Parameterschätzung (Fortsetzung)

Parameterschätzung

Beispiel:

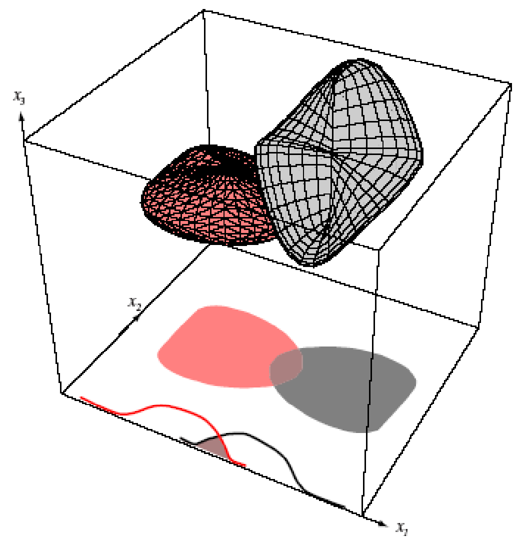


FIGURE 3.3. Two three-dimensional distributions have nonoverlapping densities, and thus in three dimensions the Bayes error vanishes. When projected to a subspace—here, the two-dimensional $x_1 - x_2$ subspace or a one-dimensional x_1 subspace—there can be greater overlap of the projected distributions, and hence greater Bayes error. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

In der Praxis leider manchmal Vergrößerung des Fehlers durch falsches Modell oder zu wenige Trainingsdaten

Überanpassung

Problem:

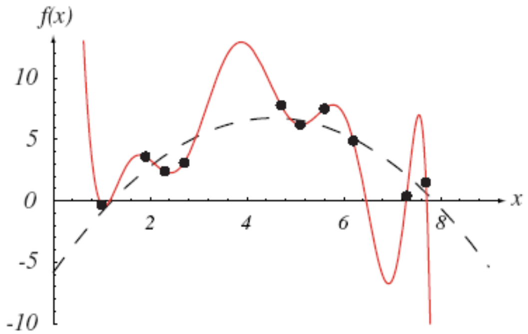


FIGURE 3.4. The "training data" (black dots) were selected from a quadratic function plus Gaussian noise, i.e., $f(x) = ax^2+bx+c+\epsilon$ where $p(\epsilon) \sim N(0, \sigma^2)$. The 10th-degree polynomial shown fits the data perfectly, but we desire instead the second-order function $f(x)$, because it would lead to better predictions for new samples. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

- Zu viele Parameter
- Zu wenige Testdaten

Lösung:

-> Vereinfachen

Überanpassung

Analoges Problem bei der Schätzung mehrdimensionaler Gaußverteilungen:

- Viele Dimensionen = viele Parameter $(\boldsymbol{\mu}_k, \mathbf{C}_k)$
- Kovarianz kritisch bei wenigen Daten (zufällige Korrelationen)

Lösungsmöglichkeiten:

- Weniger (u. vielleicht bessere) Features
- Reduktion der Parameter durch Annahme: Alle Klassen haben gleiche Kovarianz $(\mathbf{C}_k = \mathbf{C})$

Überanpassung

Lösungsmöglichkeiten (Forts.):

- Bessere Schätzung der Kovarianzen durch vernünftigen Prior $\mathbf{C}_{k,0}$:

$$\lambda \mathbf{C}_0 + (1-\lambda)\hat{\mathbf{C}}_k$$

- Begrenzung der Kovarianzen auf Maximalwert, im Extremfall auf 0.
- Shrinkage: Kombination aus gemeinsamer und klassenspezifischer Kovarianz:

$$\mathbf{C}_k(\alpha) = \frac{(1-\alpha)n_k\hat{\mathbf{C}}_k + \alpha n\hat{\mathbf{C}}}{(1-\alpha)n_k + \alpha n}$$

Dimensionsreduktion

Linearkombination von Merkmalen

Dimensionsreduktion durch Linearkombination von Merkmalen.

- Linearkombination = Projektion

Dazu notwendig ist eine

- Analyse der Trainingsdaten

Wir besprechen:

- PCA: Principal Component Analysis
- MDA: Multiple Discriminant Analysis

PCA: Principal Component Analysis

Gesucht:

Linearkombinationen, die die Daten am besten repräsentieren (im Sinne eines quadratischen Fehlers).

