Modelos Gaussianos

Índice

Referencia principal

- Cap.5 de "Python Data Science Handbook", sección "In Depth: Gaussian Mixture Models"
- Secciones 4.1, 4.2 y 11.4 de *"Machine Learning. A probabilistic perspective"*
- Documentación de Scikit-Learn sobre Modelos de mezclas de Gaussianas

---

Al terminar este tema tendremos las herramientas necesarias para crear modelos generativos de características continuas con distribuciones multimodales. Utilizaremos el algoritmo de Esperanza-Maximización.

1. Introducción

En la semana anterior introdujimos la aproximación probabilística y generativa a la construcción de clasificadores. Desde este punto de vista el objetivo es construir la densidad de probabilidad que genera conjuntamente los datos y las etiquetas $p(\mathbf{X}, \mathbf{t})$. Mediante la regla de la cadena esta distribución se puede descomponer en la distribución a priori sobre las etiquetas y la verosimilitud de los datos para cada una de ellas, es decir $p(\mathbf{X}, \mathbf{t}) = p(\mathbf{t})p(\mathbf{X}|\mathbf{t})$.

En general la construcción de $p(\mathbf{X}|\mathbf{t})$ es complicada si no se realizan algunas suposiciones. En la semana pasada asumíamos la independencia de las características dada la etiqueta, y obteníamos el clasificador NBC.

Esta semana vamos a asumir que las características están interrelacionadas de manera que cada ejemplo de una misma etiqueta es una muestra obtenida de una distribución Normal multivariada (MVN) con unos parámetros concretos, asociados a dicha etiqueta; y que los parámetros para una clase son diferentes a los parámetros para las otras.

2. La distribución normal multivariada (MVN)

La MVN es probablemente la distribución de densidad de probabilidad más empleada de todas debido a sus propiedades. En esta sección vamos a hacer un resumen de éstas previo a la construcción de modelos más complejos a partir de ellas.

Función de densidad La MVN es la generalización a n dimensiones de la distribución Normal o Gaussiana $\mathcal{N}(x)$. Para comprender mejor su expresión abajo se pueden ver y comparar ambas.

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

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

- 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 $n \times n$
- $\sqrt{\left((\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)}$ es la distancia Mahalanobis desde $\mathbf{x}$ a la media $\boldsymbol{\mu}$

Es decir que MVN tiene dos parámetros que son la generalización a n dimensiones de los parámetros de la distribución Normal. La interpretación de estos es también similar. La media es el punto del espacio en el que se alcanza la moda (el máximo) de la distribución, y la covarianza está relacionada con su anchura.

La media Cuando se utiliza la distribución normal para modelar un conjunto de m datos unidimensionales, el estimador MLE del valor esperado es el promedio de los datos.

$$\mu = \mathbb{E}[\mathbf{X}] = \frac{1}{m}\sum_{i=1}^{m} x^{(i)}$$

Cuando se trata de la distribución MVN, el parámetro $\boldsymbol{\mu}$ es un vector donde el componente j es la media de la característica j-esima.

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

La matriz de covarianza La varianza σ^2 es el valor esperado $\mathbb{E}$ de las distancias al cuadrado entre cada uno de los ejemplos (unidimensionales) del conjunto de datos y su valor medio. Esta medida está relacionada con la anchura de una distribución normal, que es prácticamente igual a 6σ; en concreto el 99.4 % de su área está en el intervalo $[-3\sigma, 3\sigma]$. Cuando se utiliza la distribución normal para modelar un conjunto de datos se define como:

$$\sigma^2 = \mathbb{E}\left[(\mathbf{X} - \mu)(\mathbf{X} - \mu)\right],$$

Cuando los ejemplos son n-dimensionales, y asumimos que tienen una distribución normal conjunta (es decir una MVN), entonces la varianza se generaliza en la matriz de covarianza $\mathbf{\Sigma}$, donde el elemento de la fila i y la columna j se define como

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

Intuitivamente, los elementos de la diagonal son la varianza de cada una de las características por separado, sin tener en cuenta la otra. Por tanto una MVN con una matriz de covarianza diagonal modela n variables independientes normalmente distribuidas, es decir

$$\text{MVN}\left(\mathbf{x};\ \boldsymbol{\mu}, \mathbf{\Sigma} = \texttt{diag}(\sigma_1^2, \ldots, \sigma_n^2)\right) = \prod_{j=1}^{n} \mathcal{N}(x_j;\ \mu_j, \sigma_j^2).$$

Cuando existe una dependencia entre ellas los elementos fuera de la diagonal son distintos de cero. Por este motivo diremos que la matriz de covarianza impone una **estructura de dependencia** entre las variables que se modelan.

