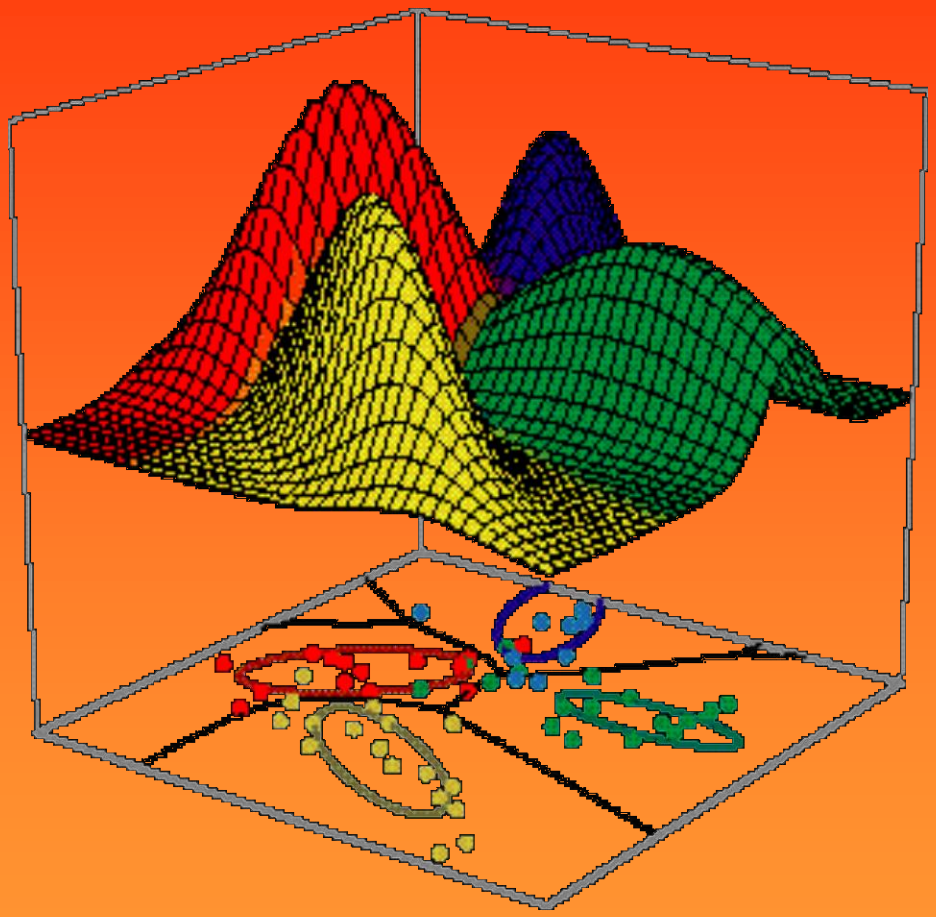

Pattern Classification

All materials in these slides were taken from
Pattern Classification (2nd ed) by R. O. Duda, P. E. Hart and D. G. Stork, John Wiley & Sons, 2000
with the permission of the authors and the publisher

Chapter 3: Maximum-Likelihood & Bayesian Parameter Estimation

- Introduccion
- Estimacion de maxima verosimilitud
  - Ejemplo de un caso especifico
  - El caso gaussiano: μ y σ desconocida
  - Sesgo
- Estimacion Bayesiana

Seccion 3.1 Introduccion

Hipotesis sobre datos en un entorno Bayesiano

- Podemos diseñar un estimador optimo si sabemos:
  - $P(\omega_j)$ (priors)
  - $P(x \mid \omega_j)$ (densidades condicionales a la clase)
- Desafortunadamente, raramente se tiene toda esta informacion completa !!

Hipotesis sobre datos en un entorno Bayesiano

- Opcion I:
  - Estimar toda la informacion faltante de la muestra y usar las estimaciones como si fuera la información real.
- Opcion II
  - Se diseña un clasificador a partir de una muestra de entrenamiento
    - Estimación a priori es usualmente factible.
    - Estimacion condicional por clase muy dificil. Muestras son a menudo muy chicas para la estimacion , el espacio de características es de dimension muy alta .
  - Se agrega información contextual del problema.

Informacion a priori sobre $P(x \mid \omega_i)$

- Supongamos que
  - Se puede suponer una forma funcional para las densidades
  - Se sabe el numero de parametros desconocidos
- El ejemplo mas importante por su utilidad es el de la distribucion normal
  - $P(x \mid \omega_i) \sim N(\mu_i, \Sigma_i)$
  - Caracterizada por 2 parametros desconocidos.
- Tecnicas de estimacion de parámetros que revisaremos
  - Maxima verosimilitud (ML) y Estimacion Bayesiana (BE)
  - Resultados son casi identicos, pero los metodos son muy diferentes

Diferencias entre ML y BE

- En estimación ML, los parámetros son fijos pero desconocidos.
- Los mejores parámetros se obtienen maximizando la probabilidad de obtener las muestras observadas.
- Los métodos Bayesianos ven los parámetros como variables aleatorias con una distribución conocida.

- En cualquiera de los métodos se usa $P(\omega_i \mid x)$ para la regla de clasificación

Aprendizaje

- Las muestras se obtienen seleccionando un estado natural aleatoriamente wi, y luego seleccionando en forma independiente un dato x como muestra de la distribucion condicional p(x|wi)
- Aprendizaje supervizado:
  - Se conoce el estado natural elegido en la muestra con la que se realiza la estimacion de parametros
- Aprendizaje no supervizado
  - No se conoce el estado natural de los datos de la muestra.

Seccion 3.2: Maximum-Likelihood Estimation

- Tiene buenas propiedades de convergencia cuando el tamaño de muestra crece
- Mas simple que cualquier otra técnica alternativa

- Principio general
  - Se asume que hay c clases con $P(x \mid \omega_j) \equiv P(x \mid \omega_j, \theta_j)$ conocida salvo por el valor de:

$$\theta_j = (\theta_j^1, \theta_j^2, ..., \theta_j^m)$$

un vector de parámetros.

- Especificando:
- Si tenemos c clases, llamaremos $D_1 \ldots D_c$ a las muestras independientes identicamente distribuidas corrrespondientes a cada clase.
- Cada $P(x \mid \omega_j, \theta_j)$ tiene forma parametrica conocida.
- Cada θ_j es desconocido.

- En estas notas suponemos independencia, es decir, que no hay informacion en la muestra D_i sobre los parámetros de la clase D_j

Ejemplo Normal

Cada $P(x \mid \omega_j) \sim N(\mu_j, \Sigma_j)$

$$P(x) = \frac{1}{(2\pi)^{d/2}\left|\Sigma_j\right|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu_j)^t \Sigma_j^{-1}(x-\mu_j)\right]$$

- $x = (x_1, x_2, \ldots, x_d)^t$
- $\mu = (\mu_1, \mu_2, \ldots, \mu_d)^t$ vector de medias
- Σ matriz de varianza covarianza
- $|\Sigma|$ y Σ^{-1} son el determinante y su inversa

$$\theta_j = (\mu_j, \Sigma_j) = (\mu_j^1, \mu_j^2, ..., \sigma_j^{11}, \sigma_j^{22}, \mathrm{cov}(x_j^m, x_j^n)...)$$

- Se usa la informacion de las muestras de entrenamiento para estimar

$$\theta = (\theta_1, \theta_2, \ldots, \theta_c),$$

- Cada θ_i esta asociado con una clase, y supongamos que la muestra i-esima no da informacion sobre la j-esima clase,
- Llamemos D a una muestra de una clase genérica que contiene n muestras $x_1, x_2, \ldots, x_n$
- El estimador de MV de θ es, por definicion el valor que maximiza $P(D \mid \theta)$

$$P(D \mid \theta) = \prod_{k=1}^{k=n} P(x_k \mid \theta) = F(\theta)$$

"Es el valor de θ que mejor ajusta con la muestra observada "

