp\left[g_+(x) - g_-(x)\right]$$

Decision boundary: determine which class a given input belongs to



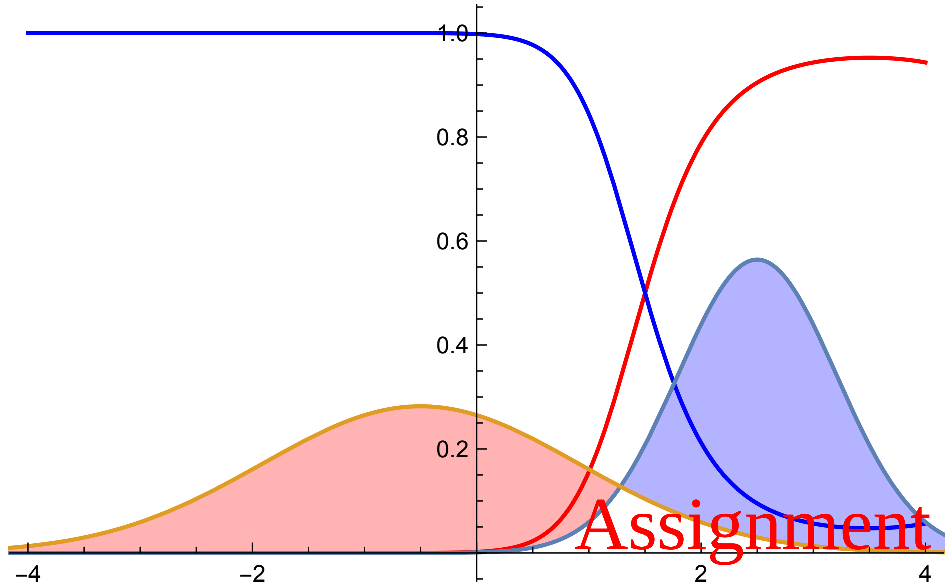
decision boundary at equal posterior probability (plotted) for either class

$$\text{odds} = 1, \ \log(\text{odds}) = 0$$

$$\text{odds} = \frac{P(C = +|X = x)}{P(C = -|X = x)} = \exp\left[g_+(x) - g_-(x)\right]$$

$$\text{odds} = \frac{P(C = +|X = x)}{P(C = -|X = x)} = \frac{P(X = x|C = +)P(C = +)}{P(X = x|C = -)P(C = -)}$$

Decision boundary: determine which class a given input belongs to



decision boundary at equal posterior probability (plotted) for either class

$$\text{odds} = 1, \; \log(\text{odds}) = 0$$

$$\text{odds} = \frac{P(C=+|X=x)}{P(C=-|X=x)} = \exp\left[g_+(x) - g_-(x)\right]$$

$$\text{odds} = \frac{P(C=+|X=x)}{P(C=-|X=x)} = \frac{P(X=x|C=+)P(C=+)}{P(X=x|C=-)P(C=-)}$$

$$\text{log-odds} = \ln\left(\frac{P(X=x|C=+)}{P(X=x|C=-)}\right) + \ln\left(\frac{P(C=+)}{P(C=-)}\right)$$

data dependent (likelihood)

data independent (prior)

Decision boundaries from probability distributions: zero log-odds: $0 = g_+(\mathbf{x}) - g_-(\mathbf{x}) = f(\mathbf{x}; \theta) + \theta_0$

Decision boundaries from probability distributions: zero log-odds: $0 = g_+(\mathbf{x}) - g_-(\mathbf{x}) = f(\mathbf{x}; \theta) + \theta_0$

- If we choose linear functions:
$g_{+}(\mathbf{x}) = w_{+0} + \mathbf{w}_{+}^{\top}\mathbf{x}, g_{-}(\mathbf{x}) = w_{-0} + \mathbf{w}_{-}^{\top}\mathbf{x}$

Decision boundaries from probability distributions: zero log-odds: $0 = g_+(\mathbf{x}) - g_-(\mathbf{x}) = f(\mathbf{x};\theta) + \theta_0$

- If we choose linear functions:
  $g_+(\mathbf{x}) = w_{+0} + \mathbf{w}_+^\top \mathbf{x}, g_-(\mathbf{x}) = w_{-0} + \mathbf{w}_-^\top \mathbf{x}$
- Then $f(\mathbf{x};\theta) = \theta_0 + \theta^\top \mathbf{x}$ is a linear decision boundary

Decision boundaries from probability distributions: zero log-odds: $0 = g_+(\mathbf{x}) - g_-(\mathbf{x}) = f(\mathbf{x};\theta) + \theta_0$

- If we choose linear functions: $g_+(\mathbf{x}) = w_{+0} + \mathbf{w}_+^\top \mathbf{x},\ g_-(\mathbf{x}) = w_{-0} + \mathbf{w}_-^\top \mathbf{x}$
- Then $f(\mathbf{x};\theta) = \theta_0 + \theta^\top \mathbf{x}$ is a linear decision boundary
- where $\theta_0 = w_{+0} - w_{-0}$, and $\theta = \mathbf{w}_+ - \mathbf{w}_-$

Decision boundaries from probability distributions: zero log-odds: $0 = g_+(\mathbf{x}) - g_-(\mathbf{x}) = f(\mathbf{x}; \theta) + \theta_0$

- If we choose linear functions:
  $g_{+}(\mathbf{x}) = w_{+0} + \mathbf{w}_{+}^{\top}\mathbf{x}, g_{-}(\mathbf{x}) = w_{-0} + \mathbf{w}_{-}^{\top}\mathbf{x}$
- Then $f(\mathbf{x}; \theta) = \theta_0 + \theta^{\top}\mathbf{x}$ is a linear decision boundary
- where $\theta_0 = w_{+0} - w_{-0}$, and $\theta = \mathbf{w}_{+} - \mathbf{w}_{-}$
- This is 2-class softmax (logistic) regression

Reminder: Logistic (2-class)/softmax (K-class) regression

Reminder: Logistic (2-class)/softmax (K-class) regression

- Choose suitable features Φ to compute scores/activations:
  $s(A_i) := \text{score}(\text{class } A_i) = f_i(\Phi(\text{data}))$

Reminder: Logistic (2-class)/softmax (K-class) regression

- Choose suitable features Φ to compute scores/activations:
  $s(A_i) :=$ score(class A_i) $= f_i(\Phi(\text{data}))$

- Recall,
  softmax$(s(A_1), \ldots, s(A_K)) = \ln[\exp(s(A_1)) + \cdots + \exp(s(A_K))]$

Reminder: Logistic (2-class)/softmax (K-class) regression

- Choose suitable features Φ to compute scores/activations:
  $s(A_i) :=$ score(class A_i) $= f_i(\Phi(\text{data}))$

- Recall,
  softmax$(s(A_1), \ldots, s(A_K)) = \ln[\exp(s(A_1)) + \cdots + \exp(s(A_K))]$

- $$P(\text{class } A_i \,|\, \text{data}) = \frac{\exp(s(A_i))}{\exp \text{softmax}(s(A_1), \ldots, s(A_K))} = \frac{\exp(s(A_i))}{\sum_{i=1}^{K} \exp(s(A_i))}$$

Reminder: Logistic (2-class)/softmax (K-class) regression

- Choose suitable features Φ to compute scores/activations:
  $s(A_i) :=$ score(class A_i) $= f_i(\Phi(\text{data}))$

- Recall,

  $\text{softmax}(s(A_1), \ldots, s(A_K)) = \ln[\exp(s(A_1)) + \cdots + \exp(s(A_K))]$

- $$P(\text{class } A_i \mid \text{data}) = \frac{\exp(s(A_i))}{\exp \text{softmax}(s(A_1), \ldots, s(A_K))} = \frac{\exp(s(A_i))}{\sum_{i=1}^{K} \exp(s(A_i))}$$

- $f_i(\text{data } x) = \mathbf{w}_i^\top \Phi(x) = \mathbf{w}_i^\top x$

Reminder: Logistic (2-class)/softmax (K-class) regression

- Choose suitable features Φ to compute scores/activations:
  $s(A_i) :=$ score(class A_i) $= f_i(\Phi(\text{data}))$

- Recall,

  $\text{softmax}(s(A_1), \ldots, s(A_K)) = \ln[\exp(s(A_1)) + \cdots + \exp(s(A_K))]$

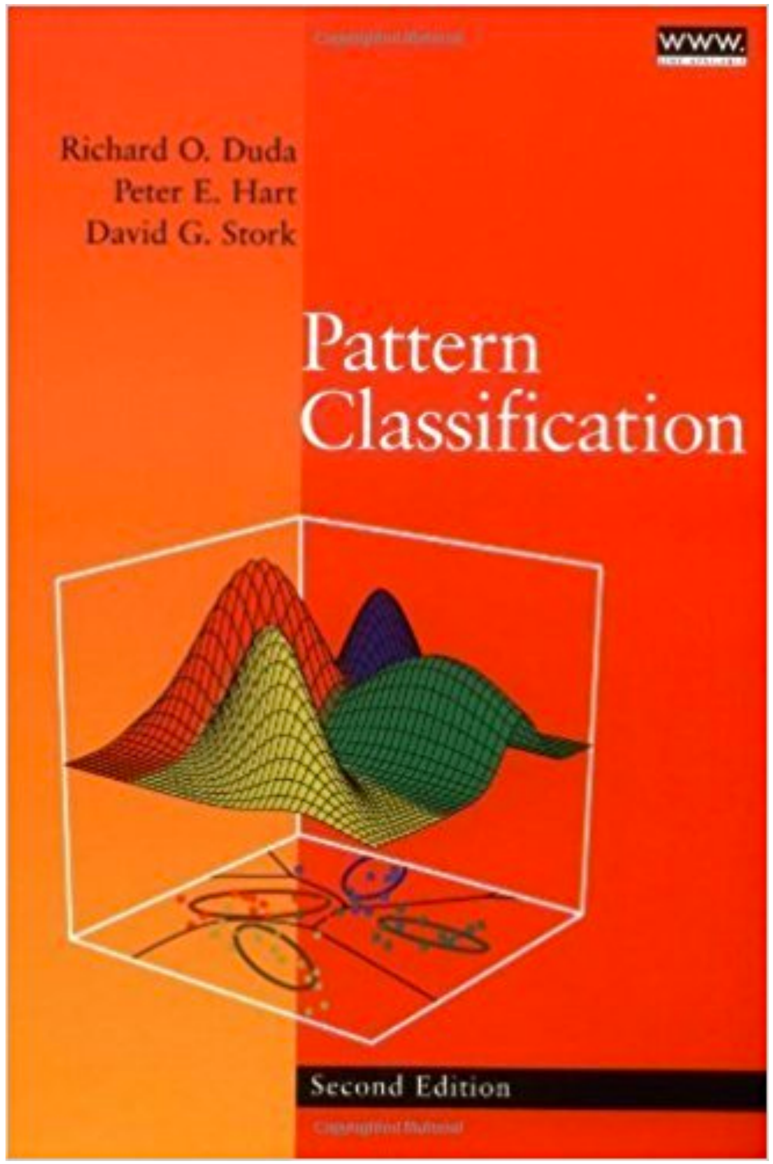
- $$P(\text{class } A_i \,|\, \text{data}) = \frac{\exp(s(A_i))}{\exp \text{softmax}(s(A_1), \ldots, s(A_K))} = \frac{\exp(s(A_i))}{\sum_{i=1}^{K} \exp(s(A_i))}$$

- $f_i(\text{data } x) = \mathbf{w}_i^\top \Phi(x) = \mathbf{w}_i^\top x$

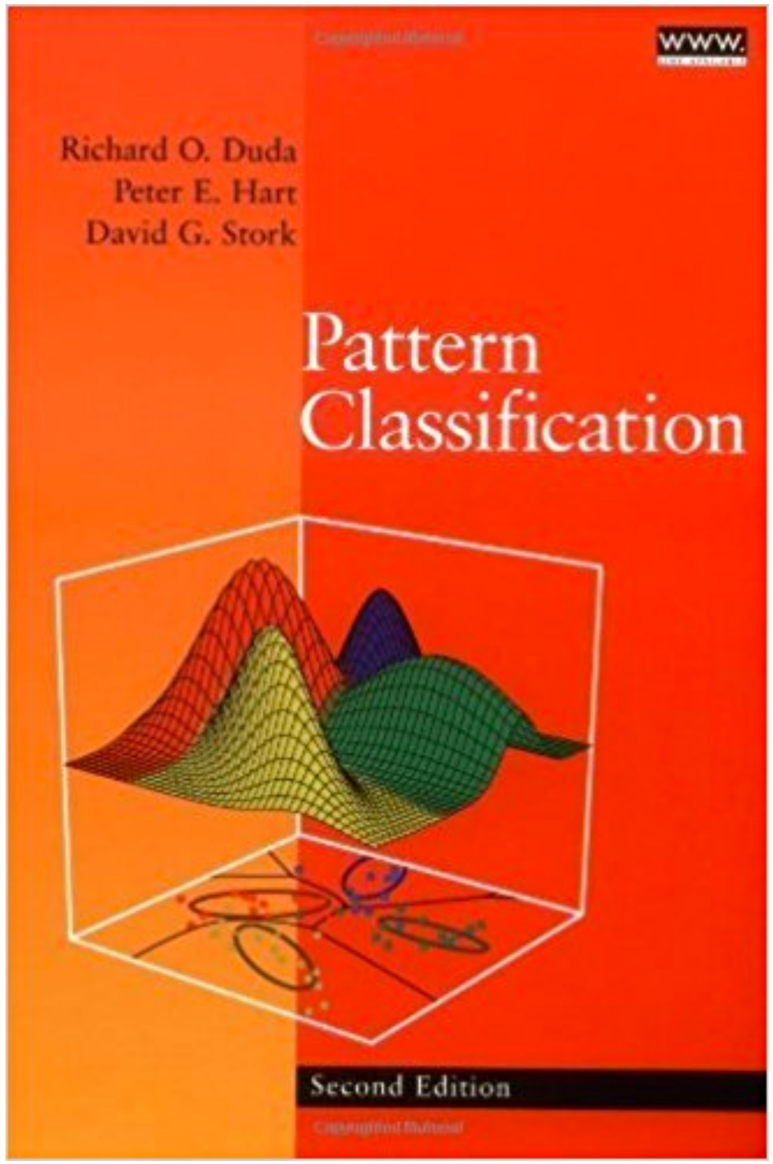
- Did not seek a decision boundary, just an argmax