A modo de ejemplo, en la Figura 1 se muestran dos MVN de dos variables, centradas en $(0, 0)$. La MVN de la izquierda tiene $\mathbf{\Sigma} = \begin{bmatrix} 3 & 0 \\ 0 & 15 \end{bmatrix}$ y la de la derecha tiene $\mathbf{\Sigma} = \begin{bmatrix} 3 & -5 \\ -5 & 15 \end{bmatrix}$

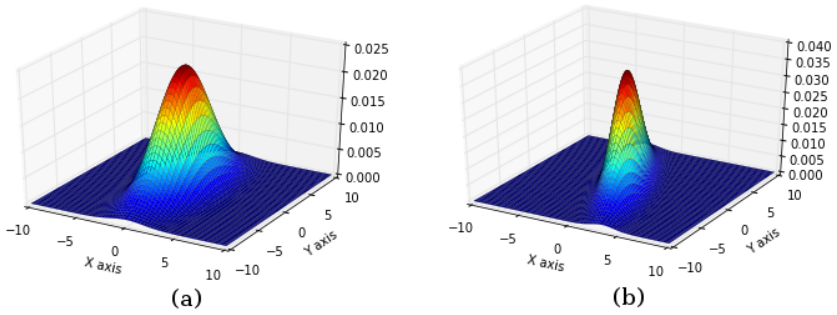


Figura 1: Dos MVN de dos variables cuya matriz de covarianza tiene la misma diagonal pero diferente Σ_{12}. (a) $\Sigma_{12} = 0$, (b) $\Sigma_{12} \neq 0$ [Fuente: Original de A. Cuesta]

Para ver mejor esta estructura vamos a obtener el contorno de la MVN para diferentes valores de su densidad.

Curvas de equidensidad Cualquier matriz cuadrada $n \times n$, $\mathbf{A}$, se puede descomponer como $\mathbf{A} = \mathbf{U}^T\mathbf{\Lambda}\mathbf{U}$, donde $\lambda = \texttt{diag}(\lambda_1, \ldots, \lambda_n)$, con $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ es la matriz formada con los autovalores ordenados de mayor a menor y $\mathbf{U}$ es la matriz cuya columna i se corresponde con el autovector normalizado ($\mathbf{u}_i^T\mathbf{u}_i = \mathbf{u}_i\mathbf{u}_i^T = 1$) del autovalor λ_i, Además, cuando todos los autovalores de una matriz son positivos la matriz se denomina "definida positiva".

La matriz de covarianza es definida positiva y de tamaño $n \times n$, por tanto admite esta descomposición y además podemos ordenar sus autovalores en orden decreciente.

Haciendo la inversa a ambos lados de la descomposición de la matriz de covarianza tenemos a

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

Ahora podemos reescribir la distancia Mahalanobis con esta expresión de $\mathbf{\Sigma}^{-1}$:

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

La expresión $(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{u}_i$ es un escalar porque $(\mathbf{x}-\boldsymbol{\mu})^T$ es un vector fila n-dimensional y $\mathbf{u}_i$ es un vector columna, también n-dimensional. El mismo escalar se obtiene también con $\mathbf{u}_i^T(\mathbf{x}-\boldsymbol{\mu})$.

Sea entonces $y_i = (\mathbf{x}-\boldsymbol{\mu})^T\mathbf{u}_i$ dicho escalar. 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 del lugar geométrico de todos los puntos $\mathbf{x}$ que están a distancia D de la media. Este lugar geométrico es una elipse centrada en $\boldsymbol{\mu}$, cuyo eje i-esimo tiene la dirección $\mathbf{u}_i$ y tamaño $\sqrt{\lambda_i}$.

Estas elipses son las "curvas de nivel" de la densidad de probabilidad de la MVN; es decir las curvas de **equidensidad** porque cada curva es el contorno de la densidad conjunta para una cierta "altura", o sea para una cierta densidad. En la Figura 2 se resumen estos resultados.

