Modelos Gaussianos

Alfredo Cuesta Infante

E. T. S. Ingeniería Informática
Universidad Rey Juan Carlos

Master Univ. en Visión Artificial
Reconocimiento de Patrones

Contenidos

Contenidos

La distribución Normal o Gaussiana

Distribución Normal

$$\mathcal{N}(x) = \frac{\exp\left(-\frac{(x-\mu)^2}{\sigma^2}\right)}{\sqrt{2\pi\sigma^2}}$$

- Tiene 2 parámetros: la media y la varianza
- x tiene 1 sola dimensión
- La media μ es un escalar
- La varianza σ^2 es un escalar positivo
- $\sqrt{(x-\mu)^2}$ es la distancia Euclidea (en 1 dim.) desde x a la media μ

Distribución MVN

$$\mathrm{MVN}(\mathbf{x}) = \frac{\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)}{\sqrt{(2\pi)^n|\boldsymbol{\Sigma}|}}$$

- Tiene 2 parámetros: la media y la covarianza
- $\mathbf{x}$ es un vector n-dimensional
- La media $\boldsymbol{\mu}$ es un vector n-dimensional
- La matriz de covarianza $\boldsymbol{\Sigma}$ es una matriz definida positiva de tamaño $n \times n$
- $\sqrt{((\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}))}$ es la distancia Mahalanobis desde $\mathbf{x}$ a la media $\boldsymbol{\mu}$

La distribución Normal o Gaussiana

La media

El estimador MLE para la MVN es el *clásico* promedio para cada componente.

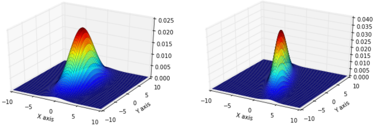
$$\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n), \text{ donde } \mu_j = \mathbb{E}[\mathbf{x}_j] = \frac{1}{m}\sum_{i=1}^{m} x_j^{(i)}, \text{ para } j = 1, \ldots, n.$$

La covarianza

Es una matriz (definida positiva) cuyos elementos expresan la varianza de un componente respecto de otro. Su estimador MLE es:

$$\Sigma_{ij} = \mathbb{E}\left[(\mathbf{x}_i - \mu_i)(\mathbf{x}_j - \mu_j)\right].$$

- Σ diagonal $\Rightarrow$ componentes independientes
- En otro caso hay dependencia lineal (correlación) entre componentes



(a) (b)

$$\boldsymbol{\Sigma} = \begin{bmatrix} 3 & 0 \\ 0 & 15 \end{bmatrix} \text{ vs. } \boldsymbol{\Sigma} = \begin{bmatrix} 3 & -5 \\ -5 & 15 \end{bmatrix}$$

Curvas de equidensidad

- Descomponer de la matriz de convarianza en autovalores y autovectores, y después tomar su inversa

$$\boldsymbol{\Sigma}^{-1} = \mathbf{U}\boldsymbol{\Lambda}^{-1}\mathbf{U}^T = \sum_{i=1}^{n} \frac{1}{\lambda_i}\mathbf{u}_i\mathbf{u}_i^T$$

- Reescribir la distancia Mahalanobis con esta expresión de $\boldsymbol{\Sigma}^{-1}$:

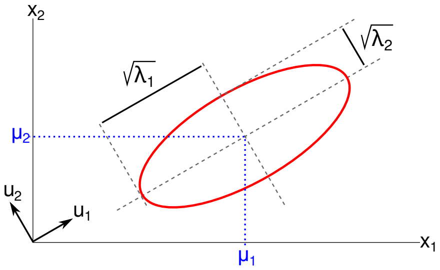
$$\begin{aligned} d_M(\mathbf{x}) = (\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) &= (\mathbf{x}-\boldsymbol{\mu})^T\left(\sum_{i=1}^{n}\frac{1}{\lambda_i}\mathbf{u}_i\mathbf{u}_i^T\right)(\mathbf{x}-\boldsymbol{\mu}) \\ &= \sum_{i=1}^{n}\frac{1}{\lambda_i}(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{u}_i\mathbf{u}_i^T(\mathbf{x}-\boldsymbol{\mu}) \end{aligned}$$