Zunächst: Reduktion auf 0 Dimensionen = Repräsentation von $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ durch einen Vektor $\mathbf{x}_0$

Quadratischer Fehler:

$$J_0\left(\mathbf{x}_0\right) = \sum_{k=1}^{n} \left\| \mathbf{x}_0 - \mathbf{x}_k \right\|^2$$

PCA: Principal Component Analysis

Minimiere $J_0(\mathbf{x}_0)$

$$\left.\frac{\partial J_0(\mathbf{x}_0)}{\partial \mathbf{x}_0}\right|_{\mathbf{m}} = \left.\frac{\partial \sum_{k=1}^{n} \left\| \mathbf{x}_0 - \mathbf{x}_k \right\|^2}{\partial \mathbf{x}_0}\right|_{\mathbf{m}} \overset{!}{=} 0$$

Lösung:

$$\mathbf{m} = \frac{1}{n} \sum_{k=1}^{n} \mathbf{x}_k$$

PCA: Principal Component Analysis

Erweiterung auf eine Dimension:

Repräsentation durch Linie durch $\mathbf{m}$: $\mathbf{x} = \mathbf{m} + a\mathbf{e}$

Quadratischer Fehler:

$$J_1(a_1, \ldots, a_n, \mathbf{m}, \mathbf{e}) = \sum_{k=1}^{n} \left\| \mathbf{m} + a_k \mathbf{e} - \mathbf{x}_k \right\|^2$$

Lösung: $a_k = \mathbf{e}^T (\mathbf{x}_k - \mathbf{m})$

PCA: Principal Component Analysis

Aber: Welche Richtung $\mathbf{e}$ hat die Linie?

Dazu setzen wir die $a_k = \mathbf{e}^T(\mathbf{x}_k - \mathbf{m})$ in die Formel

für den Fehler $J_1(a_1, \ldots, a_n, \mathbf{m}, \mathbf{e}) = \sum_{k=1}^{n} \left\| \mathbf{m} + a_k \mathbf{e} - \mathbf{x}_k \right\|^2$ ein:

$$
\boxed{
\begin{aligned}
J_1(a_1, \ldots, a_n, \mathbf{m}, \mathbf{e}) &= -\sum_{k=1}^{n} \mathbf{e}^T (\mathbf{x}_k - \mathbf{m})(\mathbf{x}_k - \mathbf{m})^T \mathbf{e} + \sum_{k=1}^{n} \left\| \mathbf{x}_k - \mathbf{m} \right\|^2 \\
&= -\mathbf{e}^T \mathbf{S} \mathbf{e} + \sum_{k=1}^{n} \left\| \mathbf{x}_k - \mathbf{m} \right\|^2
\end{aligned}
}
$$

mit $\mathbf{S} = \sum_{k=1}^{n} (\mathbf{x}_k - \mathbf{m})(\mathbf{x}_k - \mathbf{m})^T$ (Streumatrix)

PCA: Principal Component Analysis

$J_1(a_1, \ldots, a_n, \mathbf{m}, \mathbf{e})$ wird minimal, wenn $\mathbf{e}^T \mathbf{S} \mathbf{e}$ maximal, daher ist $\mathbf{e}$ Eigenvektor zu $\mathbf{S}$ mit größtem Eigenwert λ_0 :