Decision boundaries from probability distributions: zero log-odds

Decision boundaries from probability distributions: zero log-odds

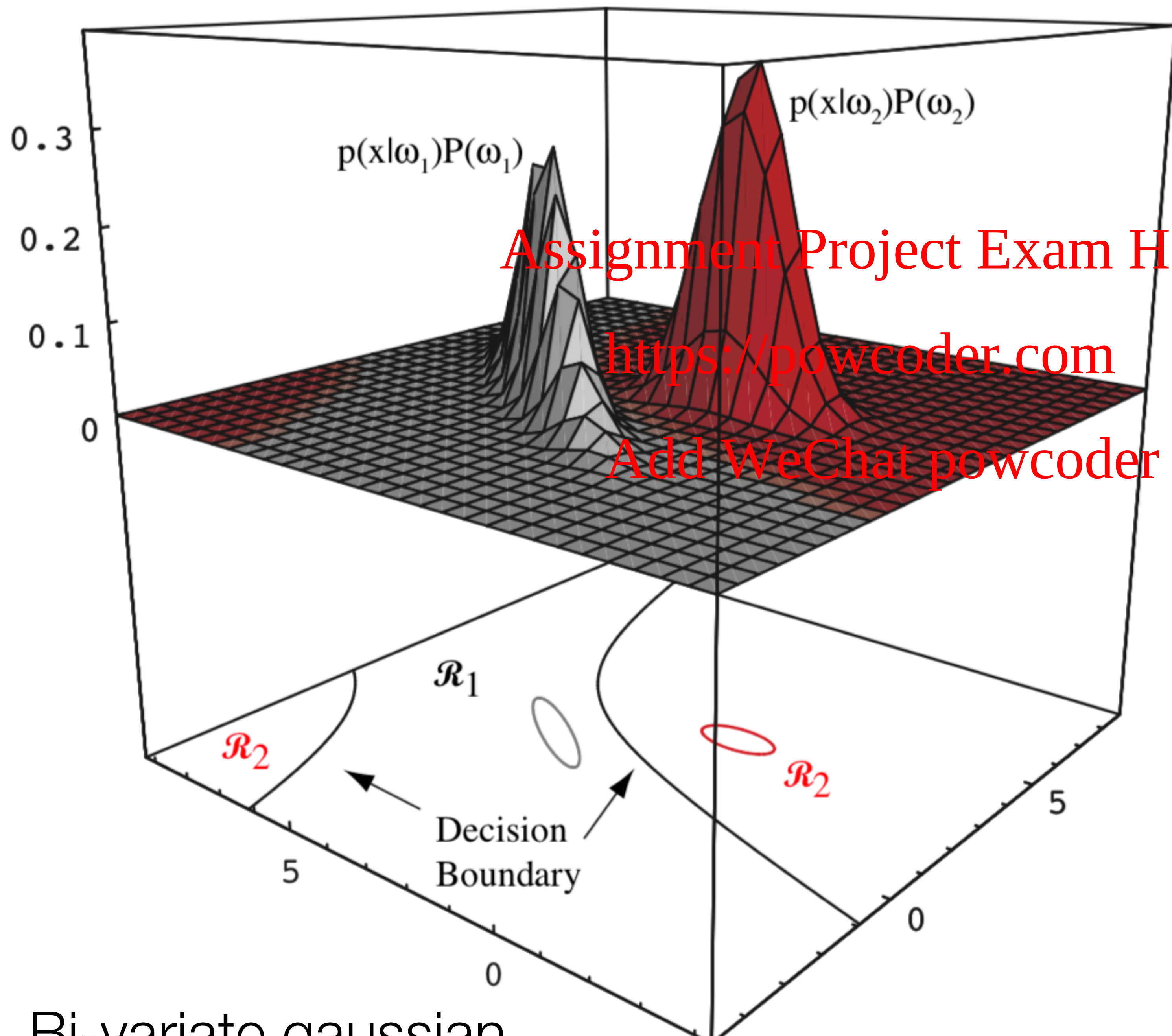


Decision boundaries from probability distributions: zero log-odds

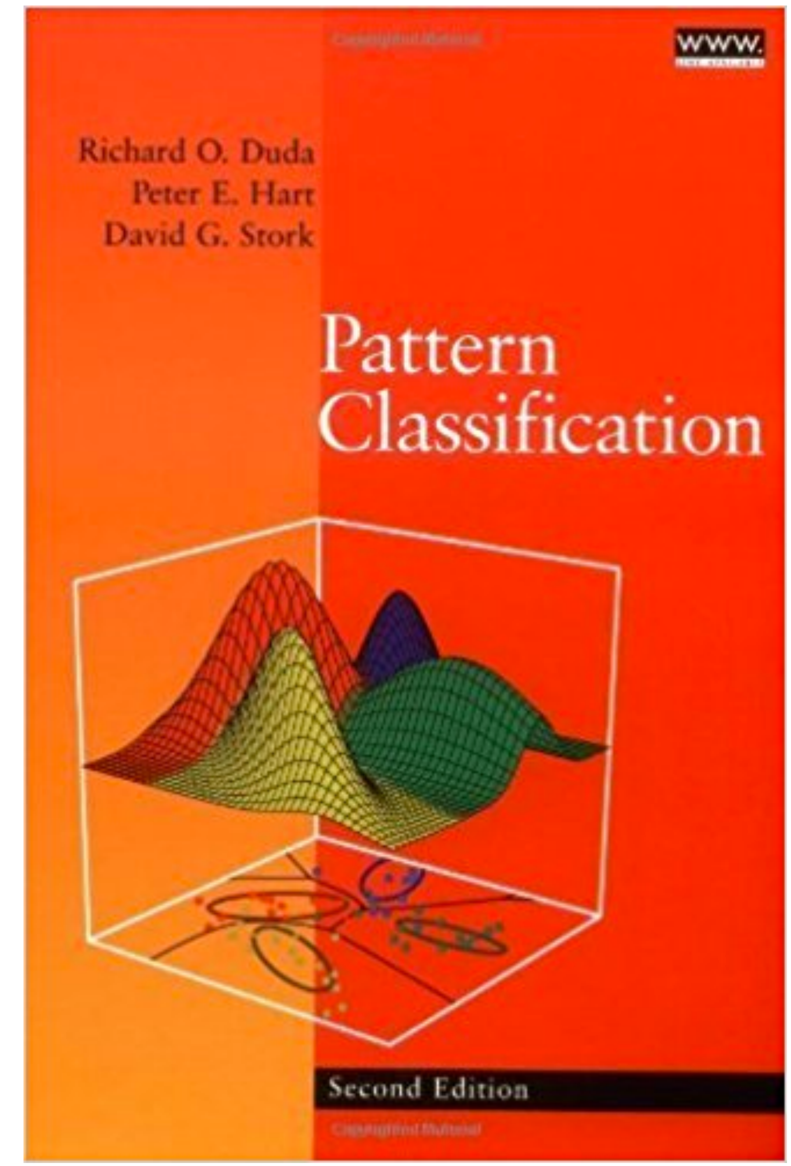


Notice the contours and line segments at the bottom of the cover graphic

Decision boundaries from probability distributions: zero log-odds



Bi-variate gaussian



Notice the contours and line segments at the bottom of the cover graphic

Log of ratio of two multi-dimensional Gaussians

Log of ratio of two multi-dimensional Gaussians

$$p(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^p|\boldsymbol{\Sigma}|}}\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

Log of ratio of two multi-dimensional Gaussians

$$p(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^p |\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

Precision matrix $\Lambda = \Sigma^{-1}$

Log of ratio of two multi-dimensional Gaussians

$$p(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^p\,|\,\boldsymbol{\Sigma}\,|}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

Precision matrix $\Lambda = \Sigma^{-1}$

For 2D features

$$p(x_1, x_2 | C = a) = \frac{\sqrt{|\Lambda_a|}}{2\pi} \exp\left\{-\frac{1}{2}\begin{bmatrix} x_1 - \mu_{a1} & x_2 - \mu_{a2}\end{bmatrix}\begin{bmatrix}\Lambda_{a11} & \Lambda_{a12} \\ \Lambda_{a21} & \Lambda_{a22}\end{bmatrix}\begin{bmatrix} x_1 - \mu_{a1} \\ x_2 - \mu_{a2}\end{bmatrix}\right\}$$

$$p(x_1, x_2 | C = b) = \frac{\sqrt{|\Lambda_b|}}{2\pi} \exp\left\{-\frac{1}{2}\begin{bmatrix} x_1 - \mu_{b1} & x_2 - \mu_{b2}\end{bmatrix}\begin{bmatrix}\Lambda_{b11} & \Lambda_{b12} \\ \Lambda_{b21} & \Lambda_{b22}\end{bmatrix}\begin{bmatrix} x_1 - \mu_{b1} \\ x_2 - \mu_{b2}\end{bmatrix}\right\}$$

Log of ratio of two multi-dimensional Gaussians

$$p(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^p\,|\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

Precision matrix $\Lambda = \Sigma^{-1}$

For 2D features

$$p(x_1, x_2 | C = a) = \frac{\sqrt{|\Lambda_a|}}{2\pi} \exp\left\{-\frac{1}{2}\begin{bmatrix} x_1 - \mu_{a1} & x_2 - \mu_{a2}\end{bmatrix}\begin{bmatrix}\Lambda_{a11} & \Lambda_{a12} \\ \Lambda_{a21} & \Lambda_{a22}\end{bmatrix}\begin{bmatrix} x_1 - \mu_{a1} \\ x_2 - \mu_{a2}\end{bmatrix}\right\}$$

$$p(x_1, x_2 | C = b) = \frac{\sqrt{|\Lambda_b|}}{2\pi} \exp\left\{-\frac{1}{2}\begin{bmatrix} x_1 - \mu_{b1} & x_2 - \mu_{b2}\end{bmatrix}\begin{bmatrix}\Lambda_{b11} & \Lambda_{b12} \\ \Lambda_{b21} & \Lambda_{b22}\end{bmatrix}\begin{bmatrix} x_1 - \mu_{b1} \\ x_2 - \mu_{b2}\end{bmatrix}\right\}$$

Log-odds (no prior) for 2-class classification

$$\ln\left(\frac{p(\boldsymbol{x}|C_+)}{p(\boldsymbol{x}|C_-)}\right) = \ln\left(\frac{|\Lambda_+|}{|\Lambda_-|}\right) + \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_-)^T\Lambda_-(\boldsymbol{x}-\boldsymbol{\mu}_-) - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_+)^T\Lambda_+(\boldsymbol{x}-\boldsymbol{\mu}_+)$$

Log of ratio of posteriors of two Gaussians: for equal and unequal covariances

Log of ratio of posteriors of two Gaussians: for equal and unequal covariances

$$
\begin{aligned}
g_-(\boldsymbol{x}) &= \ln P(C_-) + \tfrac{1}{2}\ln|\Lambda_-| - \tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_-)^T\Lambda_-(\boldsymbol{x}-\boldsymbol{\mu}_-) \\
&= \ln P(C_-) + \tfrac{1}{2}\ln|\Lambda_-| - \tfrac{1}{2}\left(\boldsymbol{x}^T\Lambda_-\boldsymbol{x} - 2\boldsymbol{\mu}_-^T\Lambda_-\boldsymbol{x} + \boldsymbol{\mu}_-^T\Lambda_-\boldsymbol{\mu}_-\right) \\
g_+(\boldsymbol{x}) &= \ln P(C_+) + \tfrac{1}{2}\ln|\Lambda_+| - \tfrac{1}{2}\left(\boldsymbol{x}^T\Lambda_+\boldsymbol{x} - 2\boldsymbol{\mu}_+^T\Lambda_+\boldsymbol{x} + \boldsymbol{\mu}_+^T\Lambda_+\boldsymbol{\mu}_+\right)
\end{aligned}
$$

Log of ratio of posteriors of two Gaussians: for equal and unequal covariances

$$
\begin{aligned}
g_-(\boldsymbol{x}) &= \ln P(C_-) + \tfrac{1}{2}\ln|\Lambda_-| - \tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_-)^T\Lambda_-(\boldsymbol{x}-\boldsymbol{\mu}_-) \\
&= \ln P(C_-) + \tfrac{1}{2}\ln|\Lambda_-| - \tfrac{1}{2}\left(\boldsymbol{x}^T\Lambda_-\boldsymbol{x} - 2\boldsymbol{\mu}_-^T\Lambda_-\boldsymbol{x} + \boldsymbol{\mu}_-^T\Lambda_-\boldsymbol{\mu}_-\right) \\
g_+(\boldsymbol{x}) &= \ln P(C_+) + \tfrac{1}{2}\ln|\Lambda_+| - \tfrac{1}{2}\left(\boldsymbol{x}^T\Lambda_+\boldsymbol{x} - 2\boldsymbol{\mu}_+^T\Lambda_+\boldsymbol{x} + \boldsymbol{\mu}_+^T\Lambda_+\boldsymbol{\mu}_+\right)
\end{aligned}
$$

For unequal covariances, $\Lambda_+ \neq \Lambda_-$ decision boundary is quadratic

Your turn: what is the quadratic term in $g_+(x) - g_-(x)$

Log of ratio of posteriors of two Gaussians: for equal and unequal covariances

$$
\begin{aligned}
g_-(\boldsymbol{x}) &= \ln P(C_-) + \tfrac{1}{2}\ln|\Lambda_-| - \tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_-)^T\Lambda_-(\boldsymbol{x}-\boldsymbol{\mu}_-) \\
&= \ln P(C_-) + \tfrac{1}{2}\ln|\Lambda_-| - \tfrac{1}{2}\left(\boldsymbol{x}^T\Lambda_-\boldsymbol{x} - 2\boldsymbol{\mu}_-^T\Lambda_-\boldsymbol{x} + \boldsymbol{\mu}_-^T\Lambda_-\boldsymbol{\mu}_-\right) \\
g_+(\boldsymbol{x}) &= \ln P(C_+) + \tfrac{1}{2}\ln|\Lambda_+| - \tfrac{1}{2}\left(\boldsymbol{x}^T\Lambda_+\boldsymbol{x} - 2\boldsymbol{\mu}_+^T\Lambda_+\boldsymbol{x} + \boldsymbol{\mu}_+^T\Lambda_+\boldsymbol{\mu}_+\right)
\end{aligned}
$$

For unequal covariances, $\Lambda_+ \neq \Lambda_-$ decision boundary is quadratic

Your turn: what is the quadratic term in $g_+(x) - g_-(x)$

Decision boundary for equal covariances $\Lambda_+ = \Lambda_- = \Lambda$ is linear:

Log of ratio of posteriors of two Gaussians: for equal and unequal covariances

$$
\begin{aligned}
g_-(\boldsymbol{x}) &= \ln P(C_-) + \tfrac{1}{2}\ln|\Lambda_-| - \tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_-)^T\Lambda_-(\boldsymbol{x}-\boldsymbol{\mu}_-) \\
&= \ln P(C_-) + \tfrac{1}{2}\ln|\Lambda_-| - \tfrac{1}{2}\left(\boldsymbol{x}^T\Lambda_-\boldsymbol{x} - 2\boldsymbol{\mu}_-^T\Lambda_-\boldsymbol{x} + \boldsymbol{\mu}_-^T\Lambda_-\boldsymbol{\mu}_-\right) \\
g_+(\boldsymbol{x}) &= \ln P(C_+) + \tfrac{1}{2}\ln|\Lambda_+| - \tfrac{1}{2}\left(\boldsymbol{x}^T\Lambda_+\boldsymbol{x} - 2\boldsymbol{\mu}_+^T\Lambda_+\boldsymbol{x} + \boldsymbol{\mu}_+^T\Lambda_+\boldsymbol{\mu}_+\right)
\end{aligned}
$$

For unequal covariances, $\Lambda_+ \neq \Lambda_-$ decision boundary is quadratic

Your turn: what is the quadratic term in $g_+(x) - g_-(x)$

Decision boundary for equal covariances $\Lambda_+ = \Lambda_- = \Lambda$ is linear:

$$
g_+(\boldsymbol{x}) - g_-(\boldsymbol{x}) = \ln\frac{P(C_+)}{P(C_-)} + (\boldsymbol{\mu}_+ - \boldsymbol{\mu}_-)^T\Lambda\boldsymbol{x} + \boldsymbol{x}\text{-independent term}
$$

Log of ratio of posteriors of two Gaussians: for equal and unequal covariances

$$
\begin{aligned}
g_-(\boldsymbol{x}) &= \ln P(C_-) + \tfrac{1}{2}\ln|\Lambda_-| - \tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_-)^T\Lambda_-(\boldsymbol{x}-\boldsymbol{\mu}_-) \\
&= \ln P(C_-) + \tfrac{1}{2}\ln|\Lambda_-| - \tfrac{1}{2}\left(\boldsymbol{x}^T\Lambda_-\boldsymbol{x} - 2\boldsymbol{\mu}_-^T\Lambda_-\boldsymbol{x} + \boldsymbol{\mu}_-^T\Lambda_-\boldsymbol{\mu}_-\right) \\
g_+(\boldsymbol{x}) &= \ln P(C_+) + \tfrac{1}{2}\ln|\Lambda_+| - \tfrac{1}{2}\left(\boldsymbol{x}^T\Lambda_+\boldsymbol{x} - 2\boldsymbol{\mu}_+^T\Lambda_+\boldsymbol{x} + \boldsymbol{\mu}_+^T\Lambda_+\boldsymbol{\mu}_+\right)
\end{aligned}
$$

For unequal covariances, $\Lambda_+ \neq \Lambda_-$ decision boundary is quadratic

Your turn: what is the quadratic term in $g_+(x) - g_-(x)$

Decision boundary for <u>equal covariances</u> $\Lambda_+ = \Lambda_- = \Lambda$ is linear:

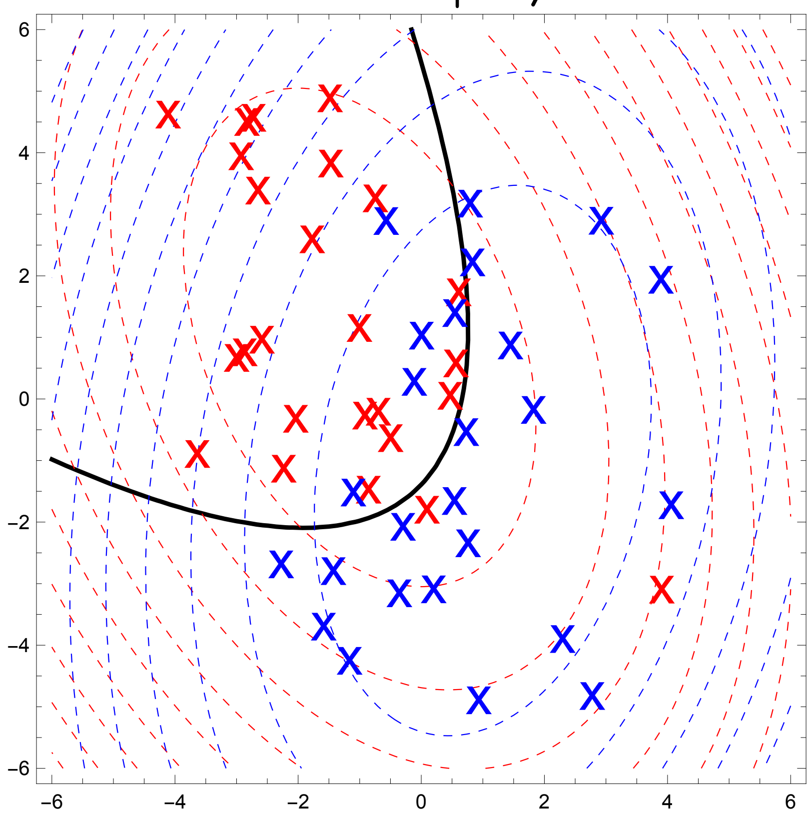
$$
g_+(\boldsymbol{x}) - g_-(\boldsymbol{x}) = \ln\frac{P(C_+)}{P(C_-)} + (\boldsymbol{\mu}_+ - \boldsymbol{\mu}_-)^T\Lambda\boldsymbol{x} + \boldsymbol{x}\text{-independent term}
$$