Estimación optima

- Sea $\theta = (\theta_1, \theta_2, \ldots, \theta_p)^t$ y sea ∇_θ el operador gradiente

$$\nabla_\theta = \left[\frac{\partial}{\partial \theta_1}, \frac{\partial}{\partial \theta_2}, \ldots, \frac{\partial}{\partial \theta_p} \right]^t$$

- Definimos $l(\theta)$ como la funcion log-likelihood

$$l(\theta) = \ln P(D \mid \theta)$$

- Estimador MV es el θ que maximiza la log-likelihood

$$\hat{\theta} = \arg\max_{\theta} l(\theta)$$

Las condiciones necesarias para que el estimador de MV se obtenga son el conjunto de p ecuaciones

$$l(\theta) = \sum_{k=1}^{n} \ln P(x_k \mid \theta)$$

$$\nabla_\theta l = \sum_{k=1}^{n} \nabla_\theta \ln P(x_k \mid \theta)$$

$$\nabla_\theta l = 0$$

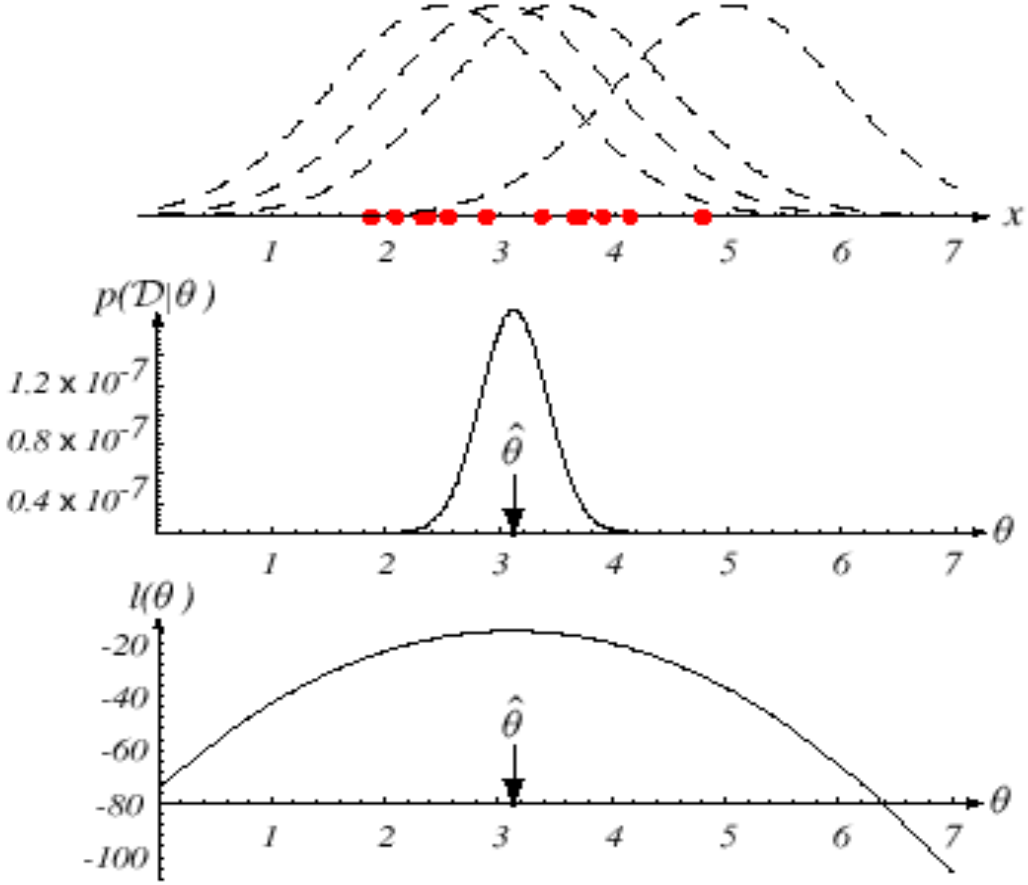

FIGURE 3.1. The top graph shows several training points in one dimension, known or assumed to be drawn from a Gaussian of a particular variance, but unknown mean. Four of the infinite number of candidate source distributions are shown in dashed lines. The middle figure shows the likelihood $p(\mathcal{D}|\theta)$ as a function of the mean. If we had a very large number of training points, this likelihood would be very narrow. The value that maximizes the likelihood is marked $\hat{\theta}$; it also maximizes the logarithm of the likelihood—that is, the log-likelihood $l(\theta)$, shown at the bottom. Note that even though they look similar, the likelihood $p(\mathcal{D}|\theta)$ is shown as a function of θ whereas the conditional density $p(x|\theta)$ is shown as a function of x. Furthermore, as a function of θ, the likelihood $p(\mathcal{D}|\theta)$ is not a probability density function and its area has no significance. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*.

Maximum a posteriori estimators

Estimadores MAP de $\theta = (\theta_1, \theta_2, \ldots, \theta_c)$ son otra clase de estimadores

Encuentran $\hat{\theta}$ tal que :

$$\begin{aligned}
\underset{\theta}{Max}\, P(\theta \mid D) &= = Max\left[P(\underset{\theta}{D} \mid \theta)P(\theta)\right] \\
&= \underset{\theta}{Max}\left[P(x_1,\ldots,x_n \mid \theta)P(\theta)\right] \\
&= \underset{\theta}{Max}\left[\prod_{k=1}^{n} P(x_k \mid \theta)P(\theta)\right] \\
&= \underset{\theta}{Max}\left[l(\theta)P(\theta)\right]
\end{aligned}$$

El estimador MV es el MAP con $P(\theta)$ ‘flat” o uniforme

Seccion 3.2.2: Normales con μ desconocida

- $P(x_i \mid \mu) \sim N(\mu, \Sigma)$, Muestras de poblaciones normales.

$$\ln P(x_k \mid \mu) = -\frac{1}{2}\ln\left[(2\pi)^d |\Sigma|\right] - \frac{1}{2}(x_k - \mu)^t \sum^{-1} (x_k - \mu)$$

$$y \quad \nabla_{\theta\mu} \ln P(x_k \mid \mu) = \sum^{-1} (x_k - \mu)$$

- $\theta = \mu$ por lo cual el estimador de MV de μ resulta:

$$\sum_{k=1}^{k=n} \Sigma^{-1}(x_k - \hat{\mu}) = 0$$

- Multiplicando por Σ y reordenando se obtiene:

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

La media aritmetica de las muestras de entrenamiento

Conclusion:

Si $P(x_k \mid \omega_j)$ (j = 1, 2, …, c) es Gaussiana en el espacio *d*-dimensional de caracteristicas , entoces se puede estimar el vector $\theta = (\theta_1, \theta_2, \ldots, \theta_c)^t$ y realizar una clasificacion optima

Section 3.2.3: Normales μ y σ desconocidos

- $\theta = (\theta_1, \theta_2) = (\mu, \sigma^2)$
- La verosimilitud de un solo punto es

$$l = \ln P(x_k \mid \theta) = -\frac{1}{2}\ln 2\pi\theta_2 - \frac{1}{2\theta_2}(x_k - \theta_1)^2$$

$$\nabla_\theta l = \begin{pmatrix} \frac{\partial}{\partial\theta_1}(\ln P(x_k \mid \theta)) \\ \frac{\partial}{\partial\theta_2}(\ln P(x_k \mid \theta)) \end{pmatrix} = 0$$