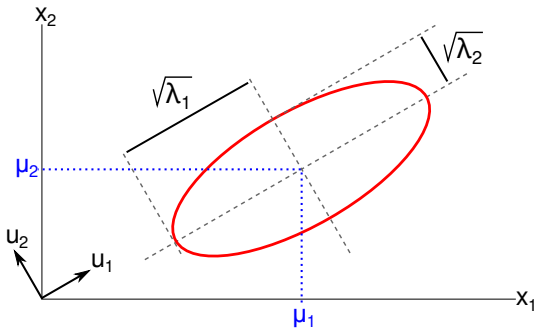


Figura 2: Curva de equidensidad de una MVN de 2 variables. [Fuente: Original de A. Cuesta]

¿Por qué se utiliza tanto la MVN? Hay varios motivos

- Por un lado los estimadores MLE de sus parámetros son muy "familiares" y tienen una interpretación sencilla.
- Cada una de las marginales de un MVN son a su vez distribuciones normales, pero de una sóla dimensión obviamente.

  Es decir que la distribución de cada cada una de sus variables por separado es a su vez una distribución normal, tanto si la matriz de covarianza es diagonal como si no.
- La distribución MVN es la distribución de máxima entropía para una media y una matriz de covarianza dadas.

  Es decir que, si se pretende modelar la distribución de unos datos imponiendo una media y una covarianza (que pueden ser estimadas de los mismos datos), entonces la MVN es la elección menos informativa, o sea que menos suposiciones hace y por tanto la más general.
- Otro motivo muy importante para algunas técnicas como el Filtro de Kalman, es que cuando marginalizamos y condicionamos MVNs obtenemos de nuevo MVNs. En este curso no vamos a ver esta propiedad, pero es conveniente ser conocerla.

3. Análisis de Discriminantes Gaussianos

Si asumimos que el subconjunto de ejemplos con etiqueta t están distribuidos según una MVN con media $\boldsymbol{\mu}_t$ y covarianza $\boldsymbol{\Sigma}_t$ entonces construir el modelo de verosimilitud es inmediato:

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

En la Figura 3 se muestran las densidades de cada clase con un mapa de calor y el resultado de clasificar todo el espacio con la etiqueta MAP, es decir:

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

En el contexto de este tema, esta maximización es equivalente a minimizar la distancia Mahalanobis de un punto $\mathbf{x}$ del espacio al centro de cada una de las MVN que modelan las diferentes clases; es decir

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

Cuando la matriz de covarianza es diagonal, entonces recuperamos el NBC Gaussiano. En caso contrario, la técnica resultante se denomina "Análisis de discriminantes Gaussianos".

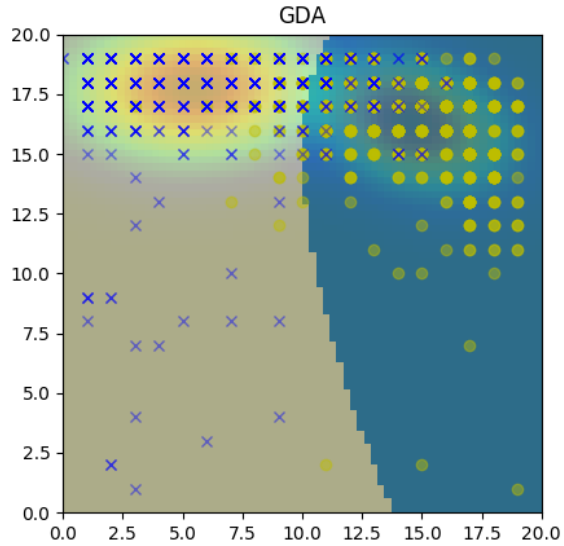


Figura 3: Superficie de decisión y densidades de las MVN para un GDA con el conjunto de datos '0' vs. '1'. [Fuente: Original de A. Cuesta]

3.1. Discriminante Gaussiano Cuadrático (QDA)

Cuando se trabaja con Gaussianas se pueden obtener fórmulas cerradas para las superficies de decisión. Para verlo supongamos un problema con dos clases $t = \{0, 1\}$, de tal manera que la distribución a posteriori de un nuevo ejemplo $\mathbf{z}$ es:

$$p(t=0|\mathbf{z}) \propto p(t=0)(2\pi|\boldsymbol{\Sigma}_0|)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{z} - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}_0^{-1}(\mathbf{z} - \boldsymbol{\mu}_0)\right)$$
$$p(t=1|\mathbf{z}) \propto p(t=1)(2\pi|\boldsymbol{\Sigma}_1|)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{z} - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}_1^{-1}(\mathbf{z} - \boldsymbol{\mu}_1)\right)$$

