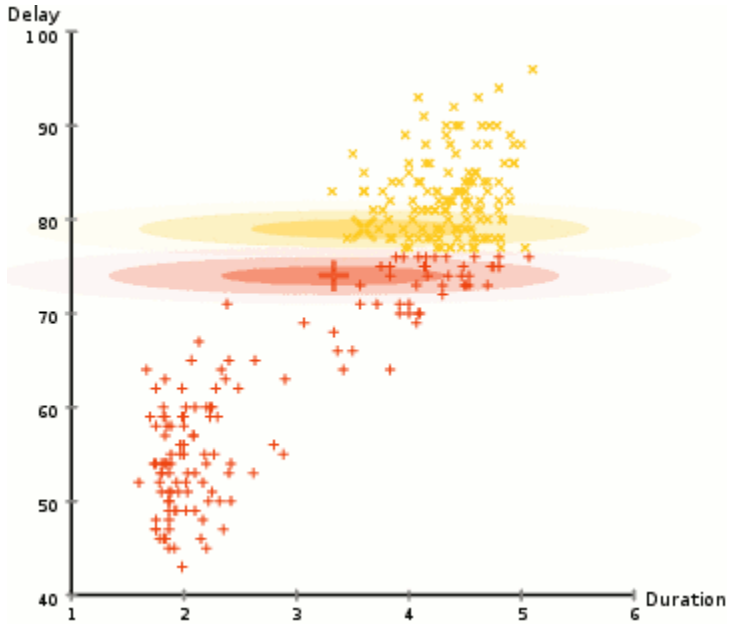
CLUSTERING. MIXTURE MODELS. EXPECTATION MAXIMIZATION

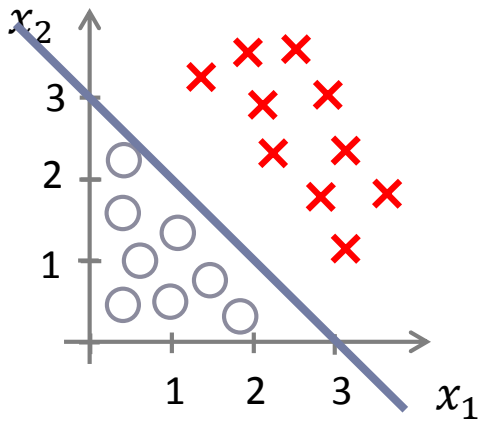
Mixture Models. Mezcla de modelos

- Suponemos que existen unos modelos que han generado los datos.
- Queremos encontrar los parámetros de esos modelos.



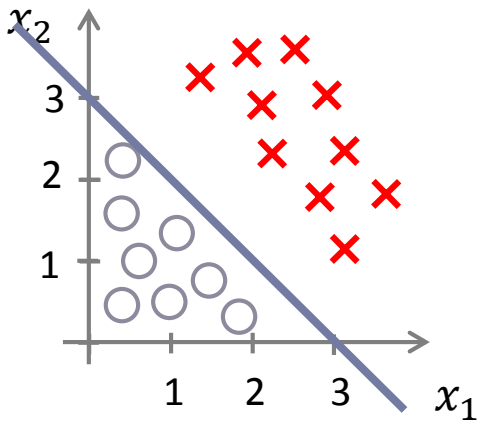
Mixture Models. Mezcla de modelos

- Antes de empezar, volvamos a clasificación:



Modelos discriminativos

- **Regresión logística**
  - Hipótesis relacionada con la probabilidad $h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}}$
  - Interpretación
    - Si $h_\theta(x) = 0.7$
      - Existe un 70% de probabilidad de que x pertenezca a la clase $y = 1$
  - P(y|x, θ)
  - Modelo discriminativo