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

Mirando todos los puntos juntos:

$$\begin{cases} \sum_{k=1}^{n} \frac{1}{\hat{\theta}_2}(x_k - \theta_1) = 0 & (1) \\ -\sum_{k=1}^{n} \frac{1}{\hat{\theta}_2} + \sum_{k=1}^{n} \frac{(x_k - \hat{\theta}_1)^2}{\hat{\theta}_2^2} = 0 & (2) \end{cases}$$

Combinando (1) y (2), se obtiene:

$$\hat{\mu} = \bar{x} = \sum_{k=1}^{n} \frac{x_k}{n} \quad ; \quad \hat{\sigma}^2 = \frac{\sum_{k=1}^{n} (x_k - \mu)^2}{n}$$

Sesgo

Estimador de MV para σ^2 es sesgado, s^2 es su correccion

$$E\left[\frac{1}{n}\sum(x_i-\bar{x})^2\right]=\frac{n-1}{n}.\sigma^2\neq\sigma^2$$

$$s^2=\frac{1}{n-1}\sum(x_i-\bar{x})^2$$

Estimadores finales usados son

$$\bar{x}=\frac{1}{n}\sum x_i$$

$$s^2=\frac{1}{n-1}\sum(x_i-\bar{x})^2$$

Caso multivariado

- Las cuentas para el caso multivariado son muy similares, pero mucho mas complejas pero los estimadores resultan las contrapartes matriciales de los anteriores

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

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

$$S = \underbrace{\frac{1}{n-1}\sum_{k=1}^{n} (x_k - \hat{\mu})(x_k - \hat{\mu})^t}_{\text{Matriz de covarianza muestral}}$$

Son los mejores estimadores?

- Nuestro problema no es estimar los parametros, sino estimar la mejor regla de clasificación.
- Usando estimadores de los parámetros y reemplazando estos valores en las densidades a priori, obtenemos una regla de clasificacion rápida, la regla ``plug in´´.
- Es esta regla óptima?
- La regla óptima con el riesgo de Bayes se basa en la probabilidad a posteriori, la regla plug in estima dicha probabilidad.

Seccion 3.3: Estimacion bayesiana

- La estimacion de parámetros Bayesiana es diferente en su planteo a ML
  - Estimacion de maxima verosimilitud supone θ fijo
  - Estimacion Bayesiana θ es una variable aleatoria

- Si $p(x|\omega_i)$ tiene una forma funcional con parametros desconocidos, hay informacion en la muestra D que debe ser extraida
- Se puede considerar que esa informacion complementa informacion a priori sobre θ

Seccion 3.3.1: Densidades condicionales por clase

- Revisemos la propuesta Bayesiana de clasificacion, incorporando la muestra de entrenamiento D para estimar los parámetros

$$P(\omega_i \mid x, \mathrm{D}) = \frac{p(x \mid \omega_i, \mathrm{D}).P(\omega_i \mid \mathrm{D})}{\sum_{j=1}^{c} p(x \mid \omega_j, \mathrm{D}).P(\omega_j \mid \mathrm{D})}$$

- Para demostrar la ecuación anterior:

$$P(\omega_i \mid x,\mathrm{D}) = \frac{p(x,\omega_i,\mathrm{D}).}{P(x,\mathrm{D})} = \frac{p(x \mid \omega_i,\mathrm{D}).P(\omega_i,\mathrm{D})}{\sum_{j=1}^{c} P(x,\omega_j,\mathrm{D})}$$

$$= \frac{p(x \mid \omega_i,\mathrm{D}).P(\omega_i \mid \mathrm{D})P(\mathrm{D})}{\sum_{j=1}^{c} p(x \mid \omega_j,\mathrm{D})P(\omega_j \mid \mathrm{D})P(\mathrm{D})} = \frac{p(x \mid \omega_i,\mathrm{D}).P(\omega_i \mid \mathrm{D})}{\sum_{j=1}^{c} p(x \mid \omega_j,\mathrm{D})P(\omega_j \mid \mathrm{D})}$$

$$Si\ P(\omega_i \mid \mathrm{D}) = P(\omega_i)$$

$$P(\omega_i \mid x,\mathrm{D}) = \frac{p(x \mid \omega_i,\mathrm{D}).P(\omega_i)}{\sum_{j=1}^{c} p(x \mid \omega_j,\mathrm{D})P(\omega_j)}$$

- Se tienen $D_1,\ldots,D_c$ muestras distintas, y D_i no tiene influencia en $p(x|\omega_i,D)$ si $i \neq j$

$$P(\omega_i \mid x,\mathrm{D}) = \frac{p(x \mid \omega_i,\mathrm{D_i}).P(\omega_i)}{\sum_{j=1}^{c} p(x \mid \omega_j,\mathrm{D_j})P(\omega_j)}$$

- En esencia, se tienen c problemas separados iguales

Usando la muestra D obtenida muestreando una probabilidad fija p(x) desconocida, se estima la densidad p(x| D) de cada clase fija por separado

Seccion 3.3.2 Distribucion de parametros

- Densidad p(x) es desconocida, pero se la supone parametrica.
- Unica cosa desconocida es el vector de parametros θ.
- Entonces p(x| θ) es totalmente conocida y se conoce una densidad a priori p(θ).
- La observacion de la muestra D permite calcular la densidad a posteriori p(θ|D), que esperemos, sea ajustada alrededor del valor desconocido θ.

- Estamos cambiando el problema de encontrar las clases por un problema de estimacion de densidades. Deseamos calcular p(x|D) que es lo mas cercano a conocer p(x) para esa clase.
- Hacemos eso integrando sobre la densidad conjunta p(x,θ|D) sobre θ.

$$\mathrm{p}(\mathrm{x}|D) = \int p(x,\theta|D)d(\theta) = \int p(x|\theta)\,p(\theta|D)d(\theta)$$

- Esta ecuacion conecta la densidad condicional por clase p(x|D) con la densidad a posteriori p(θ|D).
- Si p(θ|D) tiene un pico agudo en algun valor θ_e se obtiene

$$\mathrm{p}(\mathrm{x}|D) \cong p(x|\theta_e)$$

- Si no es asi, se promedia p(x|θ) sobre todos los valores de θ

Seccion 3.4: Estimacion Bayesiana de parametros, caso gaussiano

- Se desea estimar θ usando la densidad a posteriori $P(\theta|D)$, para luego estimar $p(x|D)$, en el caso

$$p(x|\mu) \sim N(\mu, \Sigma)$$

Seccion 3.4.1: Caso Univariado: p(μ|D)