Your turn: derive the linear equation

Discriminants (decision boundaries) for 2-class gaussians: linear if covariances equal, quadratic if not

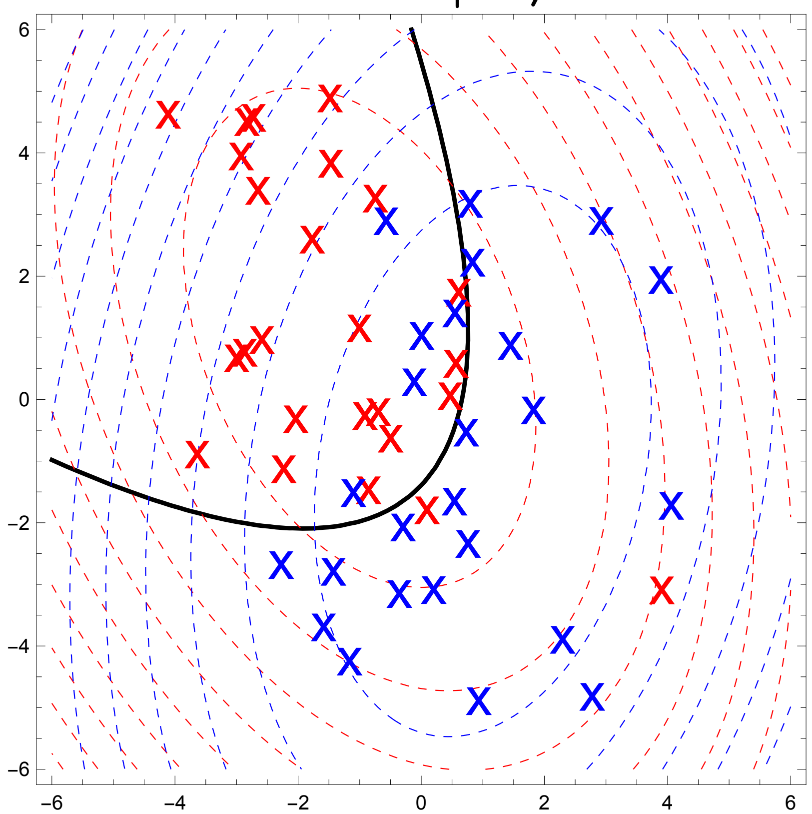
Discriminants (decision boundaries) for 2-class gaussians: linear if covariances equal, quadratic if not

$$\Sigma_+ \neq \Sigma_-$$



Discriminants (decision boundaries) for 2-class gaussians: linear if covariances equal, quadratic if not

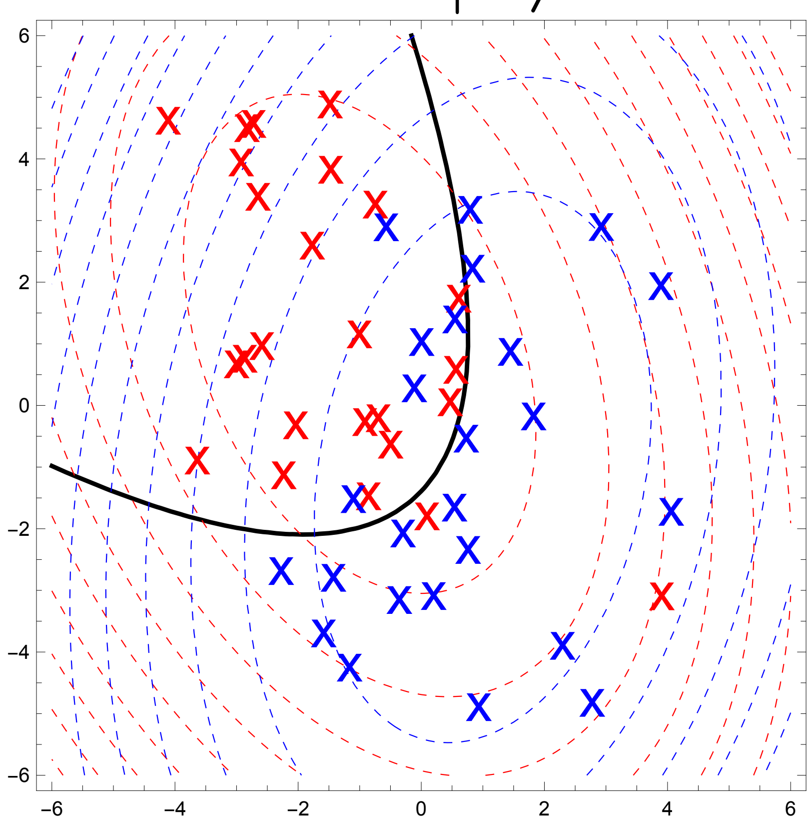
$$\Sigma_+ \neq \Sigma_-$$



If the covariance matrices are the same, the quadratic term is common to both classes, and drops out in the ratio

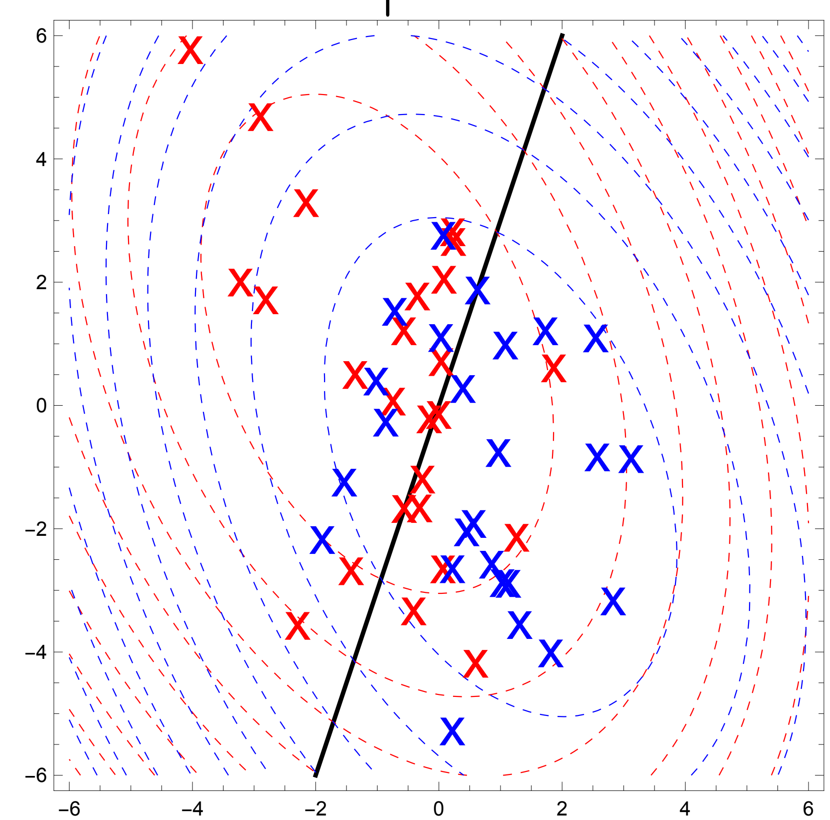
Discriminants (decision boundaries) for 2-class gaussians: linear if covariances equal, quadratic if not

$\Sigma_+ \neq \Sigma_-$



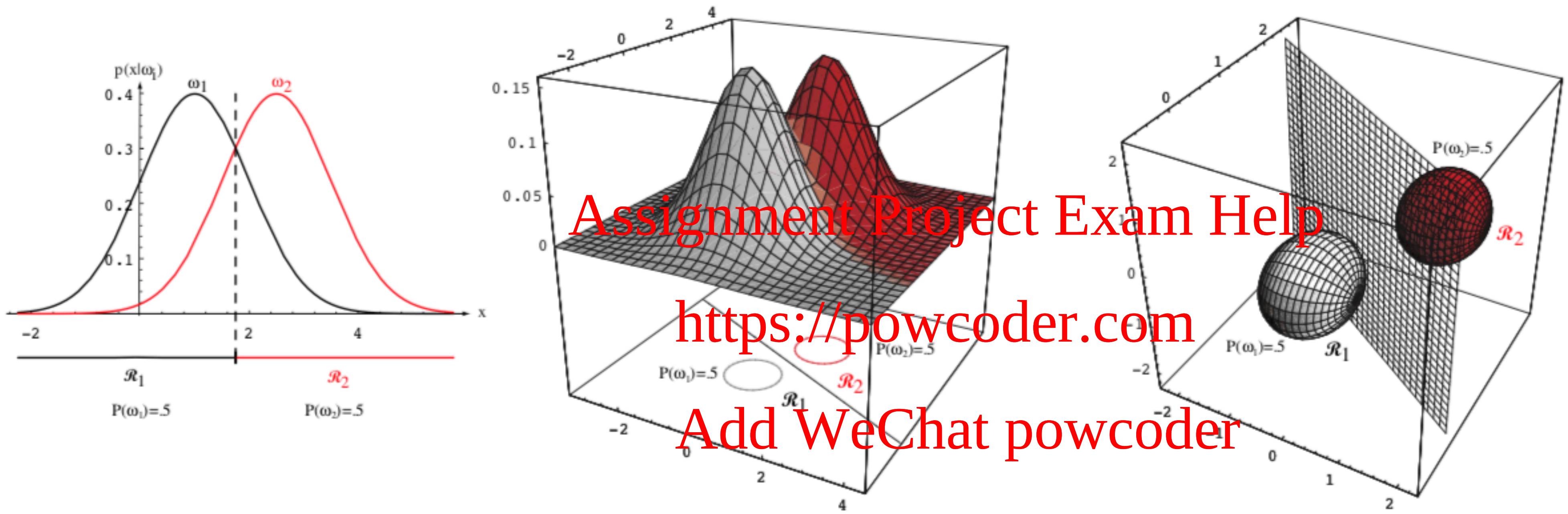
If the covariance matrices are the same, the quadratic term is common to both classes, and drops out in the ratio