Modelos generativos

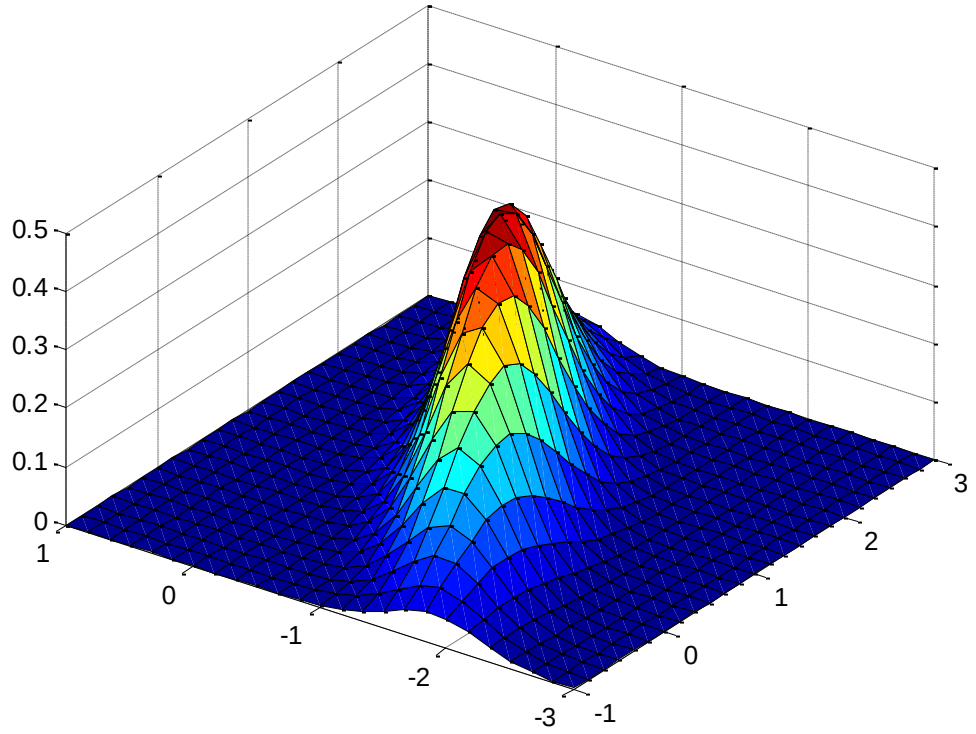
- Planteamiento diferente del problema:
  - Encontrar un modelo para cada clase
  - Para clasificar un nuevo ejemplo, vemos a cual de los modelos que hemos construido se parece más.
    - Probabilidad a priori de la clase $p(y)$
    - Modelos de la distribución de las características en cada clase $p(x|y=0)$ y $p(x|y=1)$
    - Teorema de Bayes $p(y|x) = \frac{p(x|y)p(y)}{p(x)}$
    - $argmax_y\, p(y|x)$

Modelos gaussianos

- El modelo generativo más común: modelamos $p(x|y)$ por medio de una distribución normal.
- Distribución normal multivariable
- $p(x|\mu, \Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} e^{(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu))}$
- Donde μ es la media y Σ la matriz de covarianzas