- Caso $p(x|\mu) \sim N(\mu, \sigma^2)$

- μ es el unico parametro desconocido

- Todo conocimiento previo se concentra en

$$p(\mu) \sim N(\mu_0, \sigma_0^2)$$

donde μ_0 y σ_0 son conocidos

- Pensamos que si se observo μ, entonces podemos muestrear $D=(x_1 ,\ldots,x_n)$ de la distribucion $p(x|\mu)$ para obtener

$$p(\mu|D) = \frac{p(D|\mu).p(\mu)}{\int p(D|\mu).p(\mu)d\mu} \qquad (1)$$

$$= \alpha \prod_{k=1}^{n} p(x_k|\mu).p(\mu)$$

- Reemplazando por las densidades supuestas se obtiene

$$p(\mu|D) \sim N(\mu_n, \sigma_n^2) \qquad (2)$$

con

$$\mu_n = \left(\frac{n\sigma_0^2}{n_0\sigma_0^2 + \sigma^2} \right) \bar{x}_n + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2}.\mu_0$$

$$y \;\; \sigma_n^2 = \frac{\sigma_0^2\sigma^2}{n\sigma_0^2 + \sigma^2}$$

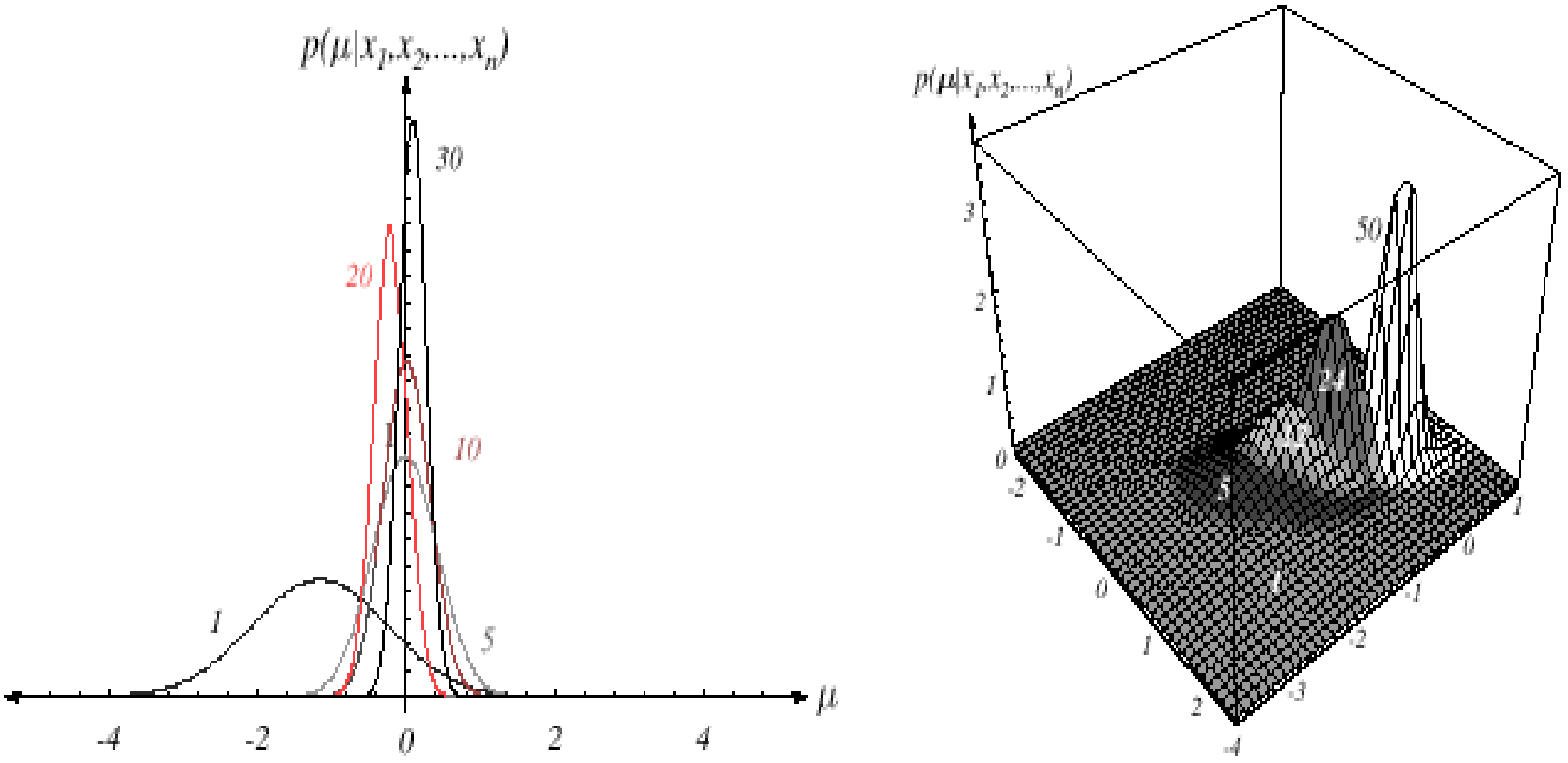

FIGURE 3.2. Bayesian learning of the mean of normal distributions in one and two dimensions. The posterior distribution estimates are labeled by the number of training samples used in the estimation. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Seccion:3.4.2: El caso univariado P(x | D)

- Recordemos que se desea estimar θ usando la densidad a posteriori $P(\theta \mid D)$, para luego estimar $p(x|D)$

- En este caso, la densidad $p(\mu \mid D)$ es la calculada, y

$$p(x \mid \mathrm{D}) = \int p(x \mid \mu).p(\mu \mid \mathrm{D})d\mu$$

Reemplazando, se ve que p(x|D) es una gaussiana

$$p(x \mid \mathrm{D}) \sim N(\mu_n, \sigma^2 + \sigma_n^2)$$

- La densidad p(x|D) es la deseada densidad condicional por clase $p(x|\omega_i,D_i)$ y junto a las probabilidades a priori $p(\omega_i)$ se tiene toda la informacion para diseñar el clasificador

Seccion 3.4.3: El caso mutivariado

- Suponiendo que

$$p(x \mid \mu) \sim N(\mu, \Sigma) \quad y \; p(\mu) \sim N(\mu_0, \Sigma_0)$$

Y observando un conjunto D de n muestras independientes, usando la formula de Bayes

$$p(\mu \mid \mathrm{D}) = \alpha \prod_k p(x_k \mid \mu)\, p(\mu)$$

$$= \alpha' \exp\left[-\frac{1}{2} \left(\mu^t (n\Sigma^{-1} + \Sigma_0^{-1})\mu - 2\mu^t (\Sigma^{-1} \sum_{k=1}^{n} x_k + \Sigma_0^{-1}\mu_0) \right) \right]$$

El caso mutivariado

- Esta formula puede reducirse a

$$p(\mu \mid \mathrm{D}) = \alpha'' \exp\left[-\frac{1}{2}(\mu-\mu_n)^t \Sigma_n^{-1}(\mu-\mu_n)\right]$$

Por lo cual

$$p(\mu \mid D) \sim N(\mu_n, \Sigma_n)$$

E igualando coeficientes se obtiene

$$\mu_n = \Sigma_0\left(\Sigma_0 + \frac{1}{n}\Sigma\right)^{-1}\bar{x}_n + \frac{1}{n}\Sigma\left(\Sigma_0 + \frac{1}{n}\Sigma\right)^{-1}\mu_0$$

$$\Sigma_n = \Sigma_0\left(\Sigma_0 + \frac{1}{n}\Sigma\right)^{-1}\frac{1}{n}\Sigma$$

Finalmente p(x | D)

- Usando la densidad p(μ | D) calculada, y la formula

$$p(x|\mathrm{D}) = \int p(x|\mu).p(\mu|\mathrm{D})d\mu$$

se ve que p(x|D) es una gaussiana

$$p(x|\mathrm{D}) \sim N(\mu_n, \Sigma^2 + \Sigma_n^2)$$

Y puede diseñarse el clasificador

Seccion 3.5: Teoria General: estimacion parametrica Bayesiana fuera del caso normal

- El computo de p(x | D) puede ser aplicado a cualquier situacion en que la densidad desconocida puede ser parametrizada. Las situaciones basicas son:
  - La forma de p(x | θ) se asume conocida pero el valor de θ no es conocido exactamente
  - Nuestro conocimiento acerca de θ se asume contenido en la densidad a priori p(θ)
  - El resto de nuestro conocimiento sobre θ se contiene en el conjunto D de n observaciones $x_1, x_2, \ldots, x_n$ muestredas independientemente de la distribucion desconocida p(x)

El problema basico es :

"Calcular la densidad a posteriori p(θ|D) para derivar p(x| D)"

$$\mathrm{p}(\mathrm{x}|D) = \int p(x|\theta)\, p(\theta|D) d(\theta)$$

Usando la formula de Bayes, se tiene :

$$p(\theta|D) = \frac{p(D|\theta).p(\theta)}{\int p(D|\theta).p(\theta) d\theta},$$

Por la suposicion de independencia:

$$p(D|\theta) = \prod_{k=1}^{n} p(x_k|\theta)$$

Relacion con MV

- Supongamos que $p(D|\theta)$ alcanza un pico en $\theta = \theta_0$
- Si la densidad a priori $p(\theta)$ no es cero en $\theta = \theta_0$ y no cambia mucho alrededor de un entorno.
- Entonces $p(\theta|D)$ tambien alcanza un pico en $\theta = \theta_0$
- Por lo cual $p(x|D)$ es aproximadamente $p(x|\theta_0)$, la solucion “plug in” usando el estimador de maxima verosimilitud.