$\Sigma_+ = \Sigma_-$

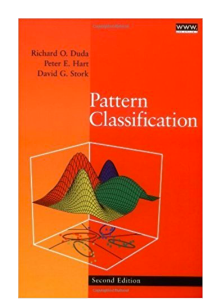


Discriminants for Gaussians for equal priors: data distribution equally divided

1-dim 2-dim 3-dim

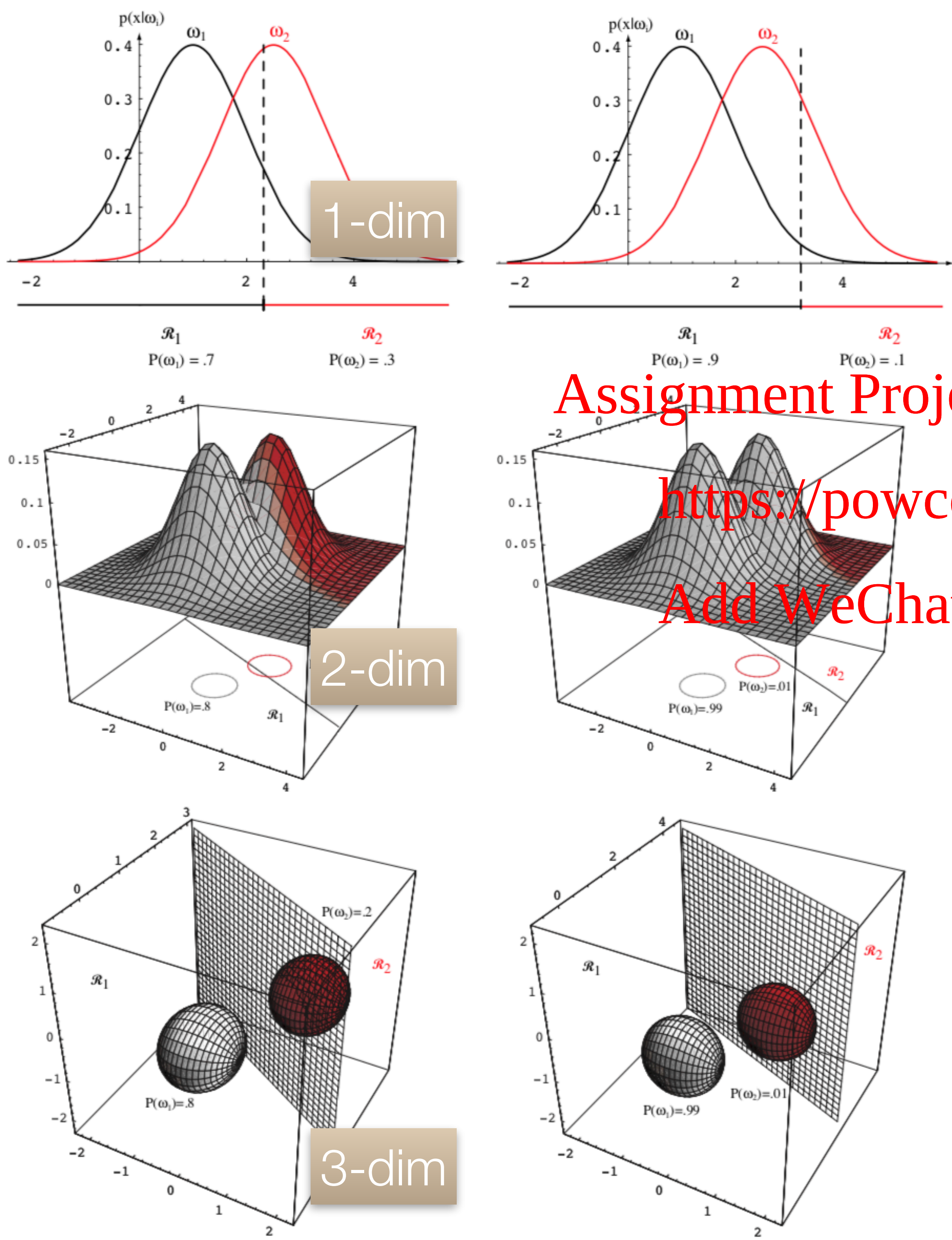


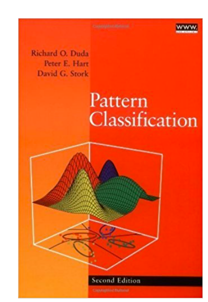
Equal priors: $\log(P_a/P_b) = 0$



From Duda and Hart

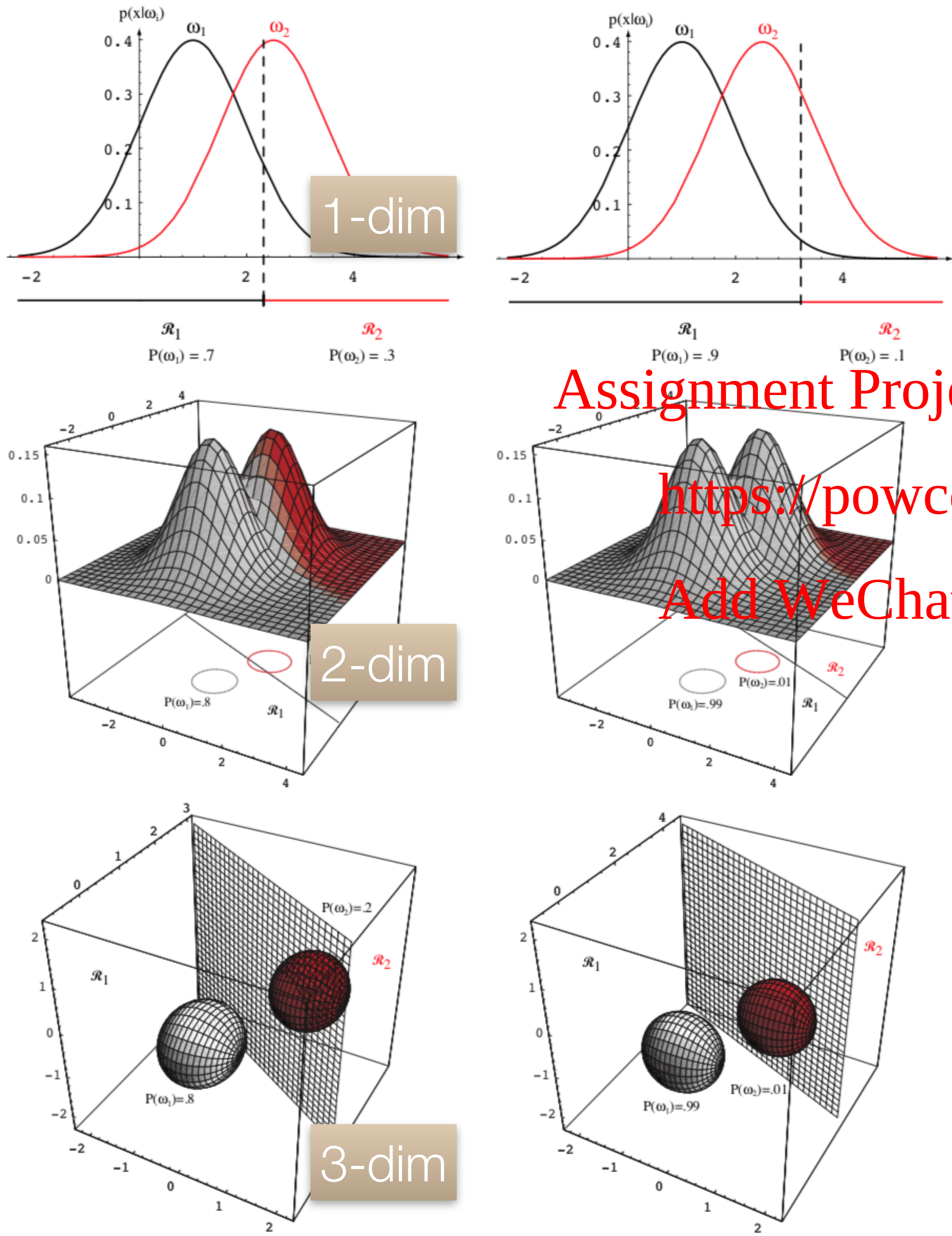
Discriminants for Gaussians: unequal priors shifts decision boundary by $\log(P_a/P_b)$



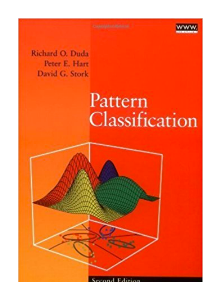


From Duda and Hart

Discriminants for Gaussians: unequal priors shifts decision boundary by $\log(P_a/P_b)$

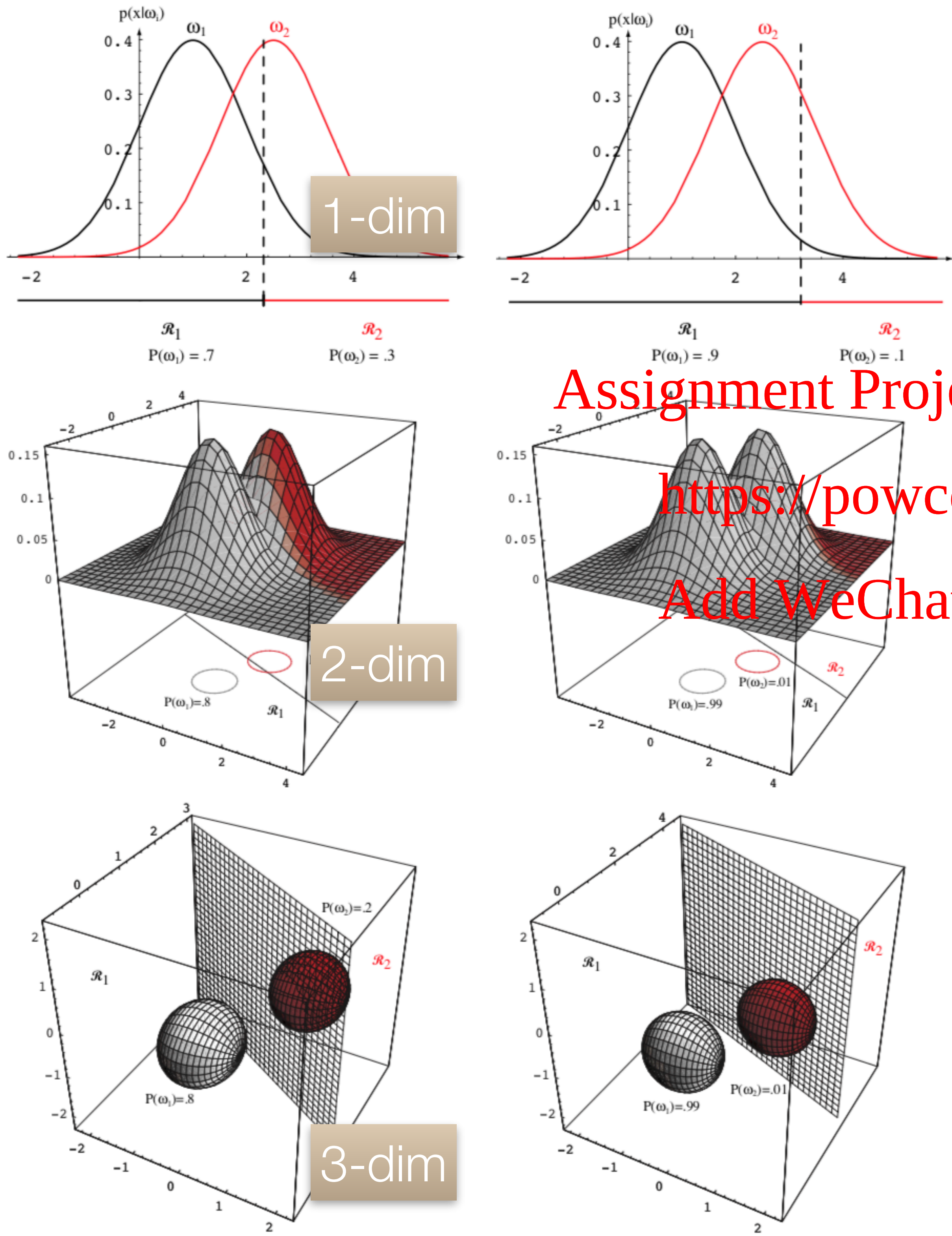


Prior overrides data distributions



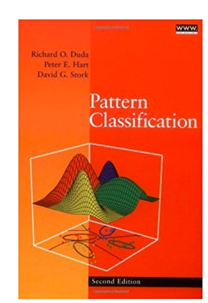
From Duda and Hart

Discriminants for Gaussians: unequal priors shifts decision boundary by $\log(P_a/P_b)$



Prior overrides data distributions

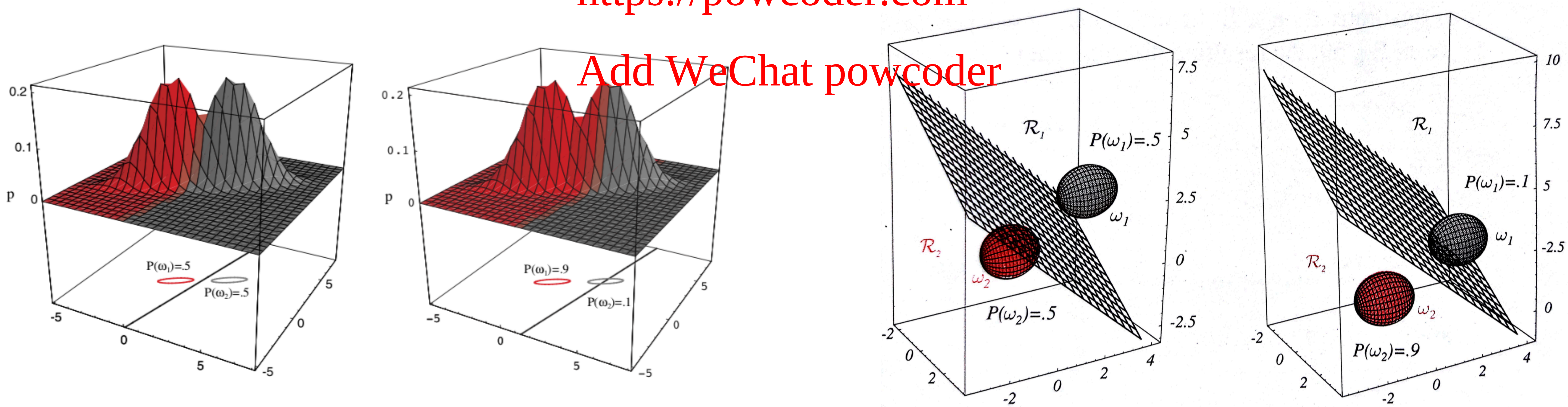
Notice how far boundary shifts

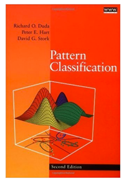


From Duda and Hart

Decision boundaries not perpendicular to line joining means if covariance matrix not proportional to the identity matrix

$$\boldsymbol{w} = \Sigma^{-1}(\boldsymbol{\mu}_+ - \boldsymbol{\mu}_-)$$



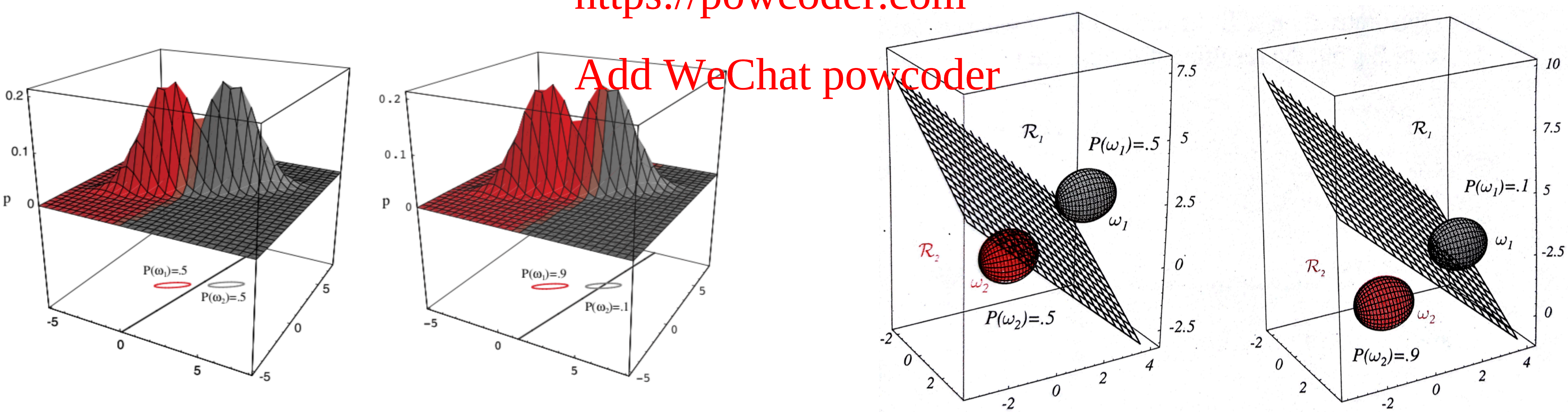


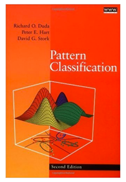
From Duda and Hart

Decision boundaries not perpendicular to line joining means if covariance matrix not proportional to the identity matrix

$$g_+(\boldsymbol{x}) - g_-(\boldsymbol{x}) \quad = \quad \ln \frac{P(C_+)}{P(C_-)} + (\boldsymbol{\mu}_+ - \boldsymbol{\mu}_-)^T \Lambda \boldsymbol{x} + \boldsymbol{x}\text{-independent term}$$