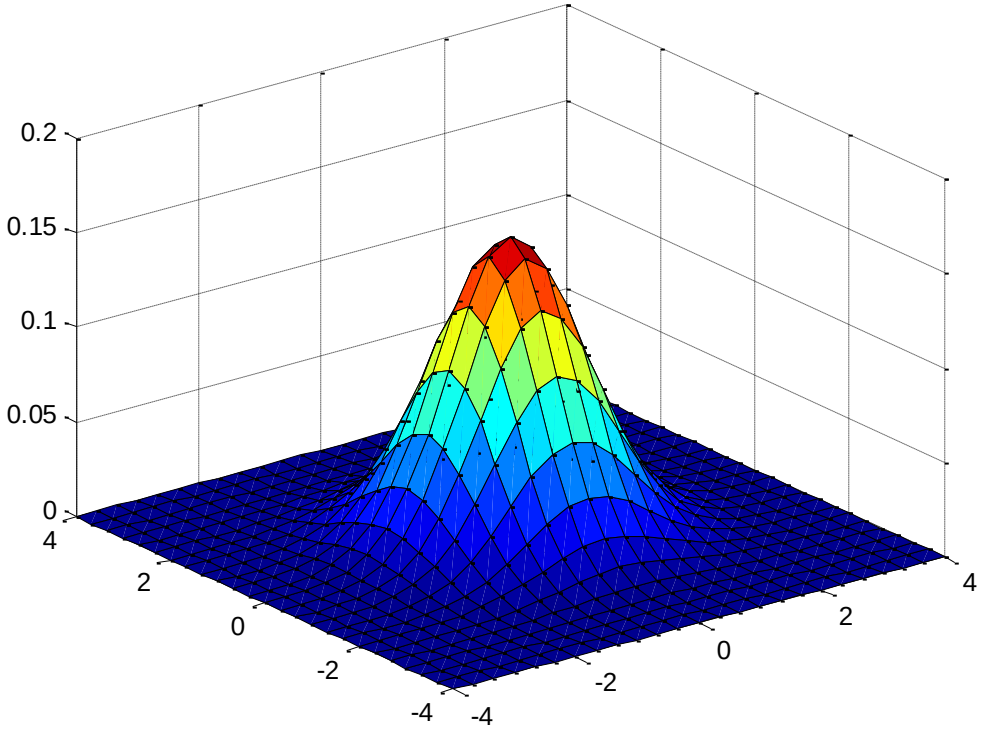
Distribución normal multivariable

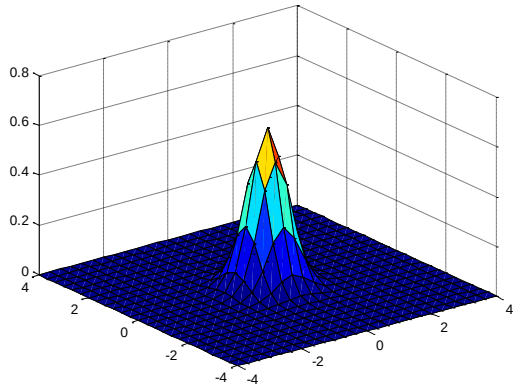
- Ejemplos:
- μ = [1 -1]; Σ = [0.9 0.4; 0.4 0.3];



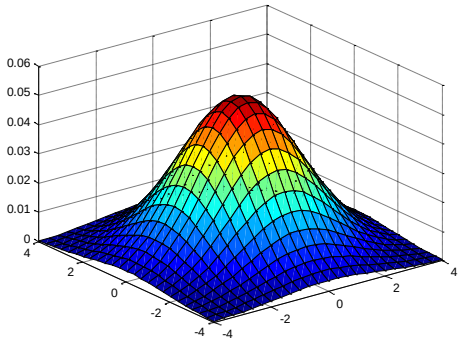
Distribución normal multivariable

- μ = [0 0]; Σ = [1 0; 0 1];



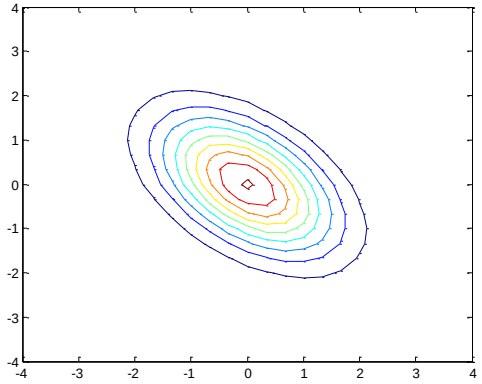


Σ = [0.25 0; 0 0.25];

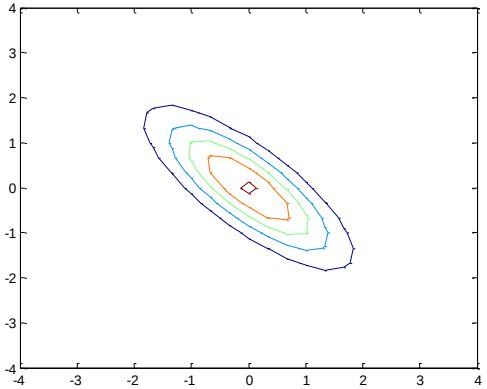


Σ = [3 0; 0 3];

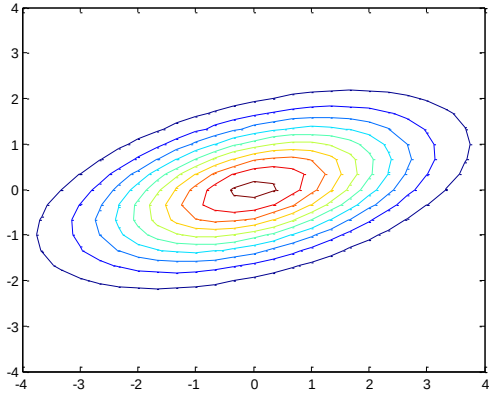
Distribución normal multivariable



Σ = [1 -0.5; -0.5 1];