De manera que el logaritmo de su cociente es:

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

En la expresión anterior no aparece el símbolo de proporcionalidad ($\propto$) sino el de igualdad ($=$). El motivo es que la probabilidad a posteriori de la etiqueta 0 y de la etiqueta 1 tienen el mismo denominador. Por tanto al dividir una por otra se cancelan. Manipulando se llega a

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

La superficie de decisión será el conjunto de puntos $\mathbf{z}$ donde $\delta = 0$. 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. Esta expresión se puede resolver, con bastante álgebra, pero lo interesante es que proporciona también una manera directa de clasificar nuevos ejemplos según la distancia Mahalanobis de estos al centro de cada clase:

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

donde d_{M0} y d_{M1} son las distancias Mahalanobis al centro de la distribución de la clase 0 y la clase 1 respectivamente. El problema multiclase se aborda igual por cada par de etiquetas.

3.2. Discriminante Gaussiano Lineal (LDA)

Es el caso particular de QDA en el que MVN($\mathbf{X}|t=0$) y MVN($\mathbf{X}|t=1$) tienen la misma covarianza, o en general que todas las clases comparten la misma covarianza $\boldsymbol{\Sigma}$. En ese caso:

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

Del mismo modo que antes, la superficie de decisión será allí donde $\delta = 0$, por lo que si igualamos a cero y agrupamos todo lo que no depende de $\mathbf{z}$ en la constante C tenemos:

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

En este punto recordamos las siguientes propiedades de la traspuesta de una matriz:

(i) $(A+B)^T = A^T + B^T$
(ii) $(AB)^T = B^T + A^T$
(iii) $(A^T)^{-1} = (A^{-1})^T$

Aplicando estas propiedades obtenemos:

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

Los términos ($\mathbf{z}^T\boldsymbol{\Sigma}^{-1}\mathbf{z}$) se cancelan y los términos ($\boldsymbol{\mu}_1^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1$) y ($\boldsymbol{\mu}_0^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_0$) no dependen de $\mathbf{z}$, por tanto se pueden asimilar en la constante C, con lo que llegamos a una expresión lineal en $\mathbf{z}$

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

A partir de esta expresión se puede construir, con bastante manipulación, un discriminante similar al del QDA.

En la Figura 4 se muestra el resultado de QDA y de LDA sobre el conjunto de datos '0' vs. '1'.

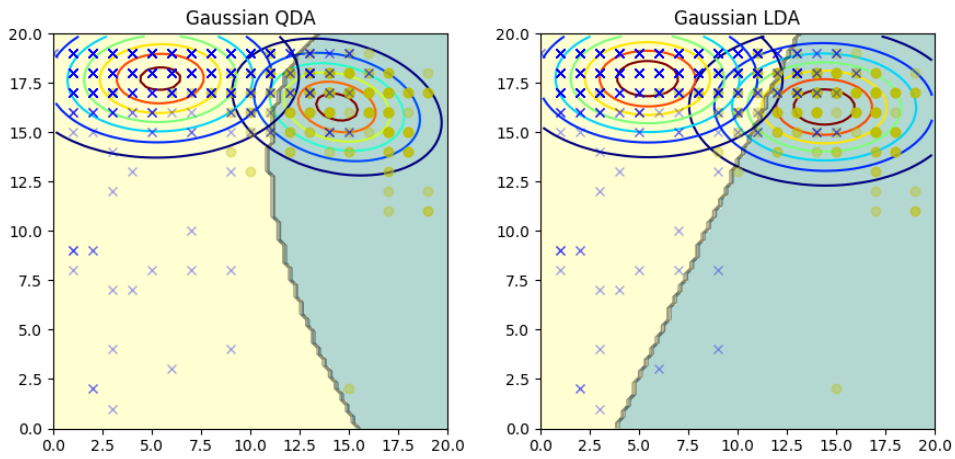


Figura 4: Superficie de decisión y curvas de contorno de las MVN de cada clase para un QDA (izq.) y un LDA (der.) [Fuente: Original de A. Cuesta]

4. Modelos de Mezclas de Gaussianas (GMM)

La MVN es una distribución unimodal, es decir sólo tiene una moda, o valor máximo de la función de densidad. ¿Qué podemos hacer si la distribución de los datos de una cierta clase tiene varias modas como en la Figura 5. Una manera de modelar este tipo de distribuciones es utilizar una