$$\mathbf{S}\mathbf{e} = \lambda_0 \mathbf{e}$$

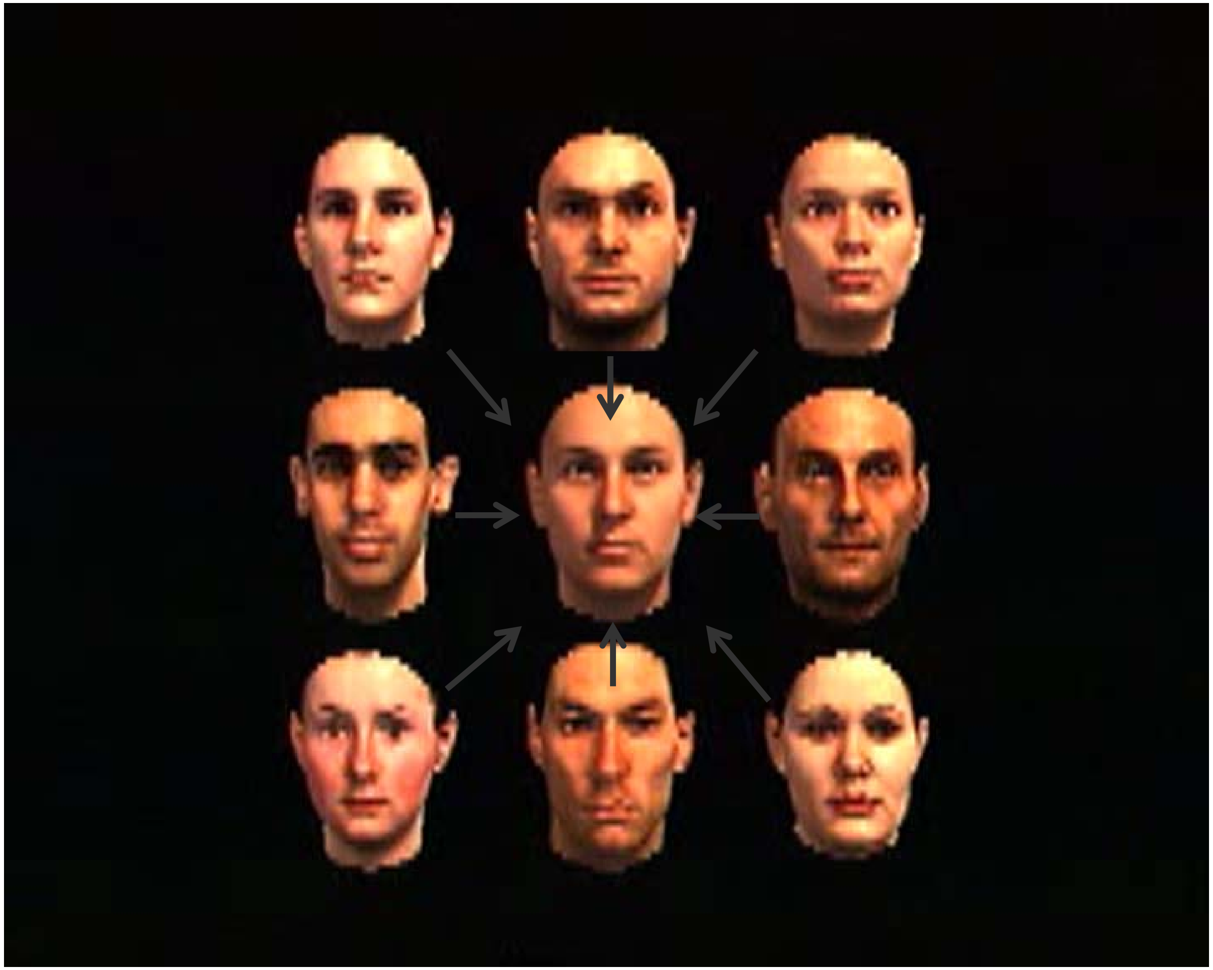
Erweiterung auf d' Dimensionen:

$$\mathbf{x} = \mathbf{m} + \sum_{i=1}^{d'} a_i \mathbf{e}_i$$

Lösung: $\mathbf{e}_i$ sind Eigenvektoren der Streumatrix, geordnet nach Größe der Eigenwerte.

PCA auf Köpfen

Beispiel:



PCA ist gut für Repräsentation.

Ist sie auch gut für Klassifizierung?

Oft Nein!

Wie findet man gute Repräsentation für Unterscheidung?

Wir betrachten eindimensionale Projektion

$$y = \mathbf{w}^T \mathbf{x}$$

Aus den Daten $\mathbf{x}_i$, die entweder zu $\mathcal{D}_1$ oder zu $\mathcal{D}_2$ gehören, werden also die Zahlen y_i