- Sea $y_i = (\mathbf{x}-\boldsymbol{\mu})^T\mathbf{u}_i$. La distancia Mahalanobis es finalmente:

$$d_M(\mathbf{x}) = \sum_{i=1}^{n}\frac{y_i^2}{\lambda_i}$$

- Si fijamos esta distancia a un valor concreto D, entonces $D = \sum_{i=1}^{n} y_i^2/\lambda_i$ es la expresión una **elipse** centrada en $\boldsymbol{\mu}$, cuyo eje i-esimo tiene la dirección $\mathbf{u}_i$ y tamaño $\sqrt{\lambda_i}$

Curvas de equidensidad



Contenidos

Análisis de Discriminantes Gaussianos

- Asumiendo que el subconjunto de ejemplos con etiqueta t está distribuido según una MVN($\boldsymbol{\mu}_t$, $\boldsymbol{\Sigma}_t$), el modelo de verosimilitud es:

$$p(\mathbf{X}|t) = \mathrm{MVN}(\mathbf{X}|t;\ \boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$$

- La etiqueta estimada MAP es:

$$\hat{t} = \arg\max_{t} \left(\log p(t) + \log \mathrm{MVN}(\mathbf{X}|t;\ \boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)\right), \text{ para cada } \mathbf{x}.$$

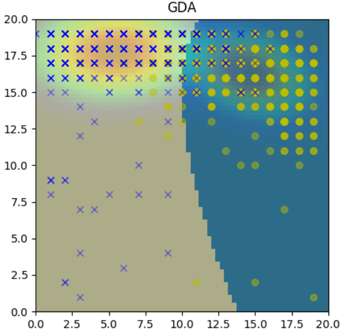


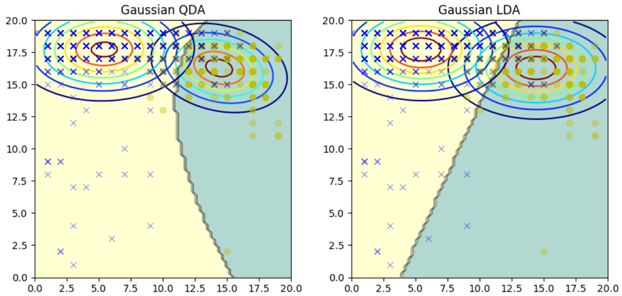
Figura: Superficie de decisión y densidades de las MVN para un GDA con el conjunto de datos '0' vs. '1'. [Fuente: Original de A. Cuesta]

Análisis de Discriminantes Gaussianos

- Equivalente a minimizar la distancia Mahalanobis de un punto $\mathbf{x}$ al centro de cada una de las MVN que modelan las diferentes clases;

$$\hat{t} = \arg\min_t \left((\mathbf{x} - \boldsymbol{\mu}_t)^T \boldsymbol{\Sigma}_t^{-1} (\mathbf{x} - \boldsymbol{\mu}_t) \right), \text{ para cada } \mathbf{x}.$$

- Cuando la matriz de covarianza es diagonal, entonces recuperamos el **NBC Gaussiano**. En caso contrario:
  - Si cada subconjunto de datos con la misma etiqueta tiene covarianzas diferentes tenemos **Discriminantes Cuadráticos** (QD)
  - Si todas son iguales tenemos **Discriminantes Lineales** (LD)



Análisis de Discriminantes Gaussianos

¿Podemos averiguar la ec. de la superficie de decisión?

