Algoritmo EM

Lección 19

Dr. Pablo Alvarado Moya

CE5506 Introducción al reconocimiento de patrones
Área de Ingeniería en Computadores
Tecnológico de Costa Rica

II Semestre, 2019

Contenido

Estimación de densidad probabilística

- Cuando revisamos el método de k-NN estudiamos varias técnicas **no paramétricas** de estimación de densidad probabilística:
  - Histogramas
  - Uso de kernels
  - Uso de k vecinos más cercanos
- La estimación de densidad es un proceso de aprendizaje **no supervisado**
- Aplicaciones de estimación: detección de atipicidades, estimación de densidades para métodos generativos, etc.
- Ahora nos abocaremos al tema de estimación **paramétrica** de densidades
- En concreto, a la estimación de una mezcla de gaussianas

Mezcla de gaussianas

- Sea un conjunto de datos de entrenamiento $\{\underline{\mathbf{x}}^{(1)}, \underline{\mathbf{x}}^{(2)}, \ldots, \underline{\mathbf{x}}^{(m)}\}$ (sin etiquetas)
- Deseamos modelar la distribución conjunta

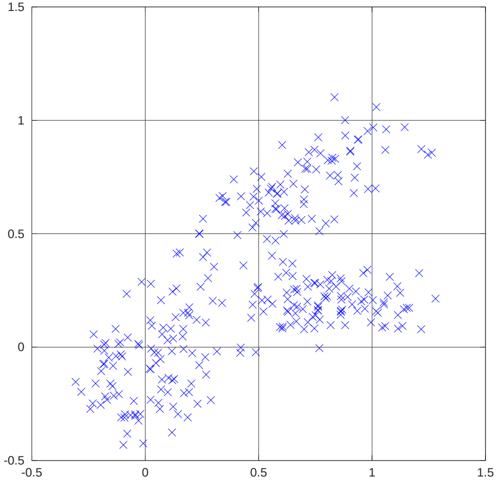
$$p(\underline{\mathbf{x}}^{(i)}, z^{(i)}) = p(\underline{\mathbf{x}}^{(i)} | z^{(i)}) p(z^{(i)}) \qquad \underline{\mathbf{x}}^{(i)} | (z^{(i)} = j) \sim \mathcal{N}(\underline{\boldsymbol{\mu}}_j, \boldsymbol{\Sigma}_j)$$

- La variable **latente** $z^{(i)} \sim \text{Multinomial}(\underline{\phi})$ indica cuál de k distribuciones gaussianas da origen al dato $\underline{\mathbf{x}}^{(i)}$

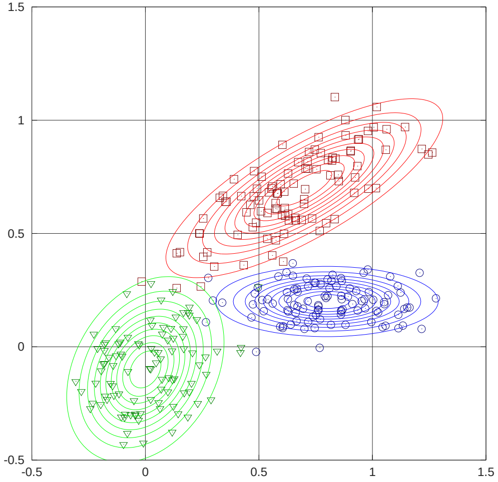
$$\phi_j > 0, \quad \sum_{j=1}^{k} \phi_j = 1, \quad \phi_j = p(z^{(i)} = j)$$

- "Latente" se refiere a que la variable no es observable
- Esas variables complican la estimación
- Esto se conoce como modelo de **mezcla de gaussianas**