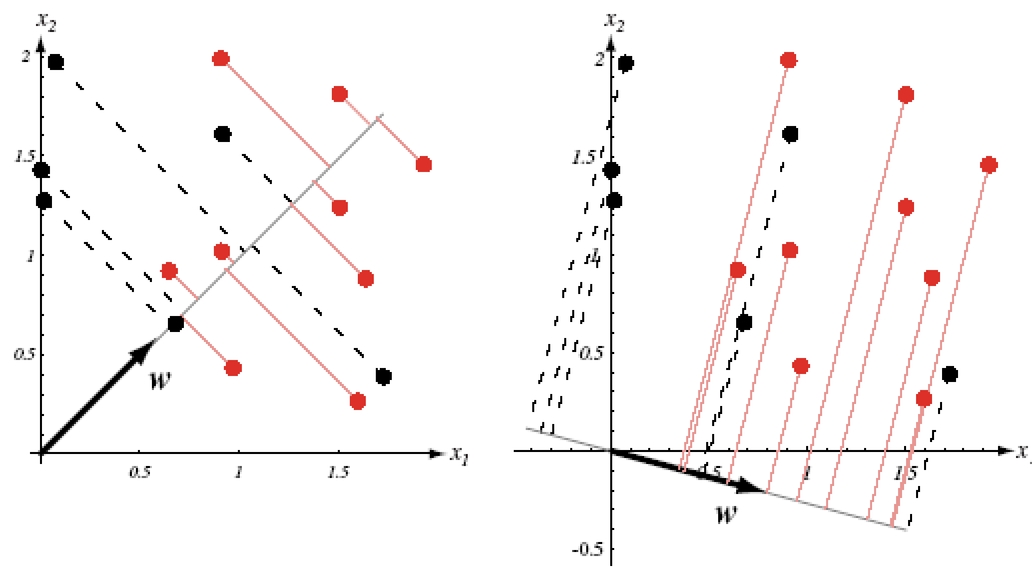


FIGURE 3.5. Projection of the same set of samples onto two different lines in the directions marked **w**. The figure on the right shows greater separation between the red and black projected points. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Kriterium: Abstand der Mittelwerte $\tilde{m}_1 - \tilde{m}_2$?
Nein, sondern
Verhältnis zwischen Abstand und Streuung:

$$J(\mathbf{w}) = \frac{\left(\tilde{m}_1 - \tilde{m}_2\right)^2}{\tilde{s}_1^2 + \tilde{s}_2^2}$$

Fisher linear discriminant

Mit $$\tilde{s}_i^2 = \sum_{y \in \mathcal{Y}_i} \left(y - \tilde{m}_i\right)^2$$

Fisher linear discriminant

Nun brauchen wir wieder die Streumatrizen:

$$\mathbf{S}_i = \sum_{\mathbf{x} \in \mathcal{D}_i} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^T$$

Wir definieren außerdem:

$$\mathbf{S}_W = \mathbf{S}_1 + \mathbf{S}_2$$ (within class scatter)

Dann gilt: $$\tilde{s}_i^2 = \sum_{\mathbf{x} \in \mathcal{D}_i} \left(\mathbf{w}^T \mathbf{x} - \mathbf{w}^T \mathbf{m}_i\right)^2 = \mathbf{w}^T \mathbf{S}_i \mathbf{w}$$

Daraus folgt: $$\tilde{s}_1^2 + \tilde{s}_2^2 = \mathbf{w}^T \mathbf{S}_W \mathbf{w}$$

Fisher linear discriminant

Genauso können wir für die Mittelwerte schreiben:

$$(\tilde{m}_1 - \tilde{m}_2)^2 = \left(\mathbf{w}^T\mathbf{m}_1 - \mathbf{w}^T\mathbf{m}_2\right)^2 = \mathbf{w}^T\mathbf{S}_B\mathbf{w}$$

Mit $\mathbf{S}_B = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T$ (Between class scatter)

Wir müssen also maximieren

$$J(\mathbf{w}) = \frac{\mathbf{w}^T\mathbf{S}_B\mathbf{w}}{\mathbf{w}^T\mathbf{S}_W\mathbf{w}}$$

Fisher linear discriminant

Die Lösung gehorcht der Bedingung

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

Wenn $\mathbf{S}_W$ invertiert werden kann gilt: $\mathbf{S}_W^{-1} \mathbf{S}_B \mathbf{w} = \lambda \mathbf{w}$