Σ = [1 -0.8; -0.8 1];



Σ = [3 0.8; 0.8 1];

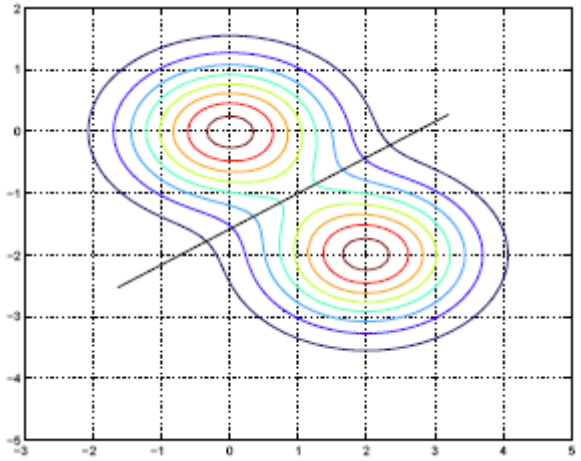
Linear Discriminant Analysis

- En LDA se asume que la distribución de cada clase k es una distribución gaussiana multivariable.
- Además en LDA, se añade una simplificación: que la matriz de covarianzas Σ es la misma para todas las clases.
- Debido a esta suposición, es un clasificador lineal.

Linear Discriminant Analysis

- Frontera de decisión. Función discriminante
- $p(y_1|x) = p(y_2|x)$

$$-\frac{1}{2}(\mathrm{x}-\mu_1)^T\Sigma^{-1}(x-\mu_1)+\log\big(p(y_1)\big)=-\frac{1}{2}(x-\mu_2)^T\Sigma^{-1}(x-\mu_2)+\log(p(y_2))$$



Linear Discriminant Analysis

- Fase de entrenamiento. Encontrar los parámetros de las distribuciones:
- Si derivamos la función de máxima verosimilitud:
- $p(y) = \frac{m_y}{m}$ % de elementos de la clase y
- $\mu_y = \frac{1}{m_y} \sum_{i=1}^{m} x^{(i)} \{y^{(i)} = y\}$ media de los ejemplos de la clase y
- $\Sigma = \frac{1}{m} \sum_{i=1}^{m} (x^{(i)} - \mu_{y^{(i)}}) (x^{(i)} - \mu_{y^{(i)}})^T$

Quadratic Discriminant Analysis

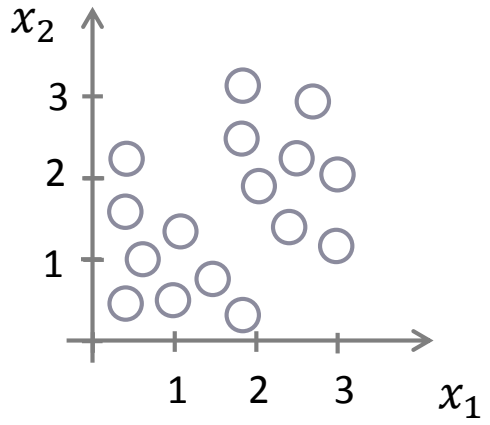
- En el QDA eliminamos la simplificación de que la matriz de covarianzas es la misma para todas las clases.
- Tendremos una matriz de covarianzas diferente para cada clase.
- $\Sigma_y = \frac{1}{m_y} \sum_{i=1}^{m} (x^{(i)} - \mu_{y^{(i)}}) \, (x^{(i)} - \mu_{y^{(i)}})^T \{y^{(i)} = y\}$

LDA y regresión logística

- $p(y = 1 | p(y), \mu_{y=1}, \mu_{y=0}, \Sigma) = \frac{1}{1+e^{-\theta^T x}}$
- Donde θ es una función de $p(y), \mu_{y=1}, \mu_{y=0}, \Sigma$
- Resulta que es la regresión logística
  - Cual es mejor?
  - LDA hace unas **suposiciones** muy fuertes
    - Distribución normal de los datos
    - Idéntica matriz de covarianzas
  - Cuando **se cumplen estas suposiciones LDA es asintóticamente eficiente** (si la cantidad de datos es grande, no hay ningún algoritmo mejor)
  - Al hacer menos suposiciones, **RL es más robusto**

Mixture Models

- Cuando no tenemos las etiquetas de los datos.
- Queremos encontrar las distribuciones que han generado esos datos.
- Modelos generativos.