- Para dos clases $t = \{0, 1\}$, la distribución a posteriori de un nuevo ejemplo $\mathbf{z}$ es:

$$p(t=0|\mathbf{z}) \propto p(t=0) \cdot (2\pi|\mathbf{\Sigma}_0|)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{z}-\boldsymbol{\mu}_0)^T \mathbf{\Sigma}_0^{-1}(\mathbf{z}-\boldsymbol{\mu}_0)\right)$$

$$p(t=1|\mathbf{z}) \propto p(t=1) \cdot (2\pi|\mathbf{\Sigma}_1|)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{z}-\boldsymbol{\mu}_1)^T \mathbf{\Sigma}_1^{-1}(\mathbf{z}-\boldsymbol{\mu}_1)\right)$$

- El log. de su cociente es:

$$\delta = \log \frac{p(t=0|\mathbf{z})}{p(t=1|\mathbf{z})} = \log p(t=0|\mathbf{z}) - \log p(t=1|\mathbf{z}),$$

- Manipulando se llega a

$$\begin{aligned}\delta = \log p(t=0) - \frac{1}{2}\log|\mathbf{\Sigma}_0| - \frac{1}{2}(\mathbf{z}-\boldsymbol{\mu}_0)^T \mathbf{\Sigma}_0^{-1}(\mathbf{z}-\boldsymbol{\mu}_0) \\ - \log p(t=1) + \frac{1}{2}\log|\mathbf{\Sigma}_1| + \frac{1}{2}(\mathbf{z}-\boldsymbol{\mu}_1)^T \mathbf{\Sigma}_1^{-1}(\mathbf{z}-\boldsymbol{\mu}_1).\end{aligned}$$

☺ La superficie de decisión será el conjunto de puntos $\mathbf{z}$ donde $\delta = 0$.

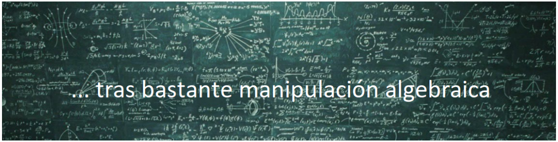
Análisis de Discriminantes Gaussianos

¿Podemos averiguar la ec. de la superficie de decisión?

- Dejando a un lado de la igualdad todo lo que depende de $\mathbf{z}$ se obtiene la siguiente forma cuadrática:

$$(\mathbf{z}-\boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}_1^{-1}(\mathbf{z}-\boldsymbol{\mu}_1) - (\mathbf{z}-\boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}_0^{-1}(\mathbf{z}-\boldsymbol{\mu}_0) = C,$$

donde $C = 2\log p(t=1) - 2\log p(t=0) + \log|\boldsymbol{\Sigma}_1| - \log|\boldsymbol{\Sigma}_0|$ es un valor constante dadas las matrices de covarianza de cada clase y la distribución a priori.



- obtenemos una manera directa de clasificar nuevos ejemplos según la distancia Mahalanobis de estos al centro de cada clase:

$$\hat{t} = \begin{cases} 0 & \text{si } d^2_{M1}(z) - d^2_{M0}(z) + C < 0 \\ 1 & \text{si } d^2_{M1}(z) - d^2_{M0}(z) + C > 0 \end{cases}$$

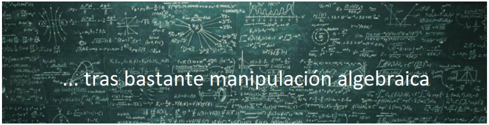
$\Rightarrow$ **Discriminante Gaussiano Cuadrático (QD)**

Análisis de Discriminantes Gaussianos

¿Podemos averiguar la ec. de la superficie de decisión?