Da $\mathbf{S}_B = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T$

bekommen wir die Lösung

$$\mathbf{w} = \mathbf{S}_W^{-1} (\mathbf{m}_1 - \mathbf{m}_2)$$

Multiple Discriminant Analysis

Multiple discriminant analysis

Wir wollen ein d-dimensionales Problem mit c Klassen auf $c-1$ Dimensionen reduzieren.
Wie muß projiziert werden, so daß das Ergebnis wieder gut für die Unterscheidung geeignet ist?
Analog zum Vorgehen bei der Fisher linear discriminant definieren wir:

$$\mathbf{S}_W = \sum_{i=1}^{c} \mathbf{S}_i$$ (Within-class scatter matrix)

wobei wieder: $\mathbf{S}_i = \sum_{\mathbf{x} \in \mathcal{D}_i} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^T$ und $\mathbf{m}_i = \frac{1}{n_i} \sum_{\mathbf{x} \in \mathcal{D}_i} \mathbf{x}$

Multiple discriminant analysis

Außerdem betrachten wir einen Gesamtmittelwert

$$\mathbf{m} = \frac{1}{n}\sum_{\mathbf{x}} \mathbf{x} = \frac{1}{n}\sum_{i=1}^{c} n_i \mathbf{m}_i$$

und eine Gesamtstreumatrix

$$\mathbf{S}_T = \sum_{\mathbf{x}} (\mathbf{x}-\mathbf{m})(\mathbf{x}-\mathbf{m})^T$$ (Total scatter matrix)

Multiple discriminant analysis

Wir erhalten:

$$\begin{aligned}
\mathbf{S}_T &= \sum_{i=1}^{c} \sum_{\mathbf{x}\in\mathcal{D}_i} (\mathbf{x}-\mathbf{m}_i+\mathbf{m}_i-\mathbf{m})(\mathbf{x}-\mathbf{m}_i+\mathbf{m}_i-\mathbf{m})^T \\
&= \sum_{i=1}^{c} \sum_{\mathbf{x}\in\mathcal{D}_i} (\mathbf{x}-\mathbf{m}_i)(\mathbf{x}-\mathbf{m}_i)^T + \sum_{i=1}^{c} \sum_{\mathbf{x}\in\mathcal{D}_i} (\mathbf{m}_i-\mathbf{m})(\mathbf{m}_i-\mathbf{m})^T \\
&= \mathbf{S}_W + \sum_{i=1}^{c} n_i (\mathbf{m}_i-\mathbf{m})(\mathbf{m}_i-\mathbf{m})^T = \mathbf{S}_W + \mathbf{S}_B
\end{aligned}$$

wobei $\mathbf{S}_B$ als generalisierte Between-class scatter matrix bezeichnet wird.

Multiple discriminant analysis

Es gilt also: $$\mathbf{S}_B = \sum_{i=1}^{c} n_i \left(\mathbf{m}_i - \mathbf{m}\right)\left(\mathbf{m}_i - \mathbf{m}\right)^T$$

und $$\mathbf{S}_T = \mathbf{S}_W + \mathbf{S}_B$$

(Diese Formel ist generell interessant zur Berechnung von Streumatrizen)

Multiple discriminant analysis

Die gesuchte Dimensionsreduktion wollen wir als Projektion darstellen:

$$y_i = \mathbf{w}_i^T \mathbf{x} \qquad\qquad i = 1, \ldots, c-1$$

bzw.

$$\mathbf{y} = \begin{pmatrix} \mathbf{w}_1^T \\ \vdots \\ \mathbf{w}_{c-1}^T \end{pmatrix} \mathbf{x} = \mathbf{W}^T \mathbf{x}$$

Multiple discriminant analysis

Mittelwerte und Streumatrizen im dimensionsreduzierten Raum der $\mathbf{y}$ nennen wir $\tilde{\mathbf{m}}_i, \tilde{\mathbf{m}}, \tilde{\mathbf{S}}_W, \tilde{\mathbf{S}}_B$

Man kann nun zeigen, dass

$$\tilde{\mathbf{S}}_W = \mathbf{W}^T \mathbf{S}_W \mathbf{W}$$

und

$$\tilde{\mathbf{S}}_B = \mathbf{W}^T \mathbf{S}_B \mathbf{W}$$

Gesucht ist nun eine Matrix $\mathbf{W}$, die das Verhältnis von Streuung zwischen den Klassen $\tilde{\mathbf{S}}_B$ zur Streuung innerhalb der Klassen $\tilde{\mathbf{S}}_W$ maximiert