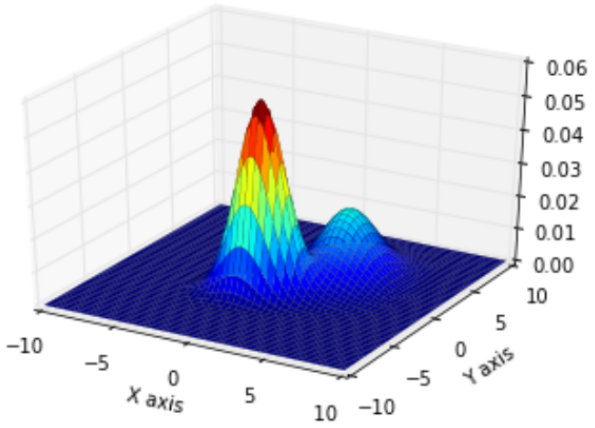


Figura 5: PDF conjunta bimodal. [Fuente: Original de A. Cuesta]

mezcla (también llamada mixtura) de Gaussianas multivariadas. Intuitivamente, esto consiste en construir la pdf conjunta como una suma MVNs de manera que cada moda se corresponde con la media de una MVN y la dispersión de los ejemplos alrededor de ella se captura con su covarianza.

Como estamos construyendo una distribución de densidad de probabilidad, la integral en todo el espacio debe ser igual a 1. Esto se consigue haciendo que cada una de las MVN que utilizamos como "ladrillo" para la construcción de la pdf conjunta tenga un cierto peso, y que la suma de todos los pesos sea igual a 1; es decir se trata de una **combinación lineal convexa**.

Por ejemplo, un GMM construido como suma de dos MVNs, la primera con un peso de 0.3 y la segunda con un peso de 0.7 significa que la primera contribuye al 30 % de la pdf conjunta, y la segunda contribuye al 70 % restante.

Formalmente, un GMM viene determinado por la combinación lineal de K MVNs, cada una de

ellas con su media y su covarianza, y ponderada por un peso π_k, para $k = 1 \dots K$, es decir:

$$p(\mathbf{x}) = \sum_{i=1}^{K} \pi_i \cdot \text{MVN}(\mathbf{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \quad \text{sujeto a} \quad \sum_{i=1}^{K} \pi_i = 1. \tag{2}$$

Clasificación vs. Agrupamiento La expresión (2) es general, en el sentido de que es válida para un ejemplo n-dimensional $\mathbf{x}$. Si lo queremos utilizar para clasificación entonces cada ejemplo tendrá asociada una etiqueta. En ese caso lo que construye el GMM es la verosimilitud $p(\mathbf{x}|t)$ es decir la distribución multimodal de los ejemplos de una determinada etiqueta. Pero esta técnica permite también realizar agrupamiento o *clustering*, es decir aprendizaje no supervisado. Recordar que en aprendizaje no supervisado la única intervención humana consiste en indicar cuantos grupos se deben formar. En ese caso, al construir el GMM de los datos con K MVNs obtendremos K grupos, de modo que el grupo i-esimo tendrá como centroide la media $\boldsymbol{\mu}_i$.

En reconocimiento de patrones estamos más interesados en aprendizaje supervisado, y será al que presetemos antención a continuación. Sin embargo, clustering con GMM se realiza exactamente con la misma técnica (Esperanza-Maximización), y además tiene aplicaciones importantes en visión artificial como la compresión de imágenes o la cuantización de colores.