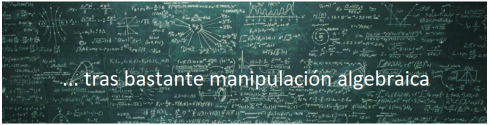
- Si $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}$, entonces

$$(\mathbf{z} - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}^{-1} (\mathbf{z} - \boldsymbol{\mu}_1) - (\mathbf{z} - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}^{-1} (\mathbf{z} - \boldsymbol{\mu}_0) = C.$$



- llegamos a una expresión lineal en $\mathbf{z}$

$$-\mathbf{z}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_1 - \boldsymbol{\mu}_1^T \boldsymbol{\Sigma}^{-1} \mathbf{z} + \mathbf{z}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_0 + \boldsymbol{\mu}_0^T \boldsymbol{\Sigma}^{-1} \mathbf{z} = C.$$



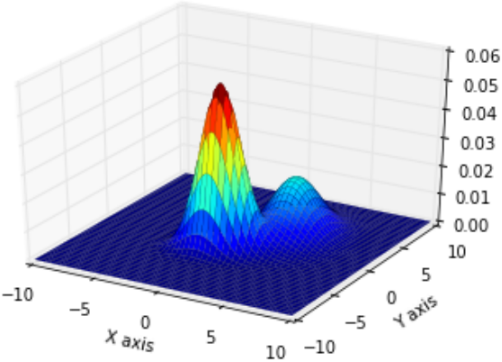
- se puede construir, un discriminante similar al del QDA pero lineal

  $\Rightarrow$ **Discriminante Gaussiano Lineal (LD)**

Contenidos

Mezclas de Gaussianas

¿Y si mis datos se *concentran* en **dos** puntos?



- Podemos utilizar **dos** MVNs, una para cada concentración, ¡¡ pero la PDF tiene que ser **válida** !!

Ej. $PDF(\mathbf{x}) = 30\,\% \cdot \mathrm{MVN}_1(\mathbf{x}) + 70\,\% \cdot \mathrm{MVN}_2(\mathbf{x})$

Mezclas de Gaussianas

GMM

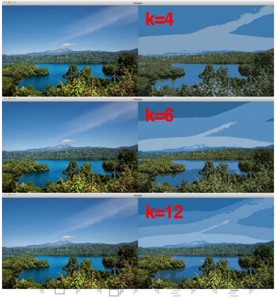
$$p(\mathbf{x}) = \sum_{i=1}^{K} \pi_i \cdot \mathrm{MVN}(\mathbf{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \quad \text{sujeto a } \sum_{i=1}^{K} \pi_i = 1.$$

Clasificación

- El GMM es la verosimilitud $p(\mathbf{x}|t)$

Agrupamiento

- Un GMM de K MVNs da lugar a K grupos
- El grupo i-esimo tendrá como centroide la media $\boldsymbol{\mu}_i$.
- Aplicación en compresión de imágenes o la cuantización de colores



Mezclas de Gaussianas

Algoritmo EM

- Parámetros del modelo:
  - la media y la covarianza de cada MVN que forma la mezcla,
  - el peso que cada MVN tiene en la mezcla,
  - para cada una de las etiquetas diferentes que hay.
- Datos: Conjunto de ejemplos etiquetados 0 ó 1

x_1	x_2	t
2,36	0,31	0
4,49	-1,42	0
1,12	2,13	0
1,23	-0,42	1
-0,66	-1,64	1
1,98	1,36	0
2,01	2,97	1
0,62	0,76	0
4,17	-1,03	0
2,60	1,89	1
3,50	0,86	1
1,24	1,35	0
2,57	0,88	1
1,86	3,41	0
2,77	3,11	1
-0,15	-1,29	1
0,67	-0,64	1
0,61	-0,37	0
0,83	0,61	0
2,02	-0,96	0
1,13	1,10	0
1,72	1,53	0
3,11	3,05	1
3,10	2,67	1
3,28	1,63	1
3,10	1,06	1
0,90	0,00	0
2,06	1,52	0
1,54	-0,14	1
0,28	2,70	0

x_1	x_2	t
2,36	0,31	0
4,49	-1,42	0
1,12	2,13	0
1,98	1,36	0
0,62	0,76	0
4,17	-1,03	0
1,24	1,35	0
1,86	3,41	0
0,61	-0,37	0
0,83	0,61	0
2,02	-0,96	0
1,13	1,10	0
1,72	1,53	0
0,90	0,00	0
2,06	1,52	0
0,28	2,70	0
1,23	-0,42	1
-0,66	-1,64	1
2,01	2,97	1
2,60	1,89	1
3,50	0,86	1
2,57	0,88	1
2,77	3,11	1
-0,15	-1,29	1
0,67	-0,64	1
3,11	3,05	1
3,10	2,67	1
3,28	1,63	1
3,10	1,06	1
1,54	-0,14	1