Multiple discriminant analysis

Als Maß für dieses Verhältnis benutzen wir das Verhältnis der Determinanten der Streumatrizen:

$$J(\mathbf{W}) = \frac{\left|\tilde{\mathbf{S}}_B\right|}{\left|\tilde{\mathbf{S}}_W\right|} = \frac{\left|\mathbf{W}^T \mathbf{S}_B \mathbf{W}\right|}{\left|\mathbf{W}^T \mathbf{S}_W \mathbf{W}\right|}$$

Es kann gezeigt werden, daß die Lösung aus einer Matrix $\mathbf{W}$ besteht, deren Spalten $\mathbf{w}_i$ die generalisierten Eigenvektoren zu den größten Eigenwerten in folgender Gleichung sind:

$$\mathbf{S}_B \mathbf{w}_i = \lambda_i \mathbf{S}_W \mathbf{w}_i$$

Multiple discriminant analysis

Diese bestimmt man durch Suchen der Lösungen der Polynomgleichung

$$\left|\mathbf{S}_B - \lambda_i \mathbf{S}_W\right| = 0$$

und dann Einsetzen dieser Lösungen in die Gleichung

$$\left(\mathbf{S}_B - \lambda_i \mathbf{S}_W\right)\mathbf{w}_i = 0$$

Multiple discriminant analysis

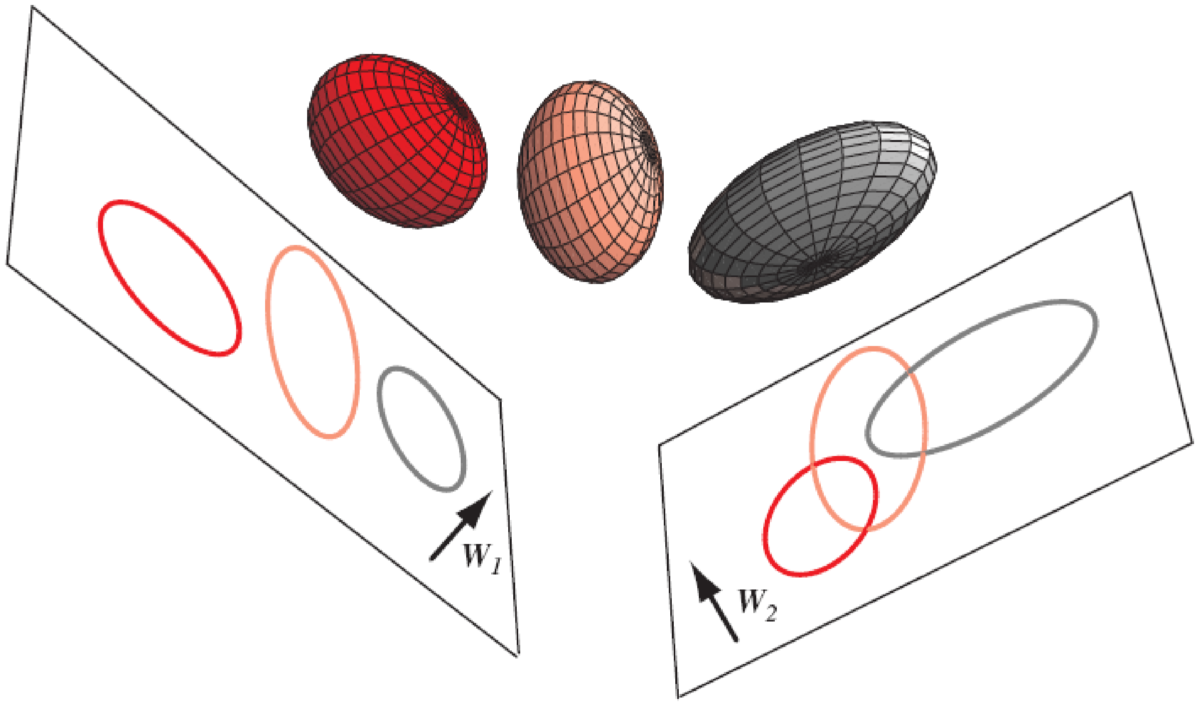


FIGURE 3.6. Three three-dimensional distributions are projected onto two-dimensional subspaces, described by a normal vectors $\mathbf{W}_1$ and $\mathbf{W}_2$. Informally, multiple discriminant methods seek the optimum such subspace, that is, the one with the greatest separation of the projected distributions for a given total within-scatter matrix, here as associated with $\mathbf{W}_1$. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*.