Recursive Bayes Approach

- Recursion

$$D^n = \{x_1, \ldots, x_n\}$$

$$p(D^n|\theta) = p(x_n|\theta)\, p(D^{n-1}|\theta)$$

$$p(\theta|D^n) = \frac{p(x_n|\theta)\, p(\theta|D^{n-1})}{\int p(x_n|\theta)\, p(\theta|D^{n-1})\, d\theta}$$

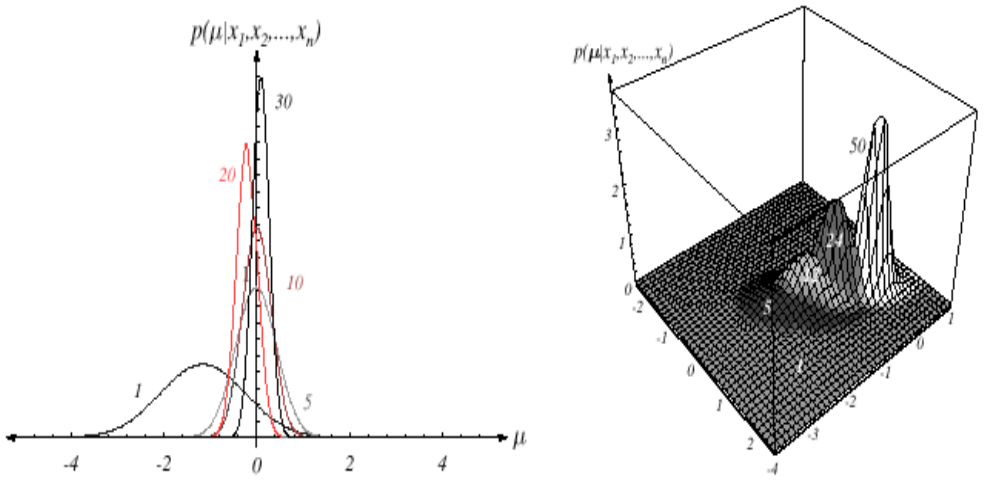

FIGURE 3.2. Bayesian learning of the mean of normal distributions in one and two dimensions. The posterior distribution estimates are labeled by the number of training samples used in the estimation. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Ejemplo 1: Aprendizaje Bayesiano recursivo

- Muestras provienen de una distribucion uniforme,

$$p(x \mid \theta) \sim U(0,\theta) = \begin{cases} 1/\theta & 0 \le x \le \theta \\ 0 & afuera \end{cases}$$

- Se supone un prior no informativo o flat

$$p(\theta) \sim U(0,10)$$

- Datos seleccionados aleatoriamente de la distribucion subyacente

$$D = \{4,7,2,8\}$$

- Antes que se tenga ningun dato

$$p(\theta \mid D^0) = p(\theta) = U(0,10)$$

- Cuando el primer dato x=4 se observa

$$p(\theta \mid D^1) \propto p(x \mid \theta)\, p(\theta \mid D^0) = \begin{cases} 1/\theta & 4 \le \theta \le 10 \\ 0 & afuera \end{cases}$$

olvidando la renormalizacion

- Cuando el siguiente dato x=7 se observa

$$p(\theta \mid D^2) \propto p(x \mid \theta)\, p(\theta \mid D^1) = \begin{cases} 1/\theta^2 & 7 \le \theta \le 10 \\ 0 & afuera \end{cases}$$

- La forma general de la solución es

$$p(\theta \mid D^n) \propto 1/\theta^n \qquad \max_x[D^n] \leq \theta \leq 10$$

- Con la muestra conjunta, el estimador MV es θ=8, lo cual implica una distribucion

$$p(\theta \mid D) \sim U(0,8)$$

- En cambio, la metodologia Bayesiana requiere una integracion

$$\mathrm{p}(\mathrm{x} \mid D) = \int p(x \mid \theta)\, p(\theta \mid D) d(\theta)$$

- La densidad solucion es uniforme hasta θ=8 y pero tiene una cola para valores mayores indicando que la influencia del prior p(θ) no ha sido descartado por la informacion de la muestra de entrenamiento.
- MV estima un punto en el espacio de θ, el metodo bayesiano estima una densidad.
- Tecnicamente, no se pueden comparar los estimadores pero si se pueden comparar las distribuciones p(x|D)

Identificabilidad

- Para la mayoria de las densidades $p(x|\theta)$, la sucesion de densidades converge a una delta de dirac.
- Esto implica, que con un gran numero de muestras, habria un solo θ que ajuste a esos datos, por lo cual θ puede ser identificado univocamente de $p(x|\theta)$.
- Cuando mas de un valor de θ ajusta los datos, decir da el mismo valor para $p(x|\theta)$, y resulta noidentificable.
- Sin embargo, $p(x|D^n)$ va a converger a $p(x)$, a pesar de la no identificabilidad de θ

3.5.1 Diferencias entre MV y Bayes Learning

- Asintoticamente, en la mayoria de los casos coinciden.
- En muestra finita, sin embargo, hay diferencias a considerar
  - Complejidad computacional
    - MV requiere calculo diferencia o busqueda por gradiente
    - BL integracion multidimensional
  - Interpretabilidad
    - MV da un solo valor
    - BL un promedio pesado de modelos, que a menudo son dificiles de interpretar

- Confianza en la informacion a priori

  - MV asume una densidad parametrica a priori, y la solucion $p(x|\theta_{MV})$ tiene esa forma.
  - BL puede no generar una p(x|D) con la forma supuesta, porque se modera con la informacion a priori sobre θ
  - Si la informacion a priori es veraz, el estimador Bayesiano es mejor
  - Si no hay informacion previa, y se supone un prior flat o uniforme, entonces BL es equivalente a ML

Errores en el clasificador

- Hay tres fuentes de error de clasificacion en el sistema final, que utiliza la densidad a posterior de cada categoria para clasificar usando el maximo a posteriori.
  - Error de Bayes: relativa al area donde las densidades por clase se cruzan. Es inherente al problema y no puede ser eliminado
  - Error de Modelo: Solo se elimina si el diseñador incluye el modelo real con que se generaron los datos. Se utiliza informacion previa sobre el problema, y raramente el metodo MV o BL influye en el error
  - Error de estimacion: parametros estimados de una muestra finita. Se reduce aumentando al muestra de entrenamiento

3.5.2 Priors no informativos

- La nocion de informacion en una distribucion a priori es compleja.
- Revela conocimiento del problema, pero en algunos casos, debe revelar el desconocimiento sobre la estructura del problema.
- Esto es, no se debe preferir regiones de parametros sobre otras.
- Priors no informativos refieren tambien a densidades (propias o impropias) que son invariantes con respecto a una caracteristica.