$$\boldsymbol{w} = \Sigma^{-1}(\boldsymbol{\mu}_+ - \boldsymbol{\mu}_-)$$

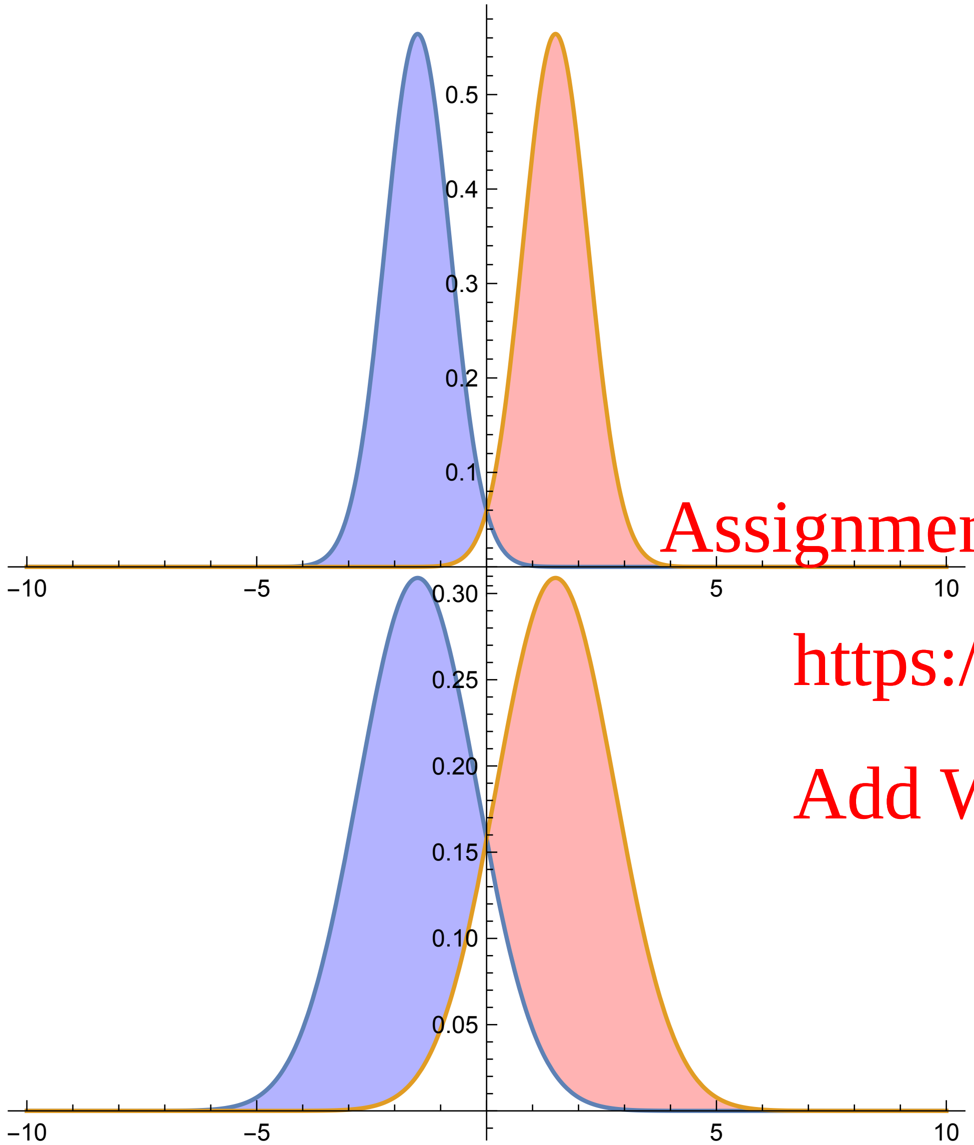




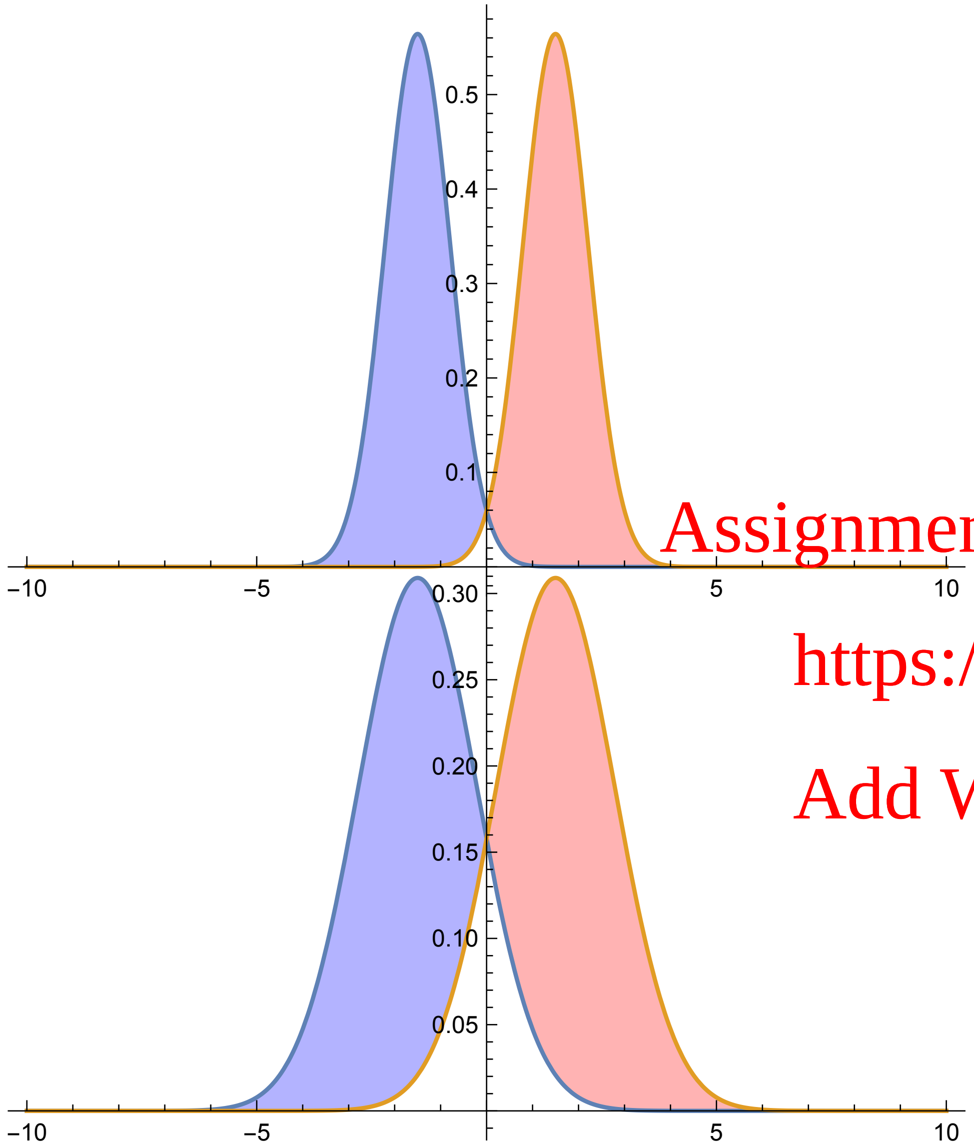
From Duda and Hart

Distinctiveness relative to natural variation measured by distance between class means scaled by standard deviation:

Distinctiveness relative to natural variation measured by distance between class means scaled by standard deviation:

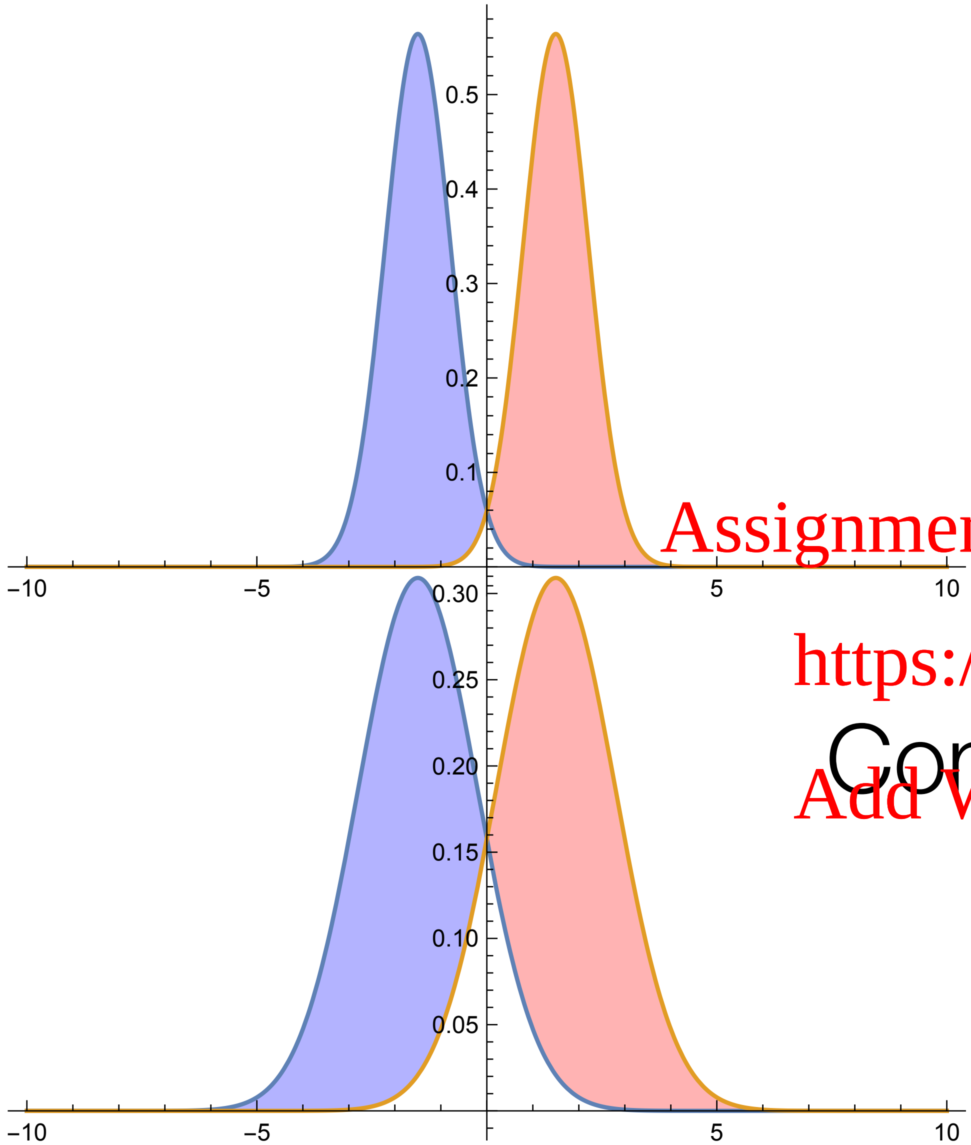


Distinctiveness relative to natural variation measured by distance between class means scaled by standard deviation:



$$d = \frac{\mu_1 - \mu_2}{\sigma}$$

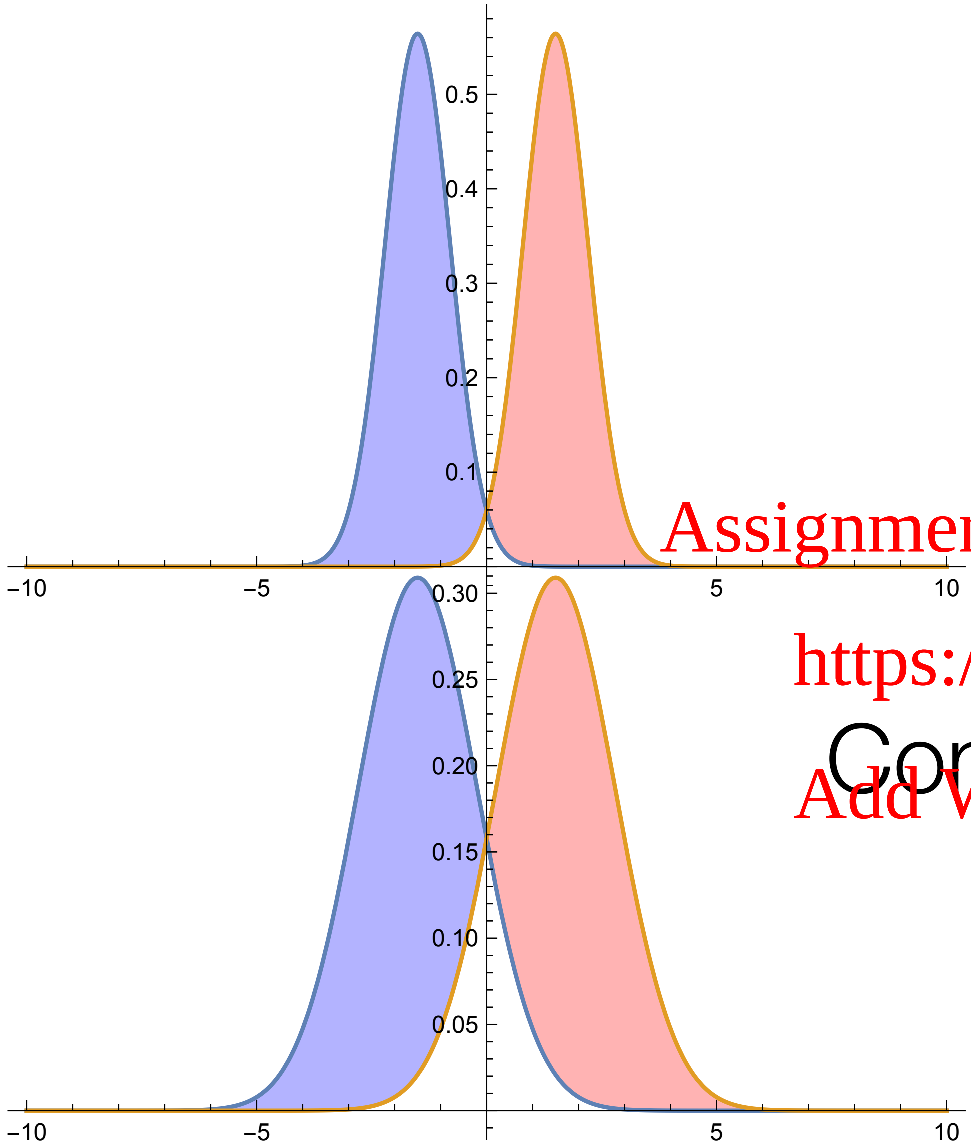
Distinctiveness relative to natural variation measured by distance between class means scaled by standard deviation:



$$d = \frac{\mu_1 - \mu_2}{\sigma}$$

Controls for overlap between classes

Distinctiveness relative to natural variation measured by distance between class means scaled by standard deviation:



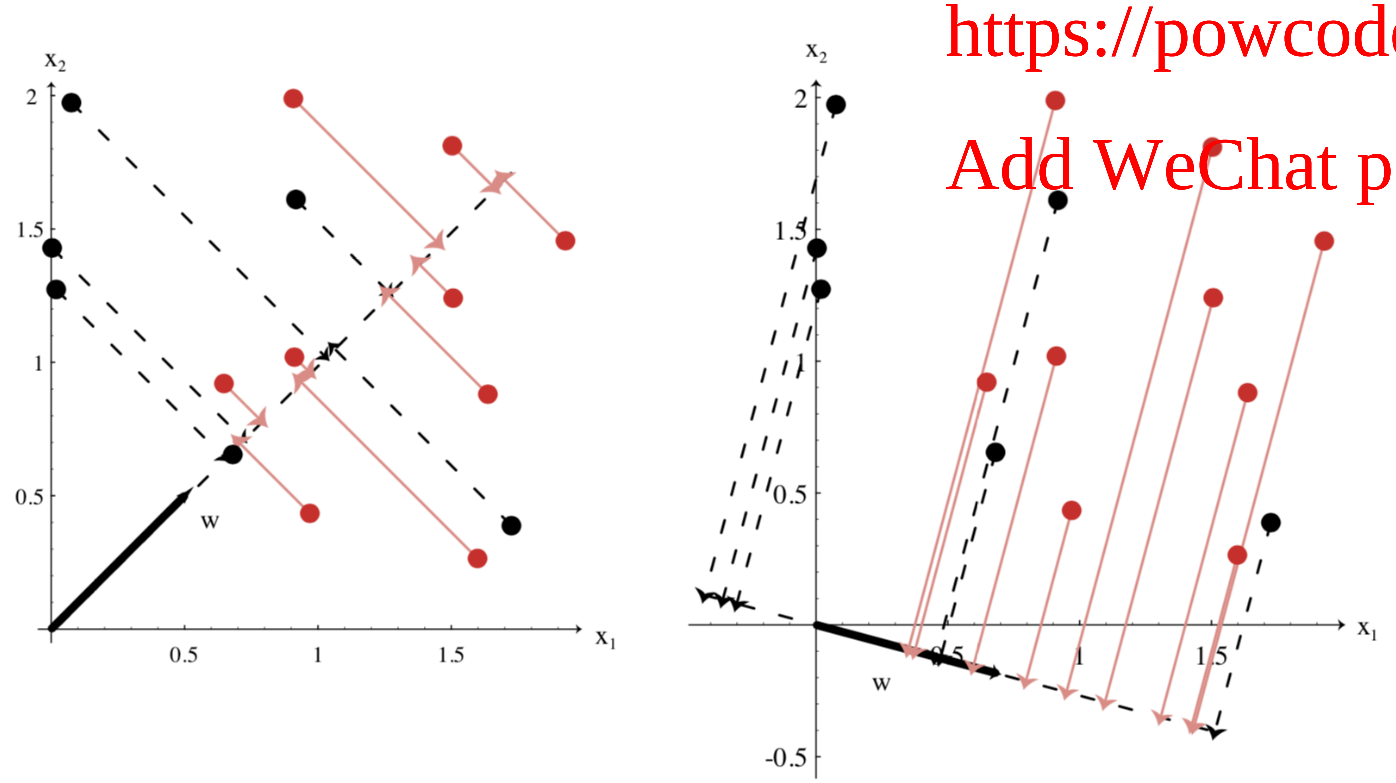
$$d = \frac{\mu_1 - \mu_2}{\sigma}$$

Controls for overlap between classes

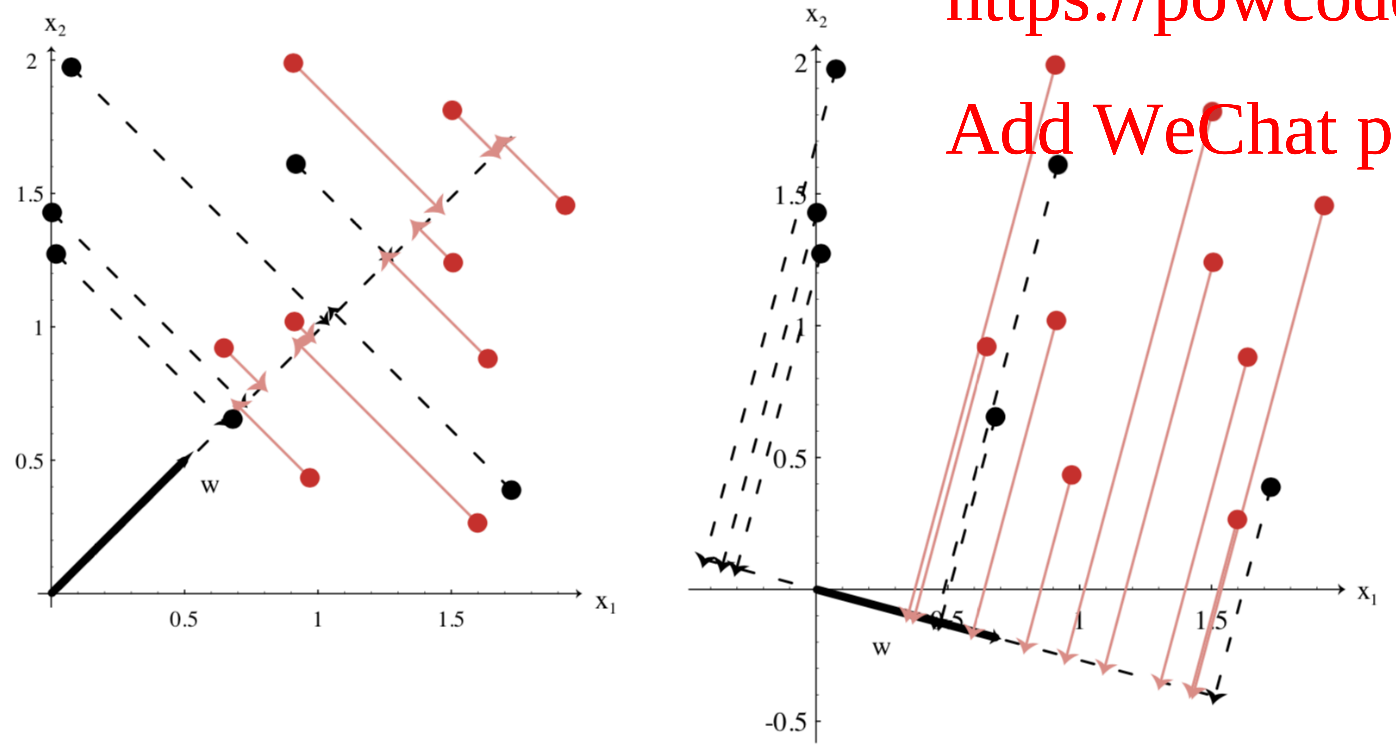
Mahalanobis distance

$$d(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sqrt{(\boldsymbol{x}_1 - \boldsymbol{x}_2)^T \Sigma^{-1} (\boldsymbol{x}_1 - \boldsymbol{x}_2)}$$