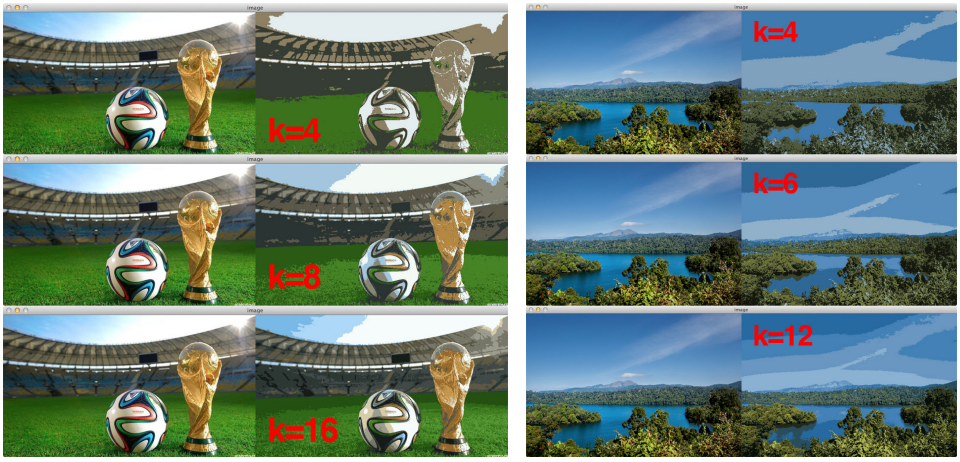


Figura 6: Ejemplos de cuantización del color utilizando *clustering*. [Fuente: PyImageSearch]

4.1. Aprendizaje con EM

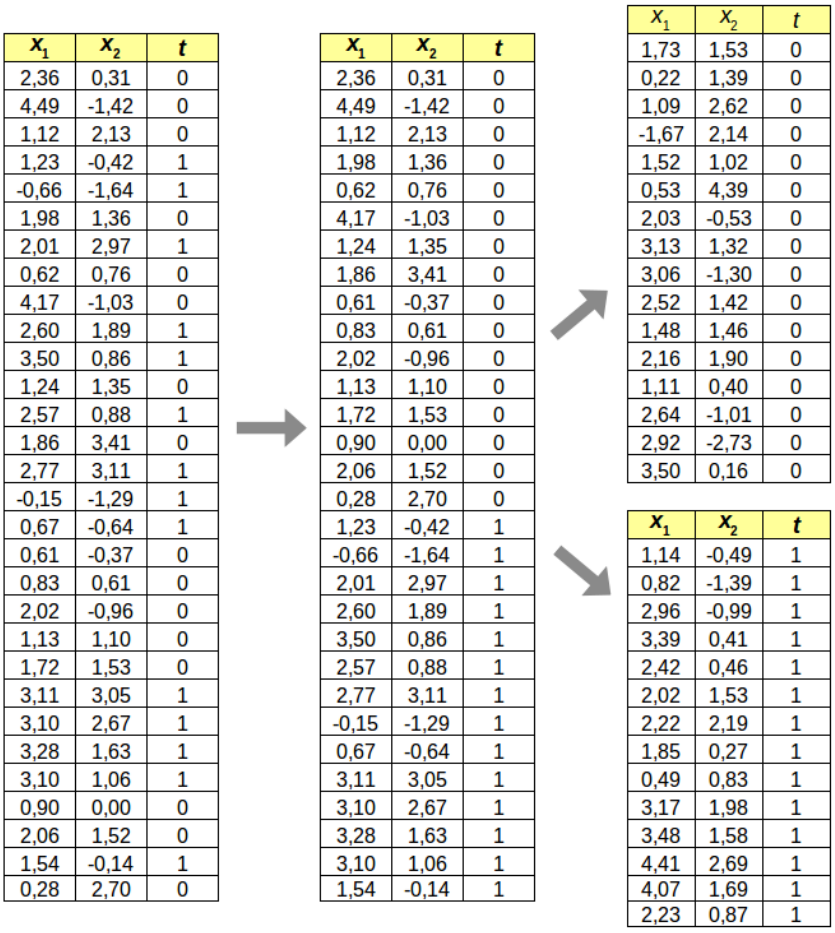
Los parámetros de nuestro modelo son:

- 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.

Es decir, si tenemos T etiquetas y construimos un GMM con K MVNs, el total del parámetros que se deben aprender es $2TK$.

El proceso de aprendizaje consiste en separar los datos por etiquetas y, para cada etiqueta t_i, aprender el GMM que modela esos datos. Esto significa estimar K medias y K covarianzas a partir de esos datos, por ejemplo con MLE. El problema, y la novedad, ahora es que no tenemos ninguna información sobre como separar estos datos (ya que todos tienen la misma etiqueta). Por ejemplo, supongamos que $K = 2$ y tenemos la tabla de ejemplos bidimensionales que se muestra en la Figura 7. Si nos fijamos en la tabla de arriba-derecha, o sea los ejemplos de la

clase "0", el siguiente paso sería estimar dos medias y la covarianzas, y el peso de cada MVN del GMM para $t = 0$. Pero ¿qué ejemplos de esa tabla tomamos para estimar $\boldsymbol{\mu}_1$ y cuáles para $\boldsymbol{\mu}_2$? Obviamente no podemos tomar todos porque entonces obtendría el mismo resultado para $\boldsymbol{\mu}_1$ y para $\boldsymbol{\mu}_2$. ¿Y cómo se estiman los pesos?