- Seccion 3.7: Problemas de Dimensionalidad
  - Problemas con 50 a 100 caracteristicas binarias
  - Exactitud de la clasificacion depende de la dimensionalidad y de la cantidad de la muestra de entrenamiento
- Caso de dos normales multivariadas con la misma covarianza.

$$P(error) = \frac{1}{\sqrt{2\pi}} \int_{r/2}^{\infty} e^{\frac{-u^2}{2}} du$$

$$donde: \quad r^2 = (\mu_1 - \mu_2)^t \Sigma^{-1} (\mu_1 - \mu_2)$$

$$\lim_{r \to \infty} P(error) = 0$$

- Si las caracteristicas son independientes entonces

$$\Sigma = diag(\sigma_1^2, \sigma_2^2, ..., \sigma_d^2)$$

$$r^2 = \sum_{i=1}^{i=d} \left(\frac{\mu_{i1} - \mu_{i2}}{\sigma_i} \right)^2$$

- Cada caracteristica contribuye a reducir el error, por lo cual es razonable aumentar las caracteristicas para reducir el error
- Se ha observado en la practica que mas haya de cierto punto la inclusion de ciertas caracteristicas adicionales empeoran la clasificacion

La razon mas frecuente es la falla de las hipotesis sobre el modelo

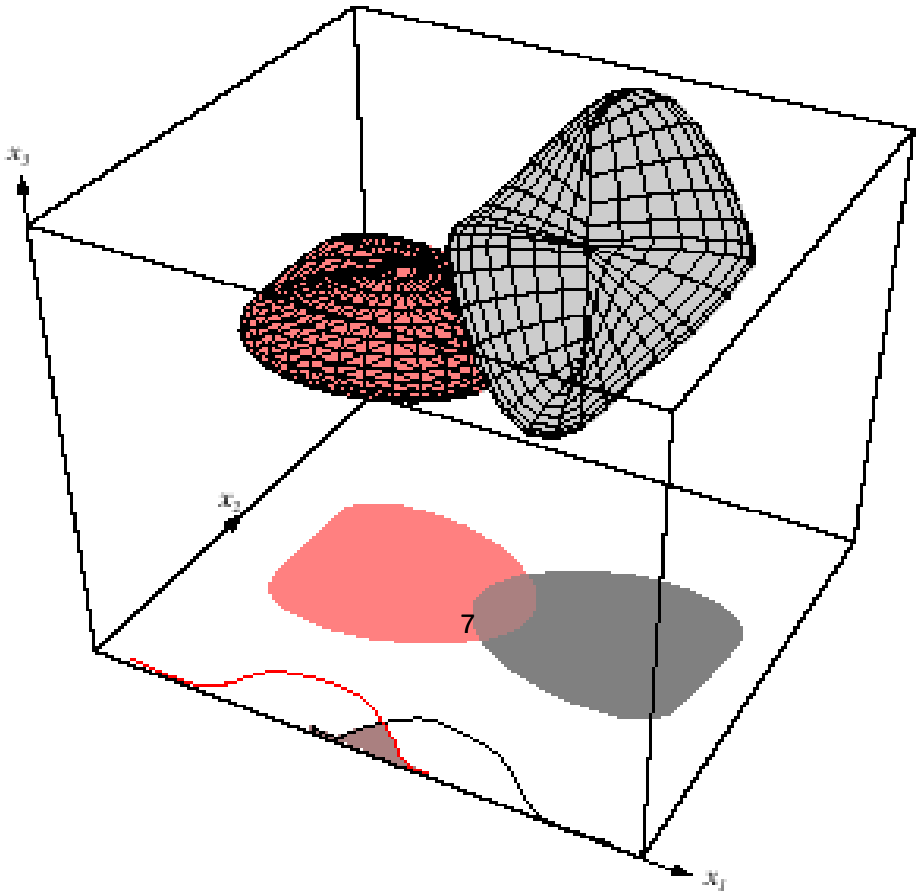

FIGURE 3.3. Two three-dimensional distributions have nonoverlapping densities, and thus in three dimensions the Bayes error vanishes. When projected to a subspace—here, the two-dimensional $x_1 - x_2$ subspace or a one-dimensional x_1 subspace—there can be greater overlap of the projected distributions, 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.

Complejidad computacional

- Nuestra metodologia de diseño esta afectada por la dificultad en el computo de los discriminantes

  - Notacion de orden de una funcion,

    f(x) = O(h(x)) "Oh grande de h(x)"

    Si:

$$\exists (c_0, x_0) \in \Re^2;\ |f(x)| \leq c_0 |h(x)|$$

f(x) crece como h(x) para x suficientemente grande!

$f(x) = 2+3x+4x^2$ $g(x) = x^2$ $f(x) = O(x^2)$

“oh grande” no es unica

- $f(x) = O(x^2)$; $f(x) = O(x^3)$; $f(x) = O(x^4)$

- Notacion “big theta”

  $f(x) = \theta(h(x))$

  Si:

$$\exists(x_0, c_1, c_2) \in \Re^3; \forall x > x_0$$

$$0 \leq c_1 g(x) \leq f(x) \leq c_2 g(x)$$

$f(x) = \theta(x^2)$ pero $f(x) \neq \theta(x^3)$

- Complejidad de la estimacion ML para el caso gausiano

- Supongamos densidades gaussianas en d dimensiones
- Clasificador con n muestras de entrenamiento para cada una de las c clases

- Para cada categoria tenemos que computar la funcion discriminante

$$g(x) = -\frac{1}{2}\underbrace{(x - \hat{\mu})^t}_{O(d.n)}\underbrace{\Sigma^{-1}}_{O(n.d^2)}(x - \hat{\mu}) - \overbrace{\frac{d}{2}\ln 2\pi}^{O(1)} - \underbrace{\frac{1}{2}\ln|\hat{\Sigma}|}_{O(d^2.n)} + \underbrace{\ln P(\omega)}_{O(n)}$$

Total = $O(d^2.n)$

Total para las c clases = $O(cd^2.n) \cong O(d^2.n)$

Costo de incremento cuando d y n son grandes !

3.7.3 Overfitting

Insuficiente cantidad de datos para calcular todos los parametros.

- Se debe combinar caracteristicas para reducir dimension
- Las combinaciones lineales son simples y
- Se proyectan los datos multidimensionales en un espacio de dimension menor
- Dos metodos clasicos

  - PCA (Principal Component Analysis)

  "Projection that best represents the data in a least- square sense"

  - MDA (Multiple Discriminant Analysis)

  "Projection that best separates the data in a least-squares sense"

Alternativas

- Suponer igualdad de covarianzas
- Suponer covarianza diagonal, independendencia.
  - Algunas veces, si la cantidad de datos es muy chica, dan mejores estimadores a pesar de la evidencia en contra de independencia
- Modelos simples tienen mejor estimados los parametros que modelos complejos