Fisher Linear Discriminant Analysis: projections along discriminative directions

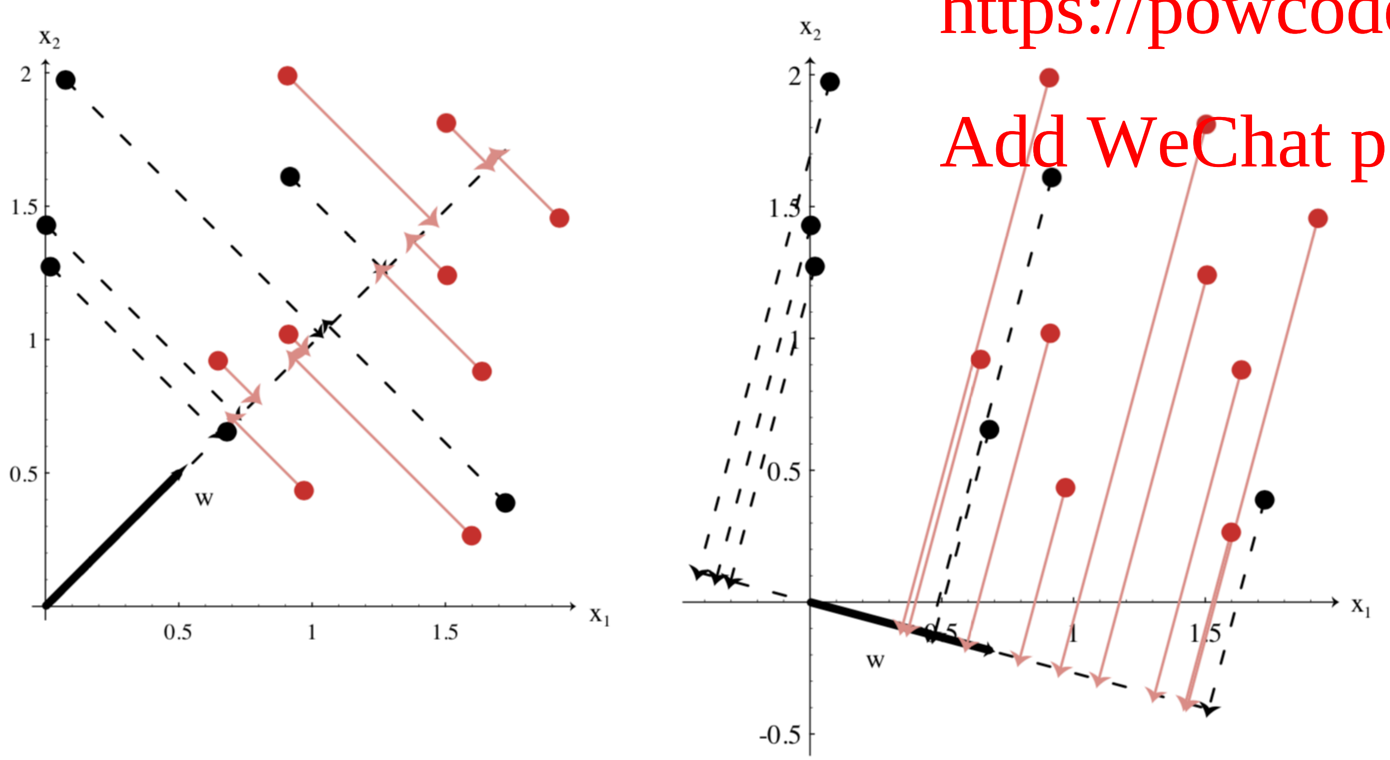


Fisher Linear Discriminant Analysis: projections along discriminative directions



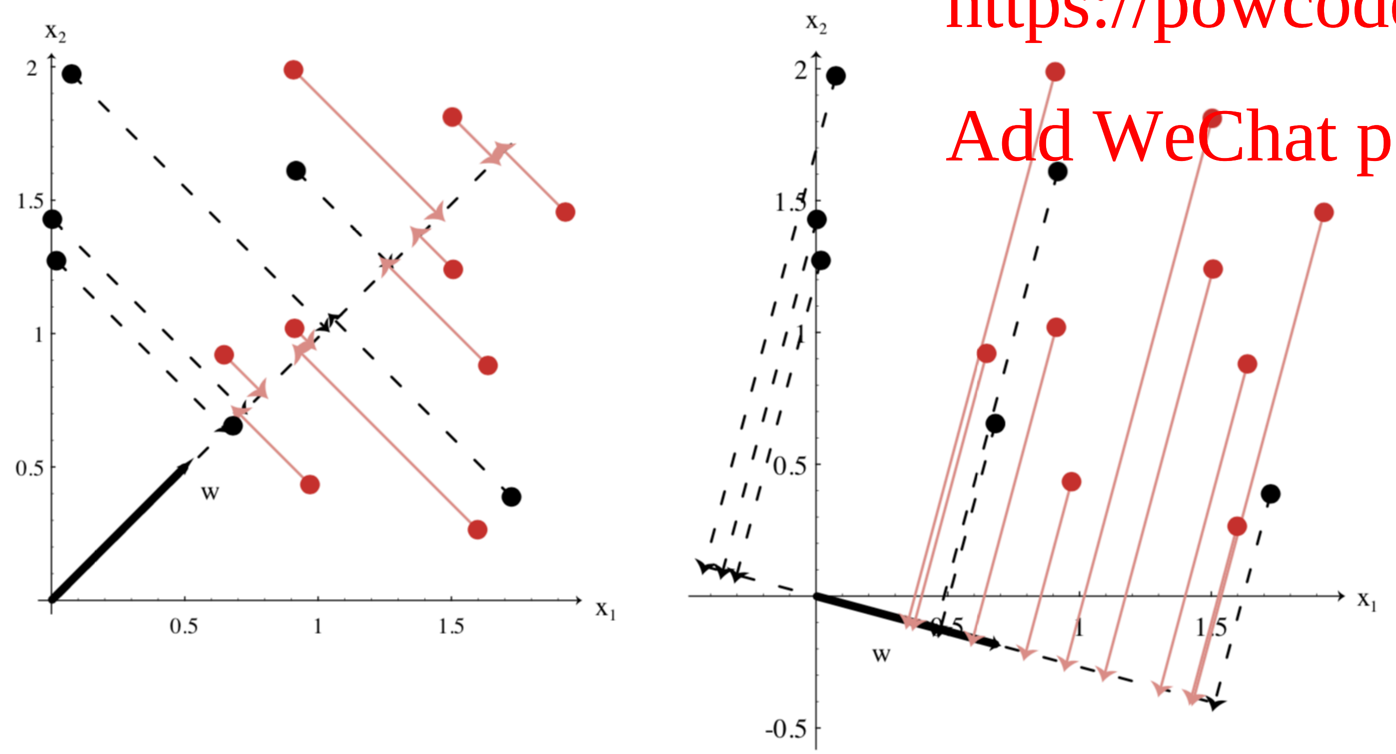
Fisher Linear Discriminant Analysis: projections along discriminative directions

$$\boldsymbol{\mu}_c = \frac{1}{n_c} \sum_{\boldsymbol{x}^n \in \mathcal{D}_c} \boldsymbol{x}^n$$



Fisher Linear Discriminant Analysis: projections along discriminative directions

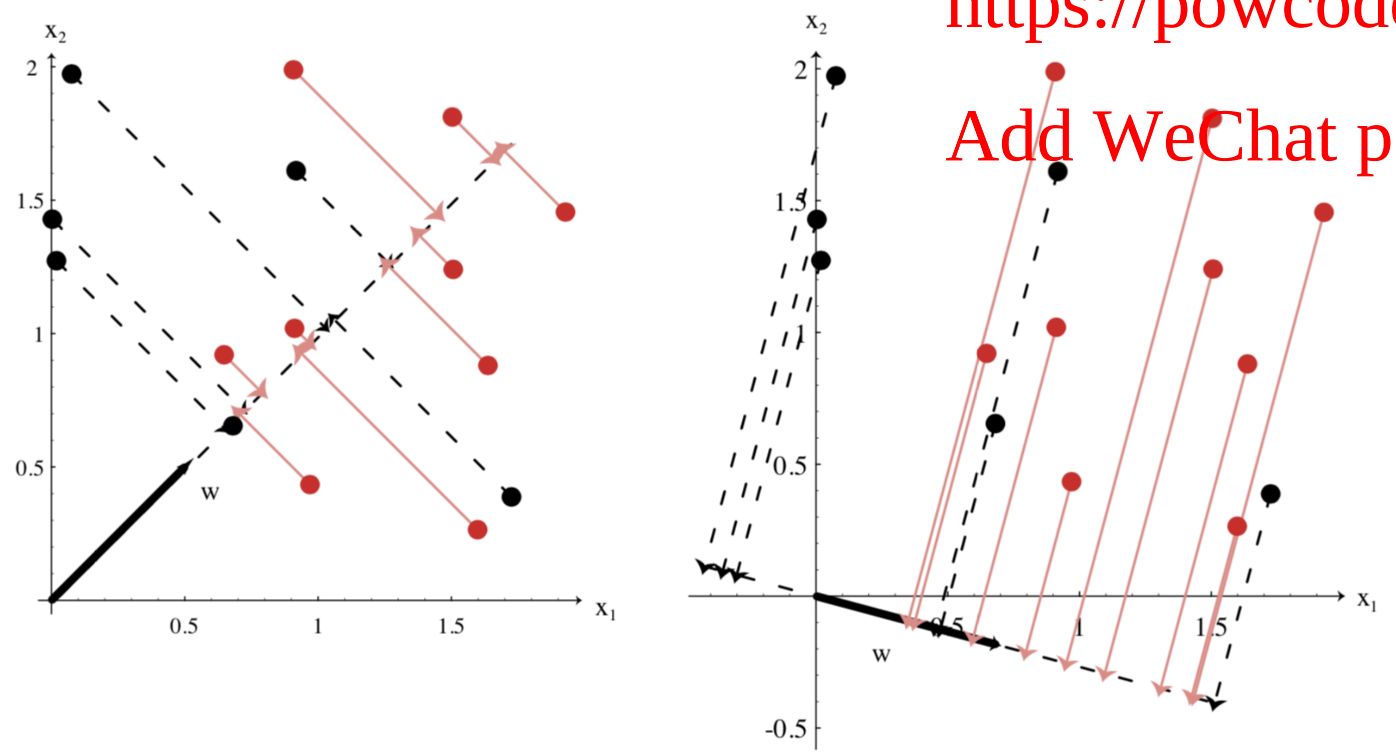
$$\boldsymbol{\mu}_c = \frac{1}{n_c} \sum_{\boldsymbol{x}^n \in \mathcal{D}_c} \boldsymbol{x}^n \qquad m_c = \boldsymbol{w}^T \boldsymbol{\mu}_c = \frac{1}{n_c} \sum_{\boldsymbol{x}^n \in \mathcal{D}_c} \boldsymbol{w}^T \boldsymbol{x}^n$$



Fisher Linear Discriminant Analysis: projections along discriminative directions

$$\boldsymbol{\mu}_c = \frac{1}{n_c} \sum_{\boldsymbol{x}^n \in \mathcal{D}_c} \boldsymbol{x}^n \qquad m_c = \boldsymbol{w}^T \boldsymbol{\mu}_c = \frac{1}{n_c} \sum_{\boldsymbol{x}^n \in \mathcal{D}_c} \boldsymbol{w}^T \boldsymbol{x}^n$$

$$\triangleq \frac{1}{n_c} \sum_{\boldsymbol{x}^n \in \mathcal{D}_c} y^n$$

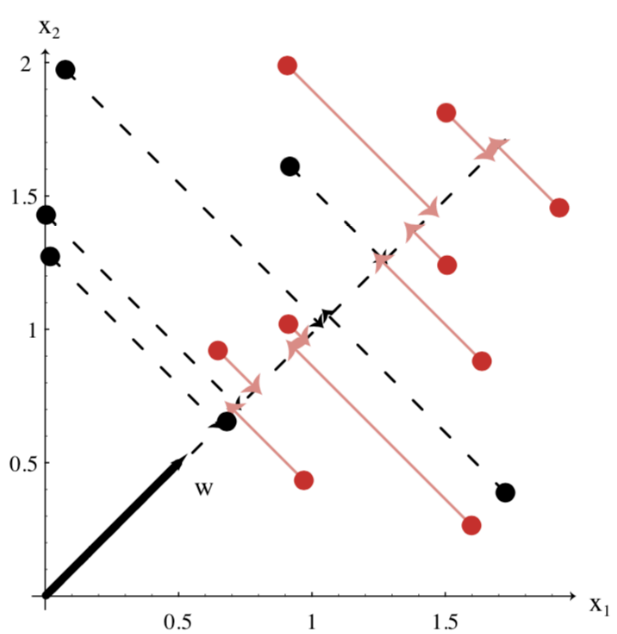


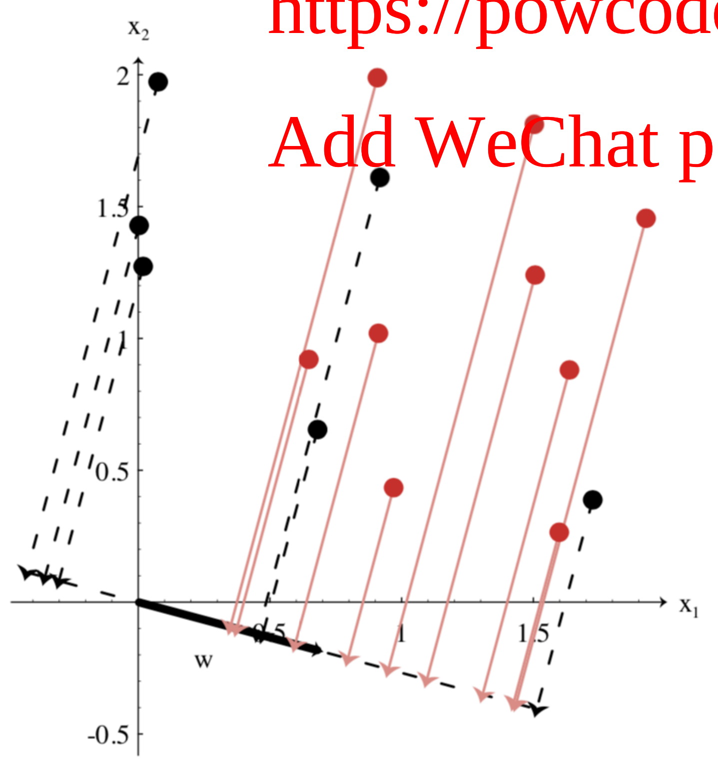
Fisher Linear Discriminant Analysis: projections along discriminative directions

$$\boldsymbol{\mu}_c = \frac{1}{n_c} \sum_{\boldsymbol{x}^n \in \mathcal{D}_c} \boldsymbol{x}^n \qquad m_c = \boldsymbol{w}^T \boldsymbol{\mu}_c = \frac{1}{n_c} \sum_{\boldsymbol{x}^n \in \mathcal{D}_c} \boldsymbol{w}^T \boldsymbol{x}^n$$

$$\triangleq \frac{1}{n_c} \sum_{\boldsymbol{x}^n \in \mathcal{D}_c} y^n$$

$$s_c^2 = \frac{1}{n_c} \sum_{\boldsymbol{x}^n \in \mathcal{D}_c} (y^n - m_c)^2$$