Ejemplo



Ejemplo



Parámetros de la mezcla de gaussianas

- Los parámetros del modelo son $\underline{\phi}$, $\underline{\boldsymbol{\mu}}$ y $\boldsymbol{\Sigma}$
- La estimación se hace con la verosimilitud logarítmica

$$\begin{aligned}\ell(\underline{\phi},\underline{\boldsymbol{\mu}},\boldsymbol{\Sigma}) &= \sum_{i=1}^{m} \ln p(\underline{\mathbf{x}}^{(i)};\underline{\phi},\underline{\boldsymbol{\mu}},\boldsymbol{\Sigma}) \\ &= \sum_{i=1}^{m} \ln \sum_{z^{(i)}=1}^{k} p(\underline{\mathbf{x}}^{(i)}|z^{(i)};\underline{\boldsymbol{\mu}},\boldsymbol{\Sigma})p(z^{(i)};\underline{\phi})\end{aligned}$$

- Si calculamos $\nabla\ell(\underline{\phi},\underline{\boldsymbol{\mu}},\boldsymbol{\Sigma}) = 0$ encontramos que no es posible encontrar los parámetros que maximizan la verosimilitud, en forma cerrada

Caso hipotético con variables latentes conocidas

- Si se conocieran las variables latentes $z^{(i)}$, el problema de máxima verosimilitud se simplifica:

$$\ell(\underline{\phi},\underline{\boldsymbol{\mu}},\boldsymbol{\Sigma}) = \sum_{i=1}^{m} \ln p(\underline{\mathbf{x}}^{(i)}|z^{(i)};\underline{\boldsymbol{\mu}},\boldsymbol{\Sigma}) + \ln p(z^{(i)};\underline{\phi})$$

- Maximizando respecto a los parámetros resultaría en

$$\phi_j = \frac{1}{m}\sum_{i=1}^{m} 1\left\{z^{(i)} = j\right\} \qquad \underline{\boldsymbol{\mu}}_j = \frac{\sum_{i=1}^{m} 1\left\{z^{(i)} = j\right\}\underline{\mathbf{x}}^{(i)}}{\sum_{i=1}^{m} 1\left\{z^{(i)} = j\right\}}$$

$$\boldsymbol{\Sigma}_j = \frac{\sum_{i=1}^{m} 1\left\{z^{(i)} = j\right\}(\underline{\mathbf{x}}^{(i)} - \underline{\boldsymbol{\mu}}_j)(\underline{\mathbf{x}}^{(i)} - \underline{\boldsymbol{\mu}}_j)^T}{\sum_{i=1}^{m} 1\left\{z^{(i)} = j\right\}}$$

- Esto es muy similar al GDA con los $z^{(i)}$ sustituyendo a las etiquetas de las clases

Algoritmo EM

- El algoritmo EM (esperanza-maximización) (*expectation-maximization*) busca estimar no solo los parámetros, sino también las variables latentes $z^{(i)}$
- Tiene dos pasos: **paso E** y **paso M** (de ahí su nombre)
- El **paso E** estima los valores de $z^{(i)}$
- El **paso M** actualiza los parámetros de las gaussianas

Algoritmo EM (pseudocódigo) (1)

repeat

    // Paso E
    foreach $(i,j) \in (1,\ldots,m) \times (1,\ldots,k)$ **do**
        $w_j^{(i)} := p(z^{(i)} = j | \underline{\mathbf{x}}^{(i)}; \underline{\phi}, \underline{\boldsymbol{\mu}}, \boldsymbol{\Sigma})$
    end
    // Paso M
    Actualice los parámetros

$$\phi_j := \frac{1}{m}\sum_{i=1}^{m} w_j^{(i)} \qquad \underline{\boldsymbol{\mu}}_j := \frac{\sum_{i=1}^{m} w_j^{(i)} \underline{\mathbf{x}}^{(i)}}{\sum_{i=1}^{m} w_j^{(i)}}$$

$$\boldsymbol{\Sigma}_j := \frac{\sum_{i=1}^{m} w_j^{(i)} (\underline{\mathbf{x}}^{(i)} - \underline{\boldsymbol{\mu}}_j)(\underline{\mathbf{x}}^{(i)} - \underline{\boldsymbol{\mu}}_j)^T}{\sum_{i=1}^{m} w_j^{(i)}}$$

