encontramos que $\hat{e}$ tiene distribución Binomial con valor medio e (como se ve calculando la esperanza) y por lo tanto $\hat{e}$ es un estimador insesgado del error medio e.

La desviación estándar de este estimador se calcula como

$$\sigma_{\hat{e}} = \sqrt{\mathrm{Var}\left[\hat{e}\right]} = \left(E\left[\hat{e}^2\right] - e^2\right)^{\frac{1}{2}}$$

Sustituyendo $\hat{e}$ y desarrollando resulta

$$\mathrm{Var}\left[\hat{e}\right] = \frac{1}{N^2}\sum_{j=1}^{N}\sum_{k=1}^{N} E\left[\eta_j \eta_k\right] - e^2 = \frac{1}{N} E\left[\eta^2\right] + \frac{N-1}{N} e^2 - e^2 = \frac{1}{N} e(1-e)$$

de donde la desviación del estimador será

$$\sigma_{\hat{e}} = \sqrt{\frac{e(1-e)}{N}} \leq \frac{1}{2\sqrt{N}} \,.$$

Resumiendo, hemos visto que podemos estimar el error medio de clasificación e presentándole a nuestro sistema de clasificación un conjunto de patrones que pertenecen a clases conocidas. El error se estima contando el número de discrepancias entre la clase verdadera y la etiqueta de clase asignada por el sistema, y dividiendo finalmente este resultado entre el número de muestras en la prueba.

Notar que si el error medio del sistema es pequeño, digamos de $\approx 1\,\%$, vamos a necesitar de un número grande de muestras de prueba para verificar este valor de desempeño con una razonable confianza relativa.

3. Apendizaje no supervisado

Es común encontrarse con situaciones en las que el sistema de clasificación de patrones debe diseñarse partiendo de un conjunto de patrones de entrenamiento $\{x_j\,; \quad j = 1, 2, \cdots, N\}$ para los cuales no conocemos sus etiquetas de clase γ_i.

Estas situaciones se presentan cuando no disponemos del conocimiento de un experto o bien cuando el etiquetado de cada muestra individual es impracticable. Esto último ocurre por ejemplo en el caso de aplicaciones con sensores remotos, como ser imágenes satelitales de terrenos donde sería muy costoso o imposible recoger información real del tipo de suelo sensado en cada punto de las imágenes. En estos casos el proceso de diseño requiere una primera etapa de análisis de las estructuras presentes en los datos de entrenamiento.

3.1. Aprendizaje no supervisado y análisis de agrupamientos

Dado un conjunto de entrenamiento suficientemente grande podemos inferir la función densidad de probabilidad conjunta $p(\mathbf{x})$ y recordando que

$$p(\mathbf{x}) = \sum_{i=1}^{m} P(\omega_i) p(\mathbf{x}|\omega_i)$$

podemos deducir que si la densidad conjunta es multimodal cada uno de los modos debería corresponderse con la distribución condicional de cada una de las clases presentes. Por lo tanto identificando estos modos en $p(\mathbf{x})$ sería en principio posible particionar el espacio de observación en regiones disjuntas $\Gamma_i\,, i = 1, \cdots, m$ asociadas con cada una de las clases presentes.

Si las distribuciones condicionales de cada clase son normales cabria la posibilidad de recuperar los parámetros de cada distribución a partir del conjunto de entrenamiento. A partir de esto podríamos seguir con el diseño del clasificador como se vio en la sección anterior. Sin embargo podemos conformarnos con recobrar