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

Figura 7: 30 ejemplos bidimensionales, separados por su etiqueta. ¿Cómo calculamos los parámetros de cada GMM? [Fuente: Original de A. Cuesta]

Un método de aprendizaje es el de **Esperanza-Maximización** (EM). Consiste en dos pasos, el paso E y el paso M, que se repiten alternativamente hasta que se alcanza un criterio de parada. Recordamos la notación:

- Un GMM es la pdf que vamos a construir para un conjunto de datos $\mathbf{x}$, todos con la misma etiqueta, es decir de la misma clase.
- Dicha GMM está compuesta por la suma convexa de K "MVN componentes", $\text{MVN}_{k=1\ldots K}$

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

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

que mide la probabilidad de que el ejemplo $\mathbf{x}$ haya sido generado por MVN_k. *Factor de pertenencia*

Paso E. Calculamos el factor de pertenencia de cada uno de los ejemplos del conjunto de datos, y para todas las MVN componentes.

De este modo obtenemos una tabla de valores $W = \{w_{i,k}\}$ para $i = 1 \dots M$, $k = 1 \dots K$; donde M es el número de ejemplos de dicho subconjunto y cada fila suma 1.

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

Y finalmente actualizamos todas las medias y covarianzas de cada MVN componente.

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

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

Arranque del algoritmo Para poder empezar necesitamos unos valores iniciales de

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

Por ej.

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

4.2. Inferencia

La etiqueta asignada a un ejemplo nuevo es la MAP. Es decir, una vez se conocen los pesos de las MVNs componentes para cada clase, a un ejemplo nuevo se le asigna la etiqueta $\hat{t}$:

$$\hat{t} = \arg\max_t \left(\frac{p(t)p(\mathbf{z}|t)}{p(\mathbf{z})} \right) = \arg\max_t p(t)p(\mathbf{z}|t).$$

donde el prior $p(t)$ se construye como el cociente de ejemplos de cada clase entre el total de ejemplos del conjunto de entrenamiento, como ya vimos la semana pasada.

Alternativamente, se puede asignar la etiqueta MLE, es decir:

$$\hat{t} = \arg\max_t (\mathbf{z}|t).$$

En la Figura 8 se muestran como clasifican cuatro GMMs construidos con diferente número de MVN componentes el espacio de muestras. Como el número de ceros es el mismo que el número de unos en el conjunto dado, la etiqueta estimada MAP y MLE coincidirán.

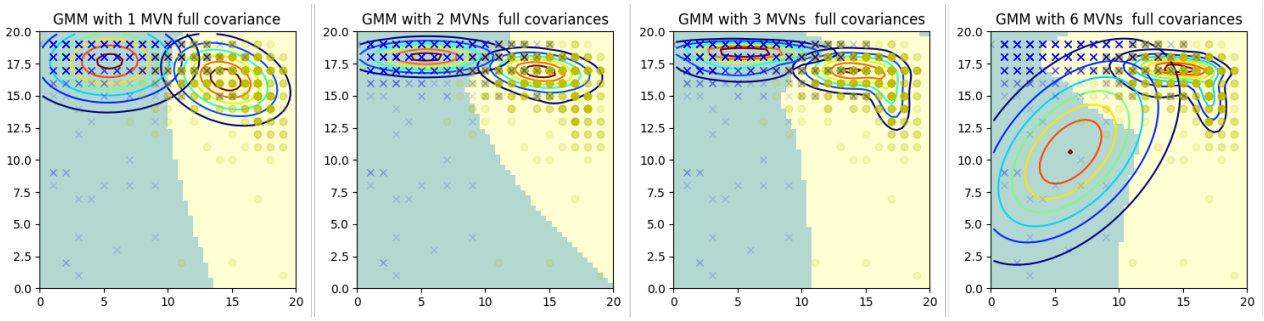


Figura 8: Cuatro GMMs con diferente número de MVN componentes para el conjunto de datos '0' vs. '1'. [Fuente: Original de A. Cuesta]

Índice alfabético