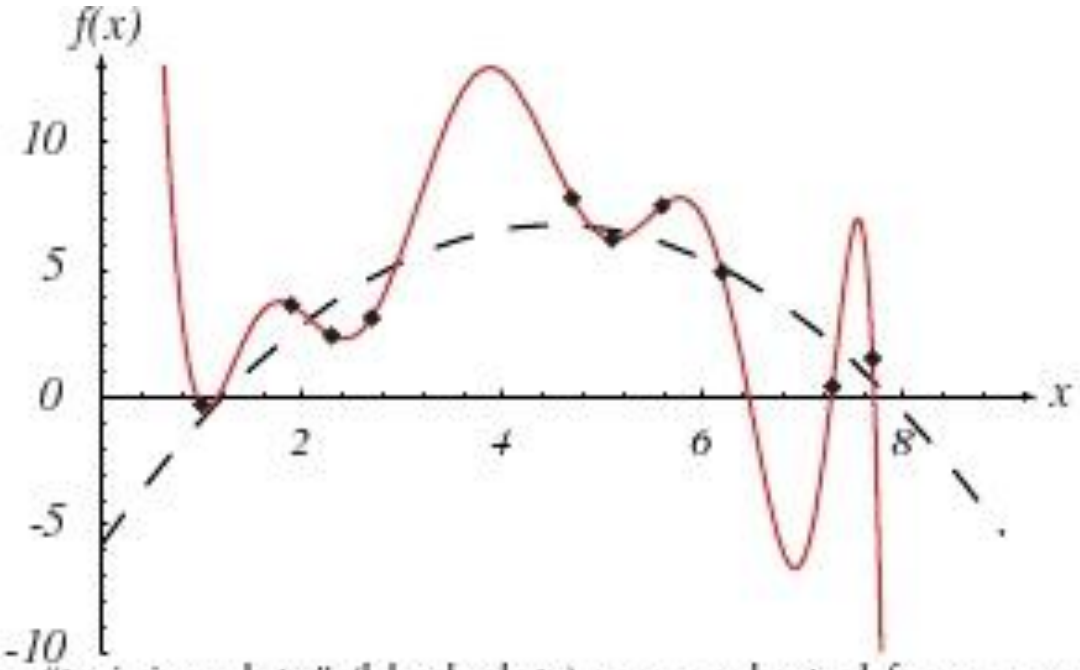

FIGURE 3.4. The "training data" (black dots) were selected from a quadratic function plus Gaussian noise, i.e., $f(x) = ax^2+bx+c+\epsilon$ where $p(\epsilon) \sim N(0, \sigma^2)$. The 10th-degree polynomial shown fits the data perfectly, but we desire instead the second-order function $f(x)$, because 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.

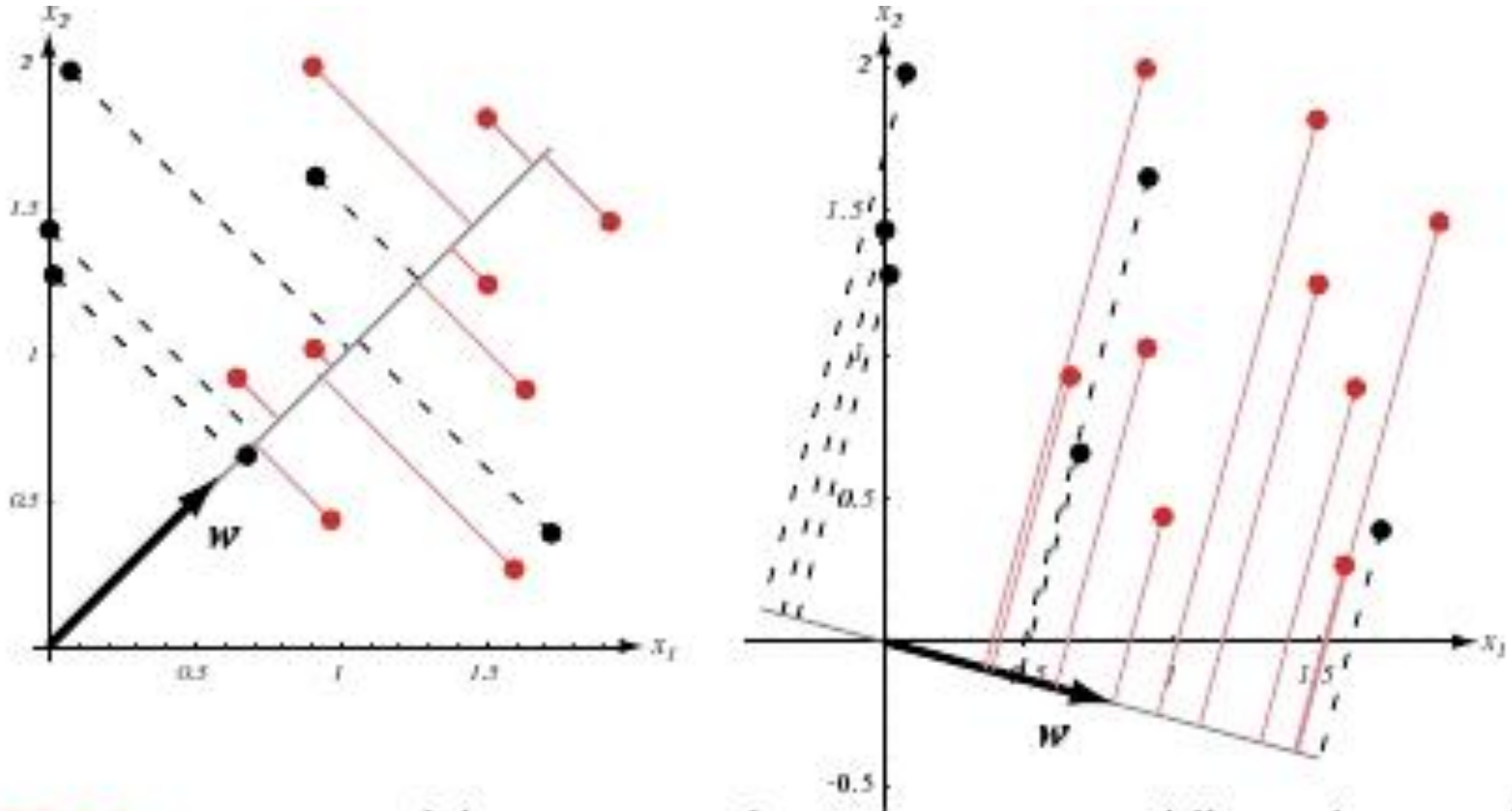

FIGURE 3.5. Projection of the same set of samples onto two different lines in the directions marked **w**. The figure on the right shows greater separation between the red and black projected points. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001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