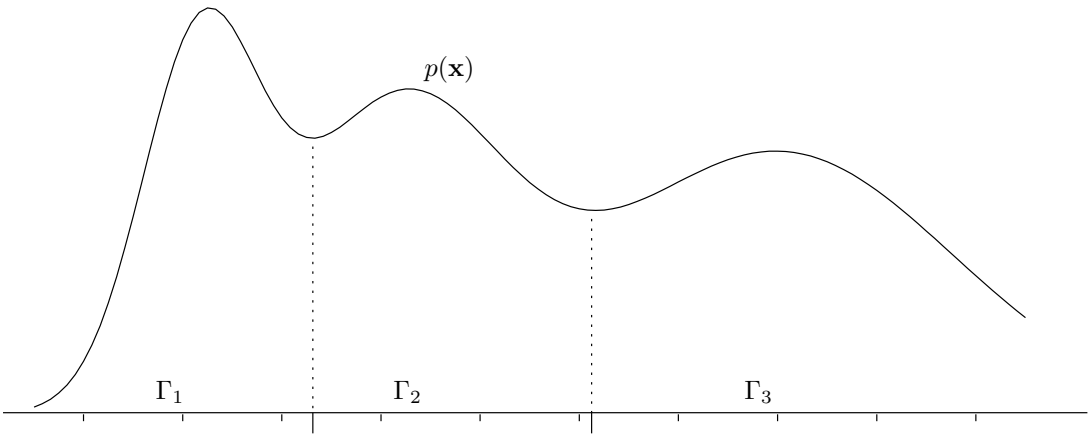

Figura 6: Distribución conjunta multimodal y regiones asociadas a cada clase..

directamente las regiones Γ_i lo cual es suficiente para nuestros intereses ya que esto puede usarse directamente para la clasificación de nuevos datos simplemente usando el criterio:

$$\text{Asignar } \mathbf{x} \text{ a } \omega_j \iff \mathbf{x} \in \Gamma_j$$

Un opción alternativa seria usar este u otro criterio para clasificar los patrones en el conjunto de entrenamiento y luego usar estas etiquetas para diseñar el sistema de reconocimiento de patrones usando un aprendizaje supervisado. En la práctica ocurre que determinar explicitamente las regiones Γ_i implicaría estimar la función de densidad conjunta y luego analizarla en un espacio de dimensión n lo que generalmente es impracticable por su complejidad computacional. Además como vimos, solo necesitamos de un método indirecto que nos permita etiquetar automáticamente los patrones de entrenamiento. Entonces lo que queremos es alguna forma de hacer una partición del conjunto de entrenamiento en clases con una misma etiqueta y esto es lo que se conoce como métodos de agrupamiento o *clustering*.

Intuitivamente podemos anticipar que las modas en la función de densidad conjunta $p(\mathbf{x})$ estarán asociadas a regiones con alta densidad de muestras en el espacio de observación. El proposito de las técnicas de agrupamiento será justamente detectar y agrupar estos *enjambres* de puntos.

3.2. Medidas de Similitud y Criterios de Agrupamiento

El propósito de los métodos de agrupamiento será analizar y extraer la estructura presente en un conjunto de patrones o muestras de entrenamiento. Diremos que un conjunto de datos está bien estructurado si contiene varios enjambres de patrones cercanos entre si, o sea regiones de alta densidad, separados por otras regiones relativamente vacias o con poca densidad.

Vemos que los puntos de un mismo agrupamiento apareceran más proximos entre ellos que a puntos en otros agrupamientos. Esta observación nos lleva a concluir que si queremos decidir si un punto $\mathbf{x}$ pertenece o no a un agrupamiento necesitaremos una medida de proximidad o similitud. Se han sugerido y estudiado un gran número de tales medidas, pero probablemente las más comunmente usadas son las medidas de distancia y en particular la distancia Euclideana.

La afinidad de un punto a un agrupamiento se puede determinar ya sea midiendo su similitud con otros puntos en el agrupamiento o bien con un modelo definido para el agrupamiento. El ejemplo más sencillo de esto último es representar un agrupamiento i por su vector medio μ_i; en este caso la afinidad entre un puntox y el agrupamiento se puede cuantificar con la distancia Euclideana al cuadrado

$$d(\mathbf{x}, \mu_i) = \left[(\mathbf{x} - \mu_i)^T (\mathbf{x} - \mu_i)\right]$$