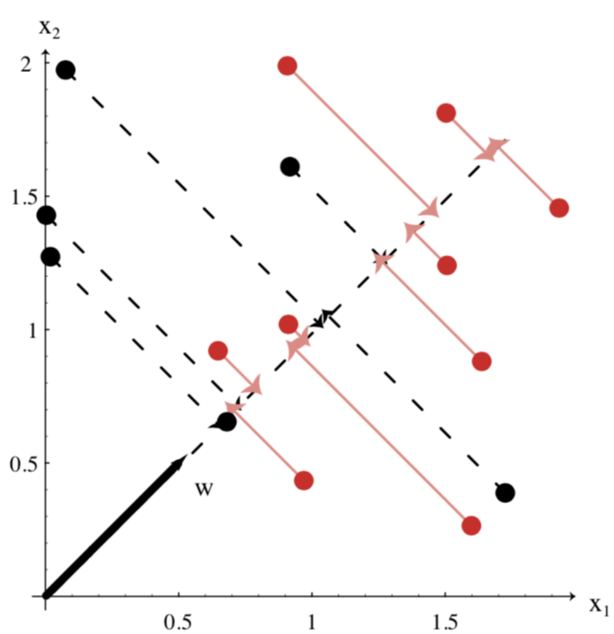


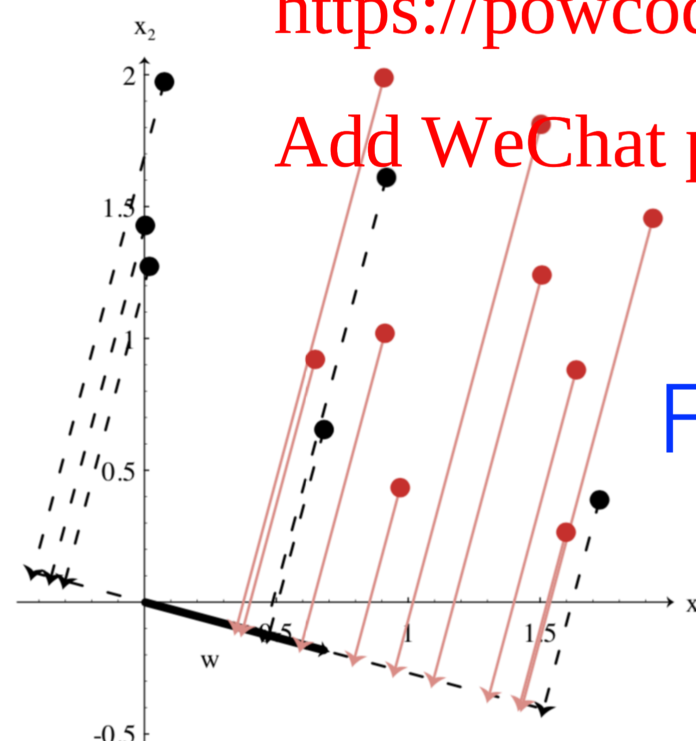
Fisher Linear Discriminant Analysis: projections along discriminative directions

$$\boldsymbol{\mu}_c = \frac{1}{n_c} \sum_{\boldsymbol{x}^n \in \mathcal{D}_c} \boldsymbol{x}^n \qquad m_c = \boldsymbol{w}^T \boldsymbol{\mu}_c = \frac{1}{n_c} \sum_{\boldsymbol{x}^n \in \mathcal{D}_c} \boldsymbol{w}^T \boldsymbol{x}^n$$

$$\triangleq \frac{1}{n_c} \sum_{\boldsymbol{x}^n \in \mathcal{D}_c} y^n$$

$$s_c^2 = \frac{1}{n_c} \sum_{\boldsymbol{x}^n \in \mathcal{D}_c} (y^n - m_c)^2$$





Fisher score:

$$J(\boldsymbol{w}) = \frac{(m_+ - m_-)^2}{n_+ s_+^2 + n_- s_-^2}$$

$$\text{Choose } \boldsymbol{w}^* = \operatorname{argmax}_{\boldsymbol{w}} J(\boldsymbol{w})$$

Rewrite Fisher score to make weight vector dependence explicit: between and within class scatter

$$s_c^2 = \frac{1}{n_c} \sum_{\boldsymbol{x}^n \in \mathcal{D}_c} (y^n - m_c)^2 = \frac{1}{n_c} \sum_{\boldsymbol{x}^n \in \mathcal{D}_c} (\boldsymbol{w}^T \boldsymbol{x}^n - \boldsymbol{w}^T \boldsymbol{\mu})^2$$

denominator:
within class scatter

numerator:
between class scatter

$$J(\boldsymbol{w}) = \frac{(m_+ - m_-)^2}{n_+ s_+^2 + n_- s_-^2} \implies \boxed{J(\boldsymbol{w}) = \frac{\boldsymbol{w}^T \boldsymbol{S}_B \boldsymbol{w}}{\boldsymbol{w}^T \boldsymbol{S}_W \boldsymbol{w}}}$$

Rewrite Fisher score to make weight vector dependence explicit: between and within class scatter

$(\boldsymbol{a} \cdot \boldsymbol{b})^2 = (\boldsymbol{b}^\top \boldsymbol{a})(\boldsymbol{a}^\top \boldsymbol{b})$

$$s_c^2 = \frac{1}{n_c} \sum_{\boldsymbol{x}^n \in \mathcal{D}_c} (y^n - m_c)^2 = \frac{1}{n_c} \sum_{\boldsymbol{x}^n \in \mathcal{D}_c} (\boldsymbol{w}^T \boldsymbol{x}^n - \boldsymbol{w}^T \boldsymbol{\mu})^2$$

denominator:
within class scatter

numerator:
between class scatter

$$J(\boldsymbol{w}) = \frac{(m_+ - m_-)^2}{n_+ s_+^2 + n_- s_-^2} \implies \boxed{J(\boldsymbol{w}) = \frac{\boldsymbol{w}^T \boldsymbol{S}_B \boldsymbol{w}}{\boldsymbol{w}^T \boldsymbol{S}_W \boldsymbol{w}}}$$

Rewrite Fisher score to make weight vector dependence explicit: between and within class scatter

$(\boldsymbol{a} \cdot \boldsymbol{b})^2 = (\boldsymbol{b}^\top \boldsymbol{a})(\boldsymbol{a}^\top \boldsymbol{b})$

$$
\begin{aligned}
s_c^2 &= \frac{1}{n_c} \sum_{\boldsymbol{x}^n \in \mathcal{D}_c} (y^n - m_c)^2 = \frac{1}{n_c} \sum_{\boldsymbol{x}^n \in \mathcal{D}_c} (\boldsymbol{w}^T \boldsymbol{x}^n - \boldsymbol{w}^T \boldsymbol{\mu})^2 \\
&= \frac{1}{n_c} \sum_{\boldsymbol{x}^n \in \mathcal{D}_c} \boldsymbol{w}^T (\boldsymbol{x}^n - \boldsymbol{\mu}) (\boldsymbol{x}^n - \boldsymbol{\mu})^T \boldsymbol{w}
\end{aligned}
$$

denominator:
within class scatter

numerator:
between class scatter

$$
J(\boldsymbol{w}) = \frac{(m_+ - m_-)^2}{n_+ s_+^2 + n_- s_-^2} \implies \boxed{J(\boldsymbol{w}) = \frac{\boldsymbol{w}^T \boldsymbol{S}_B \boldsymbol{w}}{\boldsymbol{w}^T \boldsymbol{S}_W \boldsymbol{w}}}
$$

Rewrite Fisher score to make weight vector dependence explicit: between and within class scatter

$(\boldsymbol{a}\cdot\boldsymbol{b})^2 = (\boldsymbol{b}^\top\boldsymbol{a})(\boldsymbol{a}^\top\boldsymbol{b})$

$$
\begin{aligned}
s_c^2 &= \frac{1}{n_c}\sum_{\boldsymbol{x}^n\in\mathcal{D}_c}(y^n - m_c)^2 = \frac{1}{n_c}\sum_{\boldsymbol{x}^n\in\mathcal{D}_c}(\boldsymbol{w}^T\boldsymbol{x}^n - \boldsymbol{w}^T\boldsymbol{\mu})^2 \\
&= \frac{1}{n_c}\sum_{\boldsymbol{x}^n\in\mathcal{D}_c}\boldsymbol{w}^T\left(\boldsymbol{x}^n - \boldsymbol{\mu}\right)\left(\boldsymbol{x}^n - \boldsymbol{\mu}\right)^T\boldsymbol{w} \\
&\triangleq \boldsymbol{w}^T\boldsymbol{S}\boldsymbol{w}
\end{aligned}
$$

denominator: within class scatter

numerator: between class scatter

$$
J(\boldsymbol{w}) = \frac{(m_+ - m_-)^2}{n_+s_+^2 + n_-s_-^2} \implies \boxed{J(\boldsymbol{w}) = \frac{\boldsymbol{w}^T\boldsymbol{S}_B\boldsymbol{w}}{\boldsymbol{w}^T\boldsymbol{S}_W\boldsymbol{w}}}
$$

Rewrite Fisher score to make weight vector dependence explicit: between and within class scatter

$(\boldsymbol{a}\cdot\boldsymbol{b})^2 = (\boldsymbol{b}^\top\boldsymbol{a})(\boldsymbol{a}^\top\boldsymbol{b})$

$$
\begin{aligned}
s_c^2 &= \frac{1}{n_c}\sum_{\boldsymbol{x}^n\in\mathcal{D}_c}(y^n - m_c)^2 = \frac{1}{n_c}\sum_{\boldsymbol{x}^n\in\mathcal{D}_c}(\boldsymbol{w}^T\boldsymbol{x}^n - \boldsymbol{w}^T\boldsymbol{\mu})^2 \\
&= \frac{1}{n_c}\sum_{\boldsymbol{x}^n\in\mathcal{D}_c}\boldsymbol{w}^T\left(\boldsymbol{x}^n - \boldsymbol{\mu}\right)\left(\boldsymbol{x}^n - \boldsymbol{\mu}\right)^T\boldsymbol{w} \\
&\triangleq \boldsymbol{w}^T\boldsymbol{S}\boldsymbol{w}
\end{aligned}
$$

$$
\begin{aligned}
n_+s_+^2 + n_-s_-^2 &= \boldsymbol{w}^T(n_+\boldsymbol{S}_+ + n_-\boldsymbol{S}_-)\boldsymbol{w} \\
&\triangleq \boldsymbol{w}^T\boldsymbol{S}_W\boldsymbol{w}
\end{aligned}
$$

denominator:
within class scatter

numerator:
between class scatter

$$
J(\boldsymbol{w}) = \frac{(m_+ - m_-)^2}{n_+s_+^2 + n_-s_-^2} \implies \boxed{J(\boldsymbol{w}) = \frac{\boldsymbol{w}^T\boldsymbol{S}_B\boldsymbol{w}}{\boldsymbol{w}^T\boldsymbol{S}_W\boldsymbol{w}}}
$$

Rewrite Fisher score to make weight vector dependence explicit: between and within class scatter

$(\boldsymbol{a}\cdot\boldsymbol{b})^2 = (\boldsymbol{b}^\top\boldsymbol{a})(\boldsymbol{a}^\top\boldsymbol{b})$

$$
\begin{aligned}
s_c^2 &= \frac{1}{n_c}\sum_{\boldsymbol{x}^n\in\mathcal{D}_c}(y^n - m_c)^2 = \frac{1}{n_c}\sum_{\boldsymbol{x}^n\in\mathcal{D}_c}(\boldsymbol{w}^T\boldsymbol{x}^n - \boldsymbol{w}^T\boldsymbol{\mu})^2 \\
&= \frac{1}{n_c}\sum_{\boldsymbol{x}^n\in\mathcal{D}_c}\boldsymbol{w}^T\left(\boldsymbol{x}^n - \boldsymbol{\mu}\right)\left(\boldsymbol{x}^n - \boldsymbol{\mu}\right)^T\boldsymbol{w} \\
&\triangleq \boldsymbol{w}^T\boldsymbol{S}\boldsymbol{w}
\end{aligned}
$$

$$
\begin{aligned}
n_+s_+^2 + n_-s_-^2 &= \boldsymbol{w}^T(n_+\boldsymbol{S}_+ + n_-\boldsymbol{S}_-)\boldsymbol{w} \\
&\triangleq \boldsymbol{w}^T\boldsymbol{S}_W\boldsymbol{w}
\end{aligned}
$$

denominator:
within class scatter

$$
\begin{aligned}
(m_+ - m_-)^2 &= (\boldsymbol{w}^T(\boldsymbol{\mu}_+ - \boldsymbol{\mu}_-))^2 \\
&= \boldsymbol{w}^T(\boldsymbol{\mu}_+ - \boldsymbol{\mu}_-)(\boldsymbol{\mu}_+ - \boldsymbol{\mu}_-)^T\boldsymbol{w}
\end{aligned}
$$

numerator:
between class scatter

$$
J(\boldsymbol{w}) = \frac{(m_+ - m_-)^2}{n_+s_+^2 + n_-s_-^2} \implies \boxed{J(\boldsymbol{w}) = \frac{\boldsymbol{w}^T\boldsymbol{S}_B\boldsymbol{w}}{\boldsymbol{w}^T\boldsymbol{S}_W\boldsymbol{w}}}
$$

Optimisation: find best direction, set gradients to 0

Optimisation: find best direction, set gradients to 0

Partial derivatives (gradient) set to 0

$$\frac{\partial J(\boldsymbol{w})}{\partial \boldsymbol{w}} = \frac{2}{(\boldsymbol{w}^T \boldsymbol{S}_W \boldsymbol{w})^2} \left((\boldsymbol{w}^T \boldsymbol{S}_W \boldsymbol{w}) \boldsymbol{S}_B \boldsymbol{w} - (\boldsymbol{w}^T \boldsymbol{S}_B \boldsymbol{w}) \boldsymbol{S}_W \boldsymbol{w} \right)$$

reduces to a **generalised eigenvalue problem**:

$$\boldsymbol{S}_B \boldsymbol{w} = \lambda \boldsymbol{S}_W \boldsymbol{w}$$

Optimisation: find best direction, set gradients to 0

$$\left(\frac{u}{v}\right)' = \frac{1}{v^2}\left(u'v - uv'\right)$$

Partial derivatives (gradient) set to 0

$$\frac{\partial J(\boldsymbol{w})}{\partial \boldsymbol{w}} = \frac{2}{(\boldsymbol{w}^T \boldsymbol{S}_W \boldsymbol{w})^2}\left((\boldsymbol{w}^T \boldsymbol{S}_W \boldsymbol{w})\boldsymbol{S}_B \boldsymbol{w} - (\boldsymbol{w}^T \boldsymbol{S}_B \boldsymbol{w})\boldsymbol{S}_W \boldsymbol{w}\right)$$

reduces to a **generalised eigenvalue problem**:

$$\boldsymbol{S}_B \boldsymbol{w} = \lambda \boldsymbol{S}_W \boldsymbol{w}$$

Optimisation: find best direction, set gradients to 0

$$\left(\frac{u}{v}\right)' = \frac{1}{v^2}\left(u'v - uv'\right)$$

Partial derivatives (gradient) set to 0

$$\frac{\partial J(\boldsymbol{w})}{\partial \boldsymbol{w}} = \frac{2}{(\boldsymbol{w}^T \boldsymbol{S}_W \boldsymbol{w})^2}\left((\underline{\boldsymbol{w}^T \boldsymbol{S}_W \boldsymbol{w}})\boldsymbol{S}_B \boldsymbol{w} - (\underline{\boldsymbol{w}^T \boldsymbol{S}_B \boldsymbol{w}})\boldsymbol{S}_W \boldsymbol{w}\right)$$

scalars

reduces to a **generalised eigenvalue problem**:

$$\boldsymbol{S}_B \boldsymbol{w} = \lambda \boldsymbol{S}_W \boldsymbol{w}$$

Optimisation: find best direction, set gradients to 0

$$\left(\frac{u}{v}\right)' = \frac{1}{v^2}\left(u'v - uv'\right)$$

Partial derivatives (gradient) set to 0

$$\frac{\partial J(\boldsymbol{w})}{\partial \boldsymbol{w}} = \frac{2}{(\boldsymbol{w}^T\boldsymbol{S}_W\boldsymbol{w})^2}\left((\underline{\boldsymbol{w}^T\boldsymbol{S}_W\boldsymbol{w}})\boldsymbol{S}_B\boldsymbol{w} - (\underline{\boldsymbol{w}^T\boldsymbol{S}_B\boldsymbol{w}})\boldsymbol{S}_W\boldsymbol{w}\right)$$

reduces to a **generalised eigenvalue problem**:

scalars

$$\boldsymbol{S}_B\boldsymbol{w} = \underline{\lambda}\boldsymbol{S}_W\boldsymbol{w}$$

ratio

Optimisation: find best direction, set gradients to 0

$$\left(\frac{u}{v}\right)' = \frac{1}{v^2}\left(u'v - uv'\right)$$

Partial derivatives (gradient) set to 0

$$\frac{\partial J(\boldsymbol{w})}{\partial \boldsymbol{w}} = \frac{2}{(\boldsymbol{w}^T \boldsymbol{S}_W \boldsymbol{w})^2}\left((\boldsymbol{w}^T \boldsymbol{S}_W \boldsymbol{w})\boldsymbol{S}_B \boldsymbol{w} - (\boldsymbol{w}^T \boldsymbol{S}_B \boldsymbol{w})\boldsymbol{S}_W \boldsymbol{w}\right)$$

reduces to a **generalised eigenvalue problem**:

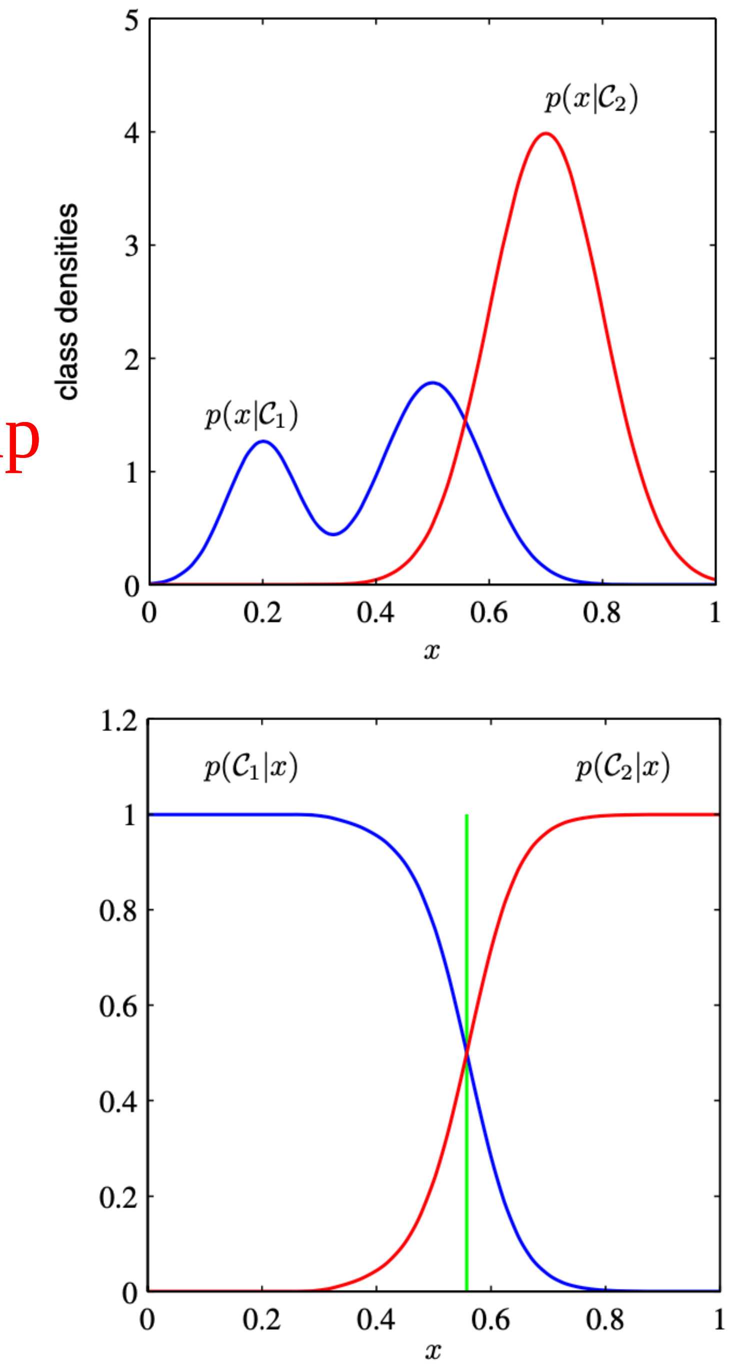
scalars

$$\boldsymbol{S}_B \boldsymbol{w} = \lambda \boldsymbol{S}_W \boldsymbol{w}$$

ratio

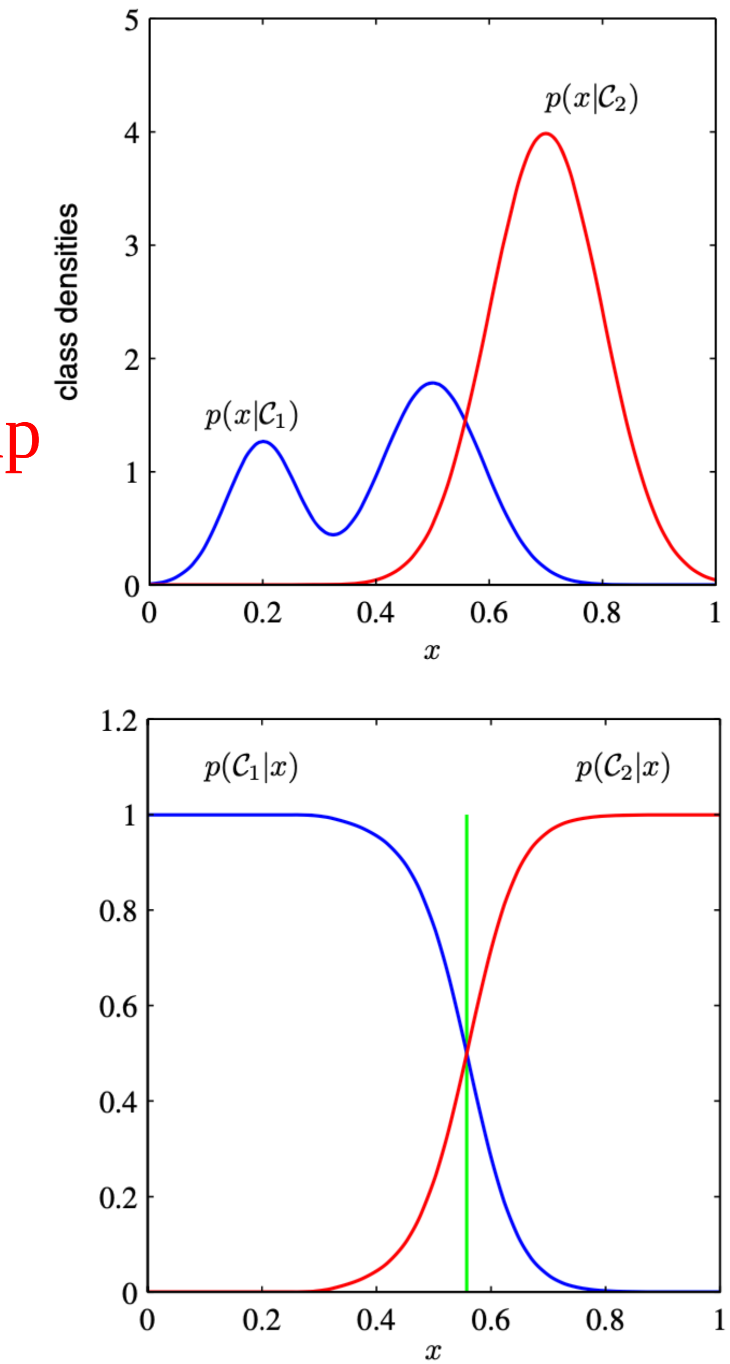
Extend to multiple classes, not just 2.

Classifiers can be generative or discriminative



Classifiers can be generative or discriminative

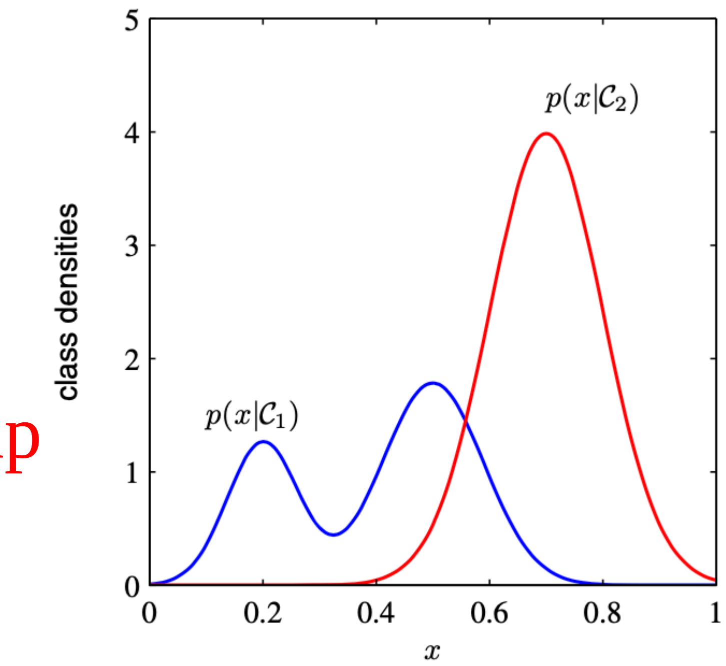
- Generative — class conditional probabilities can be used to simulate (generate) data

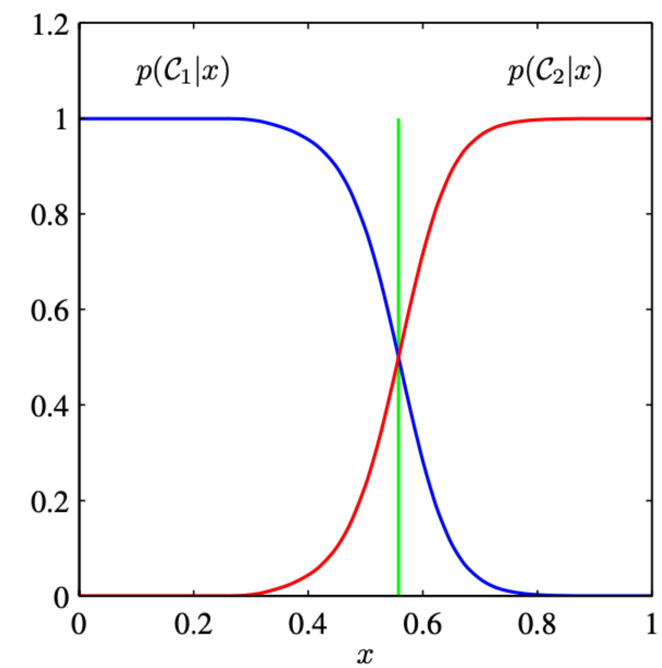


Classifiers can be generative or discriminative

- Generative — class conditional probabilities can be used to simulate (generate) data

- Generative — probabilities consistent with allocation of distance from decision boundaries



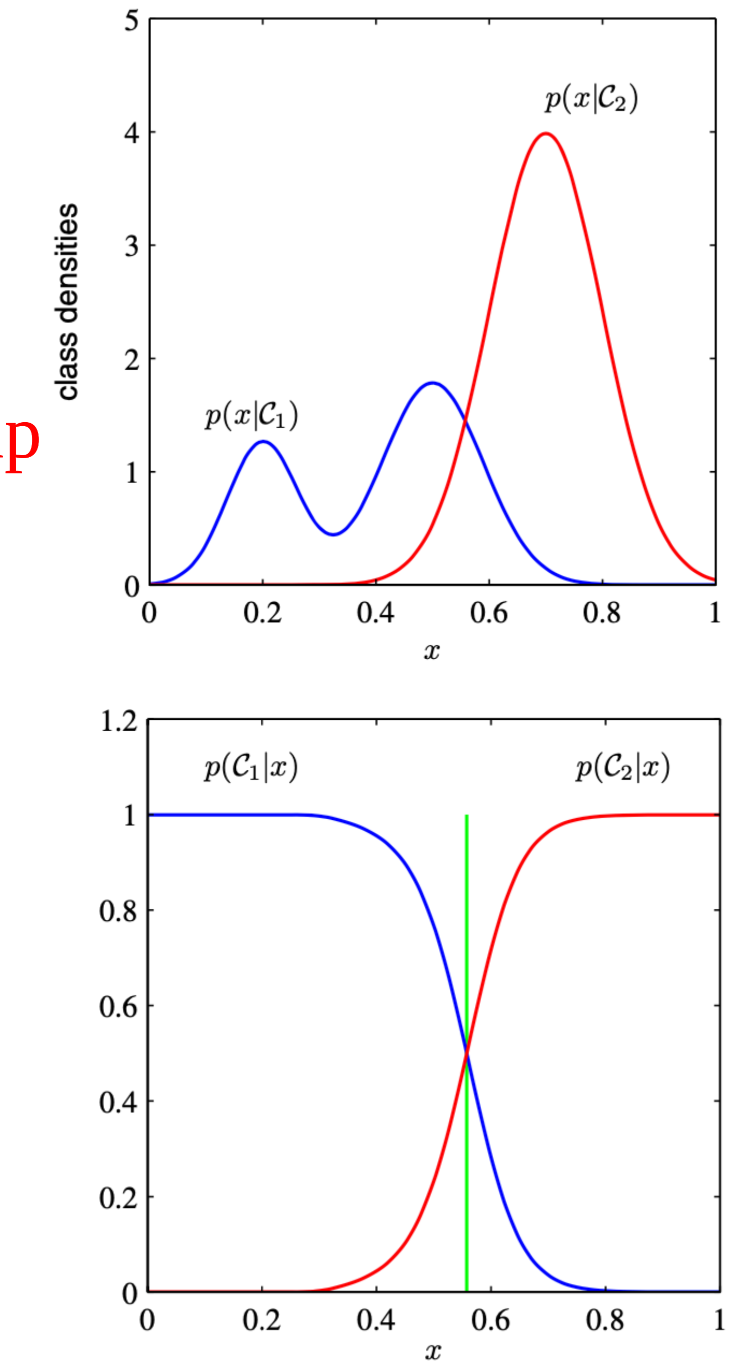


Classifiers can be generative or discriminative

- Generative — class conditional probabilities can be used to simulate (generate) data

- Generative — probabilities consistent with allocation of distance from decision boundaries

- Discriminative — does not model data within classes, only what distinguishes a class from another



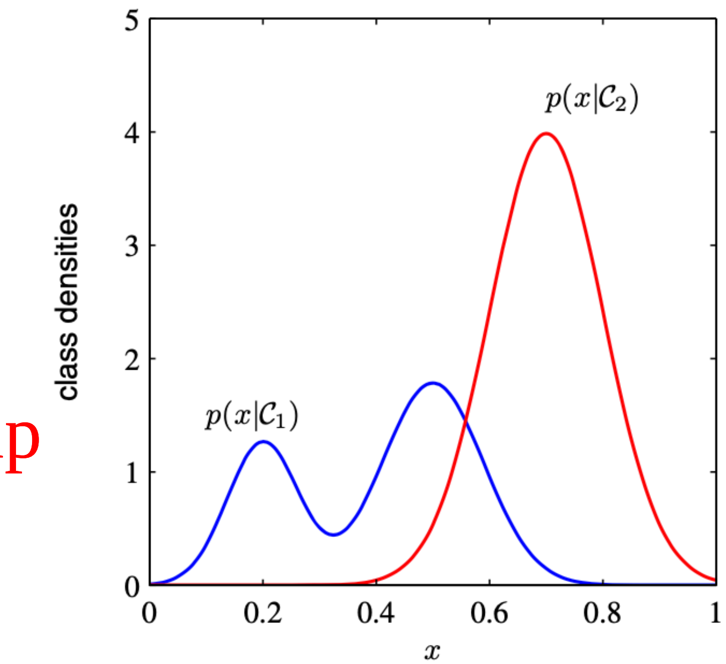
Classifiers can be generative or discriminative

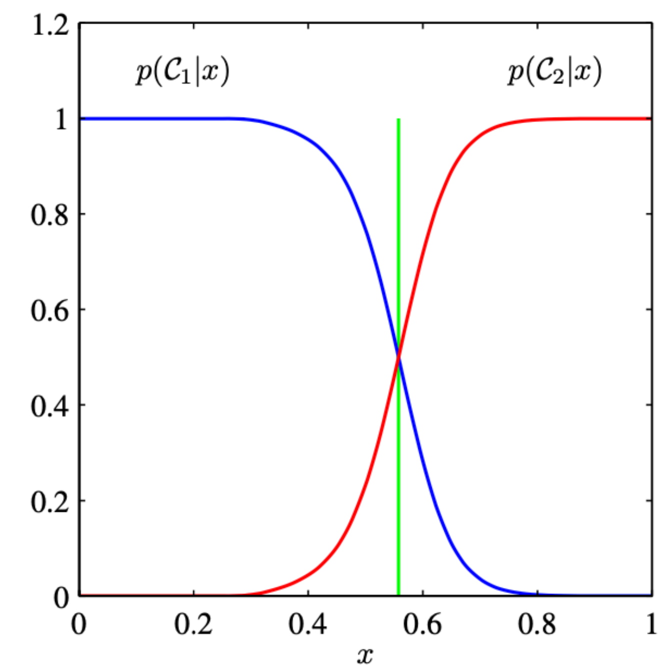
- Generative — class conditional probabilities can be used to simulate (generate) data

- Generative — probabilities consistent with allocation of distance from decision boundaries

- Discriminative — does not model data within classes, only what distinguishes a class from another

- Next time: maximum margin classifiers (Support Vector Machines)





References

- PRML, Bishop, Chapter 1, section 1.5
- PRML, Bishop, Chapter 4.
- FCML, Rogers-Girolami, Chapter 5.