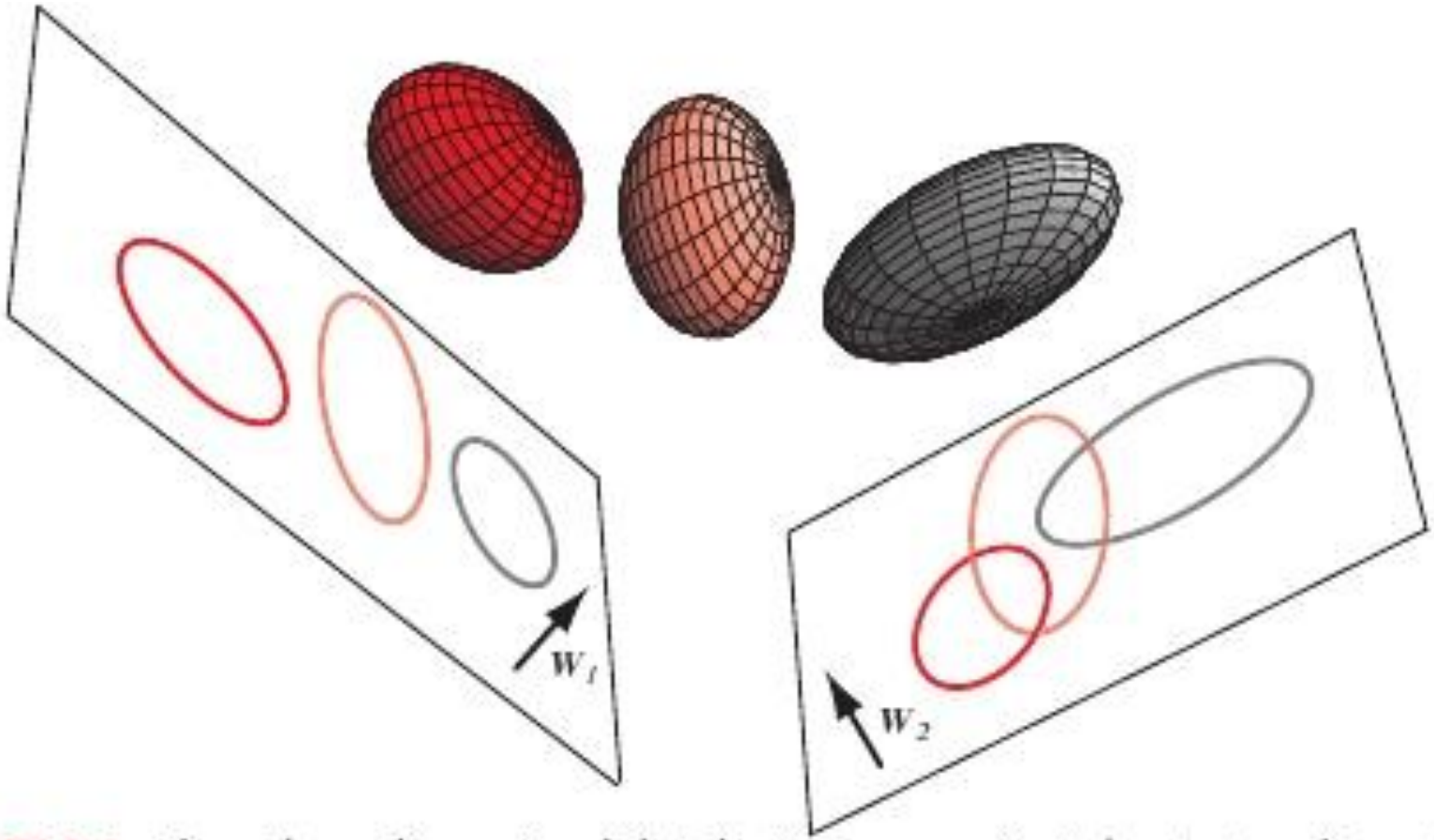

FIGURE 3.6. Three three-dimensional distributions are projected onto two-dimensional subspaces, described by a normal vectors $\mathbf{W}_1$ and $\mathbf{W}_2$. Informally, multiple discriminant methods seek the optimum such subspace, that is, the one with the greatest separation of the projected distributions 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*.

EM Expectation Maximization

- Problema:
  - datos incompletos
  - necesidad de estimar parametros
  - maxima verosimilitud es posible?
- encontrar θ tal que l(θ) es maxima, con restriccion en los datos, es posible?
- Sea

$$D = \{x_1, \ldots, x_n\} = D_g \cup D_b$$

$$Q(\theta; \theta^i) = E_{D_b}(\ln p(D_g, D_b) \mid D_g; \theta^i)$$

Algorithm

Algorithm 1 (Expectation-Maximization)

```
1 begin initialize θ^0, T, i = 0
2         do i ← i + 1
3             E step : compute Q(θ; θ^i)
5             M step : θ^{i+1} ← arg max_θ Q(θ; θ^i)
6         until Q(θ^{i+1}; θ^i) − Q(θ^i; θ^{i−1}) ≤ T
7   return θ̂ ← θ^{i+1}
8 end
```

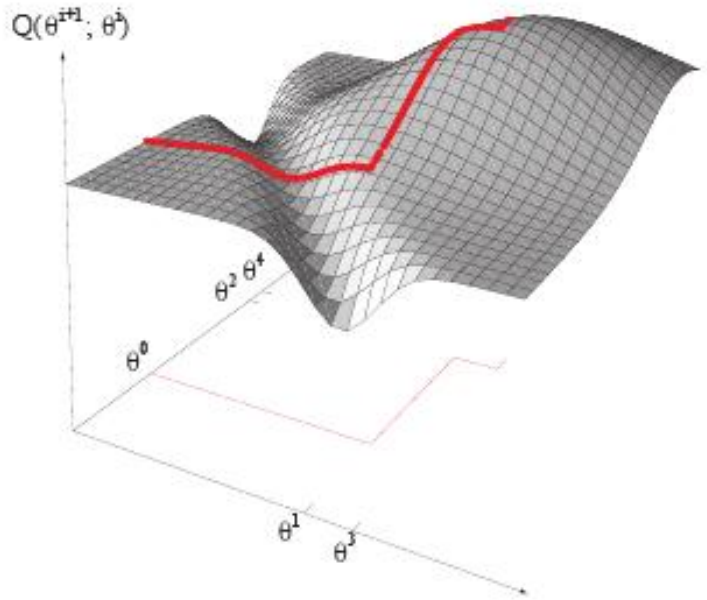

Figure 3.5: The search for the best model via the EM algorithm starts with some initial value of the model parameters, θ^0. Then, via the **M step** the optimal θ^1 is found. Next, θ^1 is held constant and the value θ^2 found which optimizes $Q(\cdot,\cdot)$. This process iterates until no value of θ can be found that will increase $Q(\cdot,\cdot)$. Note in particular that this is different from a gradient search. For example here θ^1 is the global optimum (given fixed θ^0), and would not necessarily have been found via gradient search. (In this illustration, $Q(\cdot,\cdot)$ is shown symmetric in its arguments; this need not be the case in general, however.)

Ejemplo

- Supongamos que tenemos dos distribuciones normales

$$D = \{x_1, x_2, x_3, x_4\} = \left\{\begin{pmatrix}0\\2\end{pmatrix}, \begin{pmatrix}1\\0\end{pmatrix}, \begin{pmatrix}2\\2\end{pmatrix}, \begin{pmatrix}*\\4\end{pmatrix}\right\} = D_g \cup D_b$$

$$D_b = x_{41} \quad \theta = (\mu_1, \mu_{2,}\sigma_1^2, \sigma_2^2)$$

$$\theta^0 = (0,0,1,1)$$

$$Q(\theta;\theta^0) = E_{D_b}(\ln p(D_g, D_b) \mid D_g; \theta^i)$$

$$= \int \left[\sum_k \ln p(x_k \mid \theta) + \ln p(x_4 \mid \theta)\right] p(x_{41} \mid \theta^0; x_{42} = 4) dx_{41}$$

Reemplazando por las densidades

- el paso E resulta

$$Q(\theta;\theta^0)=\sum_{k=1}^{3}\ln\, p(x_k\mid\theta)-\frac{1+\mu_1^2}{2\sigma_1^2}-\frac{(1-\mu_2)^2}{2\sigma_2^2}-\ln(2\pi\sigma_1\sigma_2)$$

- ahora se debe maximizar esta funcion (que no tiene ya mas incognitas que los parametros) y se obtiene

$$\theta^1=\begin{pmatrix}0.75\\2\\0.938\\2\end{pmatrix}$$

- Iterando se converge luego de tres pasos a

$$\theta=\begin{pmatrix}1\\2\\0.667\\2\end{pmatrix}$$

EM generalizado

- Se suele llamar EM generalizados a los algoritmos de descenso basados en la verosimilitud parcial de los datos
- El algoritmo Bayesiano que obtiene el optimo condicional a lo observado y mejorado en el paso anterior es el verdadero EM, y tiene sentido cuando la funcion Q es mas facil de calcular que la verosimilitud, en especial cuando no se tiene toda la informacion.
- Es importante destacar que el optimo para la verosimilitud con todos los datos quizas sea otro valor, pero tambien es otro problema.