x_1	x_2	t
1,73	1,53	0
0,22	1,39	0
1,09	2,62	0
-1,67	2,14	0
1,52	1,02	0
0,53	4,39	0
2,03	-0,53	0
3,13	1,32	0
3,06	-1,30	0
2,52	1,42	0
1,48	1,46	0
2,16	1,90	0
1,11	0,40	0
2,64	-1,01	0
2,92	-2,73	0
3,50	0,16	0

x_1	x_2	t
1,14	-0,49	1
0,82	-1,39	1
2,96	-0,99	1
3,39	0,41	1
2,42	0,46	1
2,02	1,53	1
2,22	2,19	1
1,85	0,27	1
0,49	0,83	1
3,17	1,98	1
3,48	1,58	1
4,41	2,69	1
4,07	1,69	1
2,23	0,87	1

Algoritmo EM

Sea el **factor de pertenencia** de un ejemplo $\mathbf{x}$ a la k-esima MVN:

$$w_k = \frac{\pi_k \cdot \mathrm{MVN}_k(\mathbf{x})}{\sum_{i=1}^{K} \pi_k \cdot \mathrm{MVN}_k(\mathbf{x})},$$

Paso E

- Calcular el factor de pertenencia de cada uno de los ejemplos
- Obtenemos una tabla de valores $W = \{w_{i,k}\}$ para $i = 1 \ldots M$, $k = 1 \ldots K$; donde M es el número de ejemplos y cada fila suma 1.

Algoritmo EM

Paso M

- Comenzamos calculando $M_k = \sum_{i=1}^{M} w_{i,k}$, es decir la suma de la k-esima columna de la tabla W, para $k = 1 \dots K$.
- A continuación actualizamos los K pesos con la regla:

$$\pi_k^{\text{next}} = M_k / M$$

- Actualizamos la media y covarianza de cada MVN_k.

$$\boldsymbol{\mu}_k^{\text{next}} = \frac{1}{M_k} \sum_{i=1}^{M} \left(w_{i,k} \mathbf{x}^{(i)} \right)$$

$$\boldsymbol{\Sigma}_k^{\text{next}} = \frac{1}{M_k} \sum_{i=1}^{M} \left(w_{i,k} (\mathbf{x}^{(i)} - \boldsymbol{\mu}_k^{\text{next}})(\mathbf{x}^{(i)} - \boldsymbol{\mu}_k^{\text{next}})^T \right)$$

Algoritmo EM

Arranque

- Dar unos valores iniciales a

$$\{\pi_1, \ldots, \pi_K,\ \boldsymbol{\mu}_1, \ldots \boldsymbol{\mu}_K,\ \boldsymbol{\Sigma}_1 \ldots \boldsymbol{\Sigma}_K\}$$

  Por ej.

  $\pi_1 = \ldots = \pi_K = 1/K$
  Por cada MNV, calcular $\boldsymbol{\mu}$ y $\boldsymbol{\Sigma}$ de un subconjunto de puntos aleatorio.

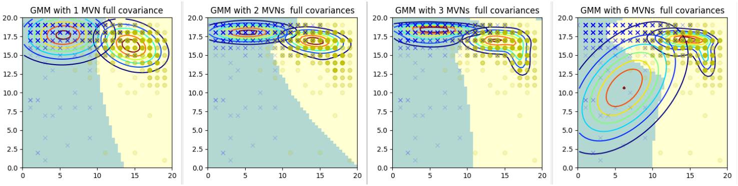


Figura: Cuatro GMMs con diferente número de MVN componentes para el conjunto de datos '0' vs. '1'. [Fuente: Original de A. Cuesta]