Máxima Verosimilitud

- Tenemos que encontrar los parámetros que definen una distribución
- Sean un conjunto de datos, pensamos que se distribuyen según una distribución de probabilidad $p(x|\Theta)$ siendo Θ los parámetros.
- Estimación: encontrar los parámetros óptimos, es decir, que más se ajusten a los datos.
- La función de verosimilitud de los parámetros dados esos datos es:

Si consideramos que los datos son independientes

$$L(\Theta|D) = \prod_{i=1}^{m} p(x^{(i)}|\Theta)$$

Máxima Verosimilitud

- Entonces los parámetros óptimos son:
- $\Theta^* = argmax \prod_{i=1}^{m} p(x^{(i)}|\Theta)$
- Los podemos calcular igualando a cero:
- $\frac{\partial}{\partial \Theta} Log\big(L(\Theta|D)\big) = 0$
- Si es una distribución normal única entonces nos encontramos con el problema de los modelos generativos. Derivamos y encontramos las ecuaciones para calcular la media y la varianza (matriz de covarianzas si es multidimensional)

Mixture Models

- ¿Qué ocurre si los datos provienen de una combinación de distribuciones, en lugar de una sola?
- Mixture Models. Mezcla de modelos:

$$P\left(x^{(i)}\right) = \sum_{j=1}^{nc} P\left(c_j\right) P\left(x^{(i)} \middle| c_j\right)$$

$$\sum_{j=1}^{nc} P\left(c_j\right) = 1$$

No podemos resolver la derivada analíticamente

Expectation-Maximization

- El algoritmo EM es un método iterativo para encontrar la estimación de máxima verosimilitud de los parámetros en modelos estadísticos, en donde el modelo depende de unas variables ocultas.
- En LDA, QDA, conocíamos $y^{(i)}$ la clase, o el clúster, al que pertenece cada dato. Ahora esas variables no las conocemos, están ocultas.

Expectation-Maximization

- El algoritmo EM repite iterativamente el siguiente proceso:
  - primero estima (E) la verosimilitud utilizando la estimación actual de los parámetros
  - y después maximiza (M), calcula los parámetros que maximizan la verosimilitud esperada calculada en el paso (E).
  - Estos parámetros estimados son usados para determinar la distribución de las variables ocultas en el siguiente paso (E).

Expectation-Maximization

- Verosimilitud: $L(\Theta|D) = \prod_{i=1}^{m} p(x^{(i)}|\Theta)$
- Las variables ocultas les llamamos Z, por tanto $L(\Theta|D, Z)$
- **Expectation**: Calcular la (log) verosimilitud, con respecto a una distribución condicional Z, dados los datos D utilizando unos parámetros estimados Θ^t

$$Q(\Theta|\Theta^t) = E_{Z|D,\Theta^t}[\log(L(\Theta|D, Z)]$$

- **Maximization**: Encontrar los parámetros que maximicen este valor

$$\Theta^{t+1} = argmax_{\Theta} Q(\Theta|\Theta^t)$$

Expectation Maximization. Mezcla de gaussianas

- Si sabemos que son distribuciones gaussianas $\Theta = (\mu_1, \sigma_1, .., \mu_K, \sigma_K)$
- En este caso, las variables Z no son 0 ó 1, como en el K-means, sino que cada ejemplo tendrá una probabilidad para pertenecer a cada cluster.
- Es lo que se llaman probabilidades de pertenencia.
- $P(z^{(i)} = j | x^{(i)}, \Theta^t)$
- Lo podemos calcular con el teorema de Bayes:
- $$P\left(z^{(i)} = j \middle| x^{(i)}, \Theta^t\right) = \frac{P\left(z^{(i)}=j\right) P(x^{(i)}, \Theta^t | z^{(i)}=j)}{P(x^{(i)})}$$