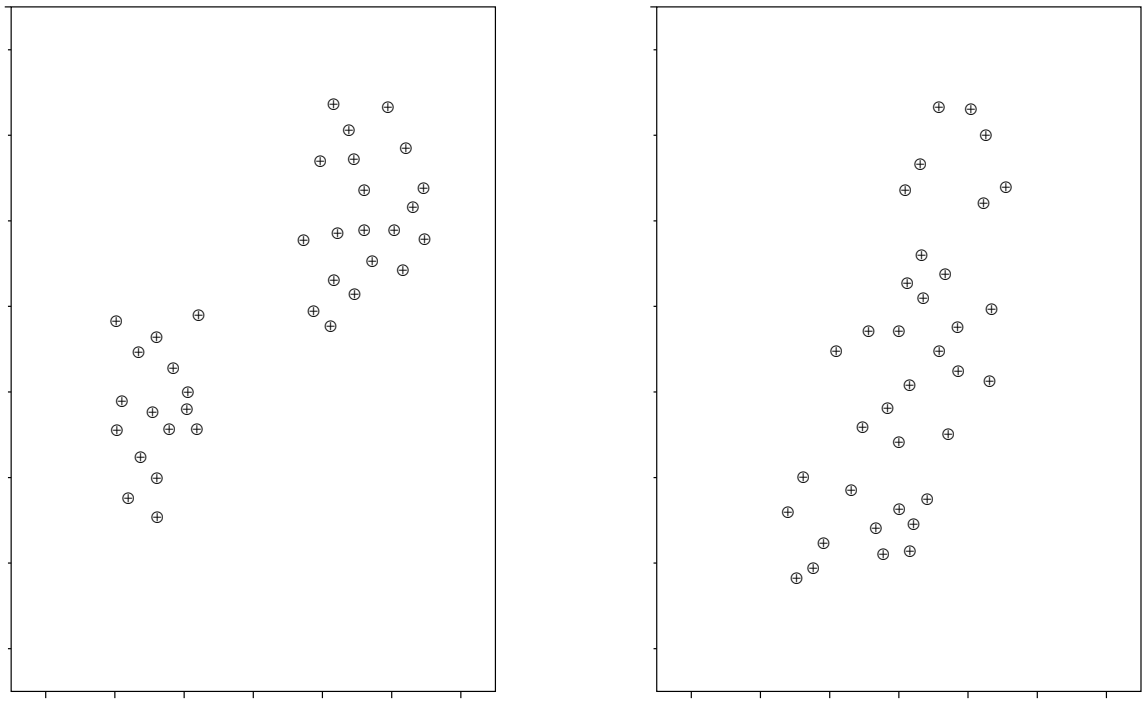

Figura 7: Datos structurados vs. no estructurados.

Pero para particionar un conjunto de puntos en agrupamientos de una manera óptima no nos alcanza con una medida de afinidad o similitud sino que además necesitamos algun *criterio de agrupamiento* que nos permita definir cuantitativamente cuando una partición es mejor que otra. Obviamente tanto el criterio de agrupamiento que definamos tanto como el algoritmo de agrupamiento asociado, estarán intimamente relacionados con la medida de similitud usada y se definirán a partir de esta.

En la siguiente sección veremos algunos ejemplos de métodos de agrupamiento que se basan en los conceptos anteriores.

3.3. Algoritmo de k-medias (k-means).

Supondremos que el conjunto de datos $\mathbf{X}$ contiene k agrupamientos y que cada uno de estos subconjuntos $\mathbf{X}_i$ puede representarse adecuadamente con su valor medio μ_i. Como se menciona anteriormente, en este caso podemos usar la distancia Euclideana como una medida de similitud. Se deduce que un criterio de agrupamiento adecuado en este caso es considerar la suma total sobre el conjunto de entrenamiento de la distancia cuadrática de cada punto al vector valor medio de su agrupamiento.

El objetivo del algoritmo de agrupamiento será encontrar entre todas las particiones de $\mathbf{X}$ en k conjuntos $\{\mathbf{X}_i\ ;\quad i = 1, 2, \cdots, k\}$ aquella que minimice el criterio de agrupamiento elegido.

Dicho formalmente, queremos encontrar los agrupamientos $\{\mathbf{X}_i\}$ que minimizan la función

$$J = \sum_{i=1}^{k} J_i = \sum_{i=1}^{k} \sum_{j=1}^{N_i} d(\mathbf{x}_{ij}, \mu_i) \quad \text{siendo} \quad \mathbf{x}_{ij} \in \mathbf{X}_i,\ N_i = \#\mathbf{X}_i$$

entre todas las posibles particiones de $\mathbf{X}$ en k subconjuntos.