Expectation Maximization. Mezcla de gaussianas

- **Paso E:** Para todos los ejemplos y todos los clústeres

$$P\left(x^{(i)}\middle|z^{(i)}=j\right)=N(\mu_j,\sigma_j)$$

$$P\left(z^{(i)}=j\middle|x^{(i)}\right)=\frac{P(c_j)P(x^{(i)},|z^{(i)}=j)}{\sum_{k=1}^{K}P(z^{(i)}=k)P(x^{(i)},|z^{(i)}=k)}$$

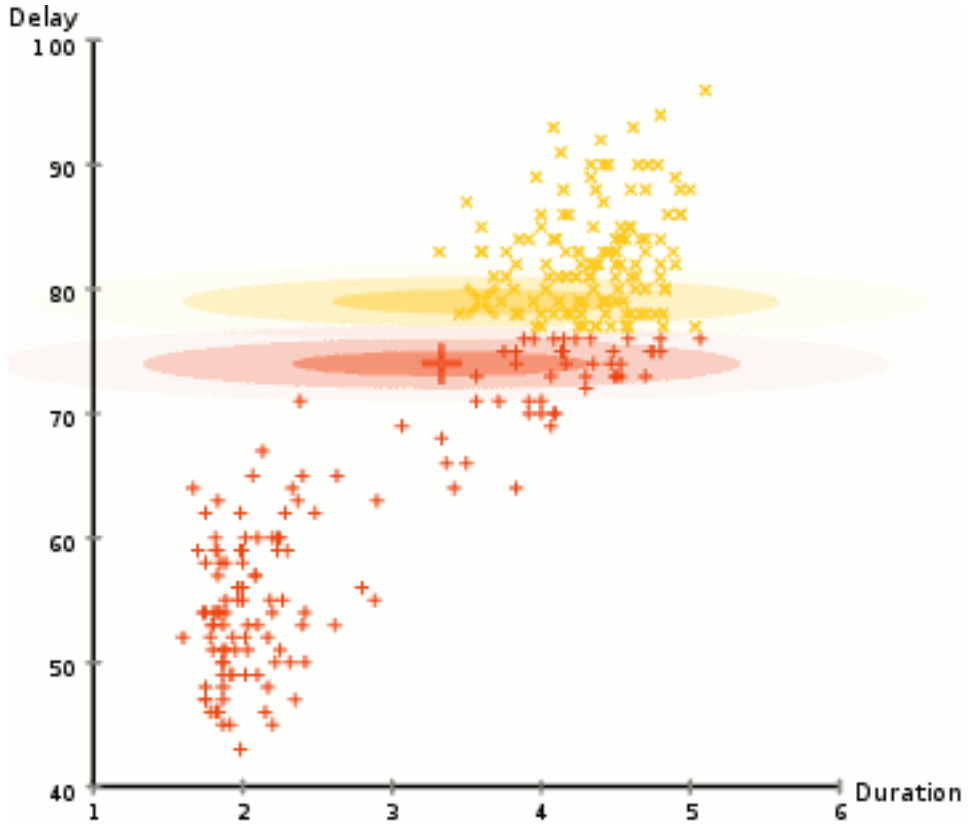
- **Paso M:** Para todos los clústeres

$$P(c_j)=\frac{1}{m}\sum_{i=1}^{m}P\left(z^{(i)}=j\middle|x^{(i)}\right)$$

$$\mu'_j=\frac{\sum_{i=1}^{m}x^{(i)}P\left(z^{(i)}=j\middle|x^{(i)}\right)}{\sum_{i=1}^{m}P\left(z^{(i)}=j\middle|x^{(i)}\right)}$$

$$\sigma'^{2}_j=\frac{\sum_{i=1}^{m}(x^{(i)}-\mu_j)^2P\left(z^{(i)}=j\middle|x^{(i)}\right)}{\sum_{i=1}^{m}P\left(z^{(i)}=j\middle|x^{(i)}\right)}$$

Expectation Maximization. Mezcla de gaussianas



EM y K-means

- Distribución normal multivariable:
- $p(x|\mu,\Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} e^{(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu))}$
- Donde μ es la media y Σ la matriz de covarianzas
- Σ va a ser una matriz diagonal

- ¿Existe relación entre EM y K-means?
- K-means es un caso especial de EM:
- Si todos los $P\left(z^{(i)} = j\right)$ son iguales y la varianza es la identidad