Un algoritmo para minimizar J puede deducirse considerando el efecto de un cambio minimal o atómico en la configuración de agrupamientos, que consiste en sacar un punto $\mathbf{x}$ que este en el agrupamiento $\mathbf{X}_l$ para pasarlo a otro agrupamiento $\mathbf{X}_r$.

Claramente esta reasignación afectara solo a los agrupamientos l y r cuyos valores medios pasarán a ser

$$\bar{\mu}_l = \mu_l + \frac{1}{N_l - 1}(\mu_l - \mathbf{x}) \quad \text{y} \quad \bar{\mu}_r = \mu_r - \frac{1}{N_r + 1}(\mu_r - \mathbf{x})$$

respectivamente.

Para deducir la primera ecuación calculamos el valor medio de $\mathbf{X}_i$ antes y despues de la reasignación

$$\mu_l = \frac{1}{N_l}\sum_{j=1}^{N_l} \mathbf{x}_j \qquad \bar{\mu}_l = \frac{1}{N_l - 1}\sum_{j=1}^{N_l - 1} \mathbf{x}_j = \frac{1}{N_l - 1}\left(\sum_{j=1}^{N_l} \mathbf{x}_j - \mathbf{x}\right)$$

donde hemos asumido que el punto reasignado es el último en la sumatoria. De aqui resulta que

$$(N_l - 1)\bar{\mu}_l = N_l\,\mu_l - \mathbf{x} \quad \Rightarrow \quad \bar{\mu}_l = \frac{N_l}{N_l - 1}\,\mu_l - \frac{1}{N_l - 1}\,\mathbf{x} \quad \Rightarrow \quad \bar{\mu}_l = \mu_l + \frac{1}{N_l - 1}(\mu_l - \mathbf{x})$$

y análogamente se verifica la segunda identidad.

Por lo tanto para calcular el cambio global en el valor de J bastará calcular los cambios en las contribuciones de J_l y J_r. Para el nuevo agrupamiento l-esimo tendremos

$$\begin{aligned}
\bar{J}_l &= \sum_{j=1}^{N_l - 1} d(\mathbf{x_j}, \mathbf{mu}_l) = \sum_{j=1}^{N_l - 1} (\mathbf{x}_j - \mu_l)^T(\mathbf{x}_j - \mu_l) = \\
&= \sum_{j=1}^{N_l} \big(\mathbf{x}_j - \mu_l + \frac{\mu_l - \mathbf{x}}{N_l - 1}\big)^T \big(\mathbf{x}_j - \mu_l + \frac{\mu_l - \mathbf{x}}{N_l - 1}\big) - \big(\mathbf{x} - \mu_l + \frac{\mu_l - \mathbf{x}}{N_l - 1}\big)^T \big(\mathbf{x} - \mu_l + \frac{\mu_l - \mathbf{x}}{N_l - 1}\big) = \\
&J_l - \frac{2}{N_l - 1}(\mu_l - \mathbf{x}) \underbrace{\sum_{j=1}^{N_l}(\mathbf{x}_j - \mu_l)}_{0} + \frac{N_l}{(N_l - 1)^2}(\mu_l - \mathbf{x})^T(\mu_l - \mathbf{x}) + \frac{N_l^2}{(N_l - 1)^2}(\mu_l - \mathbf{x})^T(\mu_l - \mathbf{x})
\end{aligned}$$

de donde luego de agrupar concluimos que

$$\bar{J}_l = J_l - \frac{N_l}{N_l - 1}(\mu_l - \mathbf{x})^T(\mu_l - \mathbf{x}) = J_l - \frac{N_l}{N_l - 1}\,d(\mathbf{x}, \mu_l)$$

y análogamente para el agrupamiento r se obtiene

$$\bar{J}_r = J_r + \frac{N_r}{N_r - 1}(\mu_r - \mathbf{x})^T(\mu_r - \mathbf{x}) = J_r + \frac{N_r}{N_r - 1}\,d(\mathbf{x}, \mu_r)$$