until *convergencia*

Probablidad a posteriori de $z^{(i)}$

- En el paso E calculamos la probabilidad a posteriori de $z^{(i)}$, dados $\underline{\mathbf{x}}^{(i)}$ y la configuración actual de parámetros $\underline{\boldsymbol{\mu}}$ y $\boldsymbol{\Sigma}$:

$$p(z^{(i)}\!=\!j|\underline{\mathbf{x}}^{(i)};\underline{\phi},\underline{\boldsymbol{\mu}},\boldsymbol{\Sigma}) = \frac{p(\underline{\mathbf{x}}^{(i)}|z^{(i)}\!=\!j;\underline{\boldsymbol{\mu}},\boldsymbol{\Sigma})p(z^{(i)}\!=\!j;\underline{\phi})}{\sum_{l=1}^{k} p(\underline{\mathbf{x}}^{(i)}|z^{(i)}\!=\!l;\underline{\boldsymbol{\mu}},\boldsymbol{\Sigma})p(z^{(i)}\!=\!l;\underline{\phi})}$$

donde

$$p(\underline{\mathbf{x}}^{(i)}|z^{(i)}\!=\!j;\underline{\boldsymbol{\mu}},\boldsymbol{\Sigma}) = G(\underline{\mathbf{x}}^{(i)};\underline{\boldsymbol{\mu}}_j,\boldsymbol{\Sigma}_j)$$
$$p(z^{(i)}\!=\!j;\underline{\phi}) = \phi_j$$

- Los pesos $w_j^{(i)}$ calculados en el paso E son las estimaciones "suaves" (inciertas) de los valores $z^{(i)}$
- Note que $w_j^{(i)}$ reemplaza a $1\left\{z^{(i)} = j\right\}$ en el caso hipotético

Motivación para convergencia de EM

- El algoritmo EM es muy sensible a la inicialización
- Ahora correponde extender el estudio del enfoque EM más allá de la mezcla de gaussianas y analizar su convergencia.
- Iniciaremos revisando la llamada desigualdad de Jensen

Funciones convexas

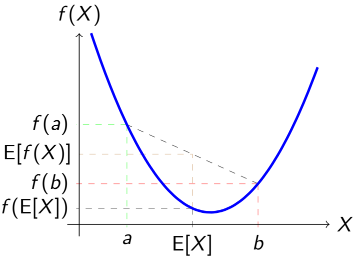
- Sea f una función con dominio real $\mathbb{R}$ o $\mathbb{R}^n$
- $f : \mathbb{R} \to \mathbb{R}$ es convexa si $f''(x) \geq 0$, para todo $x \in \mathbb{R}$
- $f : \mathbb{R}^n \to \mathbb{R}$ es convexa si su matriz hessiana $\mathbf{H}$ es positiva semi-definida.
- f es **estríctamente** convexa si $f''(x) > 0$ o si $\mathbf{H}$ es positiva definida.

Desigualdad de Jensen

- Sea f una función convexa, y X una variable aleatoria:

$$\mathrm{E}[f(X)] \geq f(\mathrm{E}[X])$$

- Si f es **estrictamente** convexa, entonces $\mathrm{E}[f(X)] = f(\mathrm{E}[X])$ solo si $X = \mathrm{E}[X]$ (esto es, si X es constante).

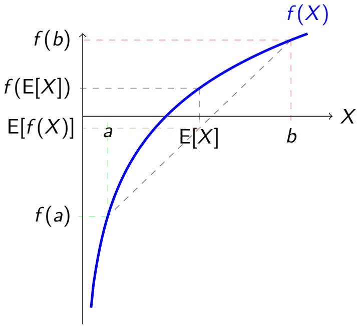


- En ejemplo:
  - Supongamos $p(X = a) = p(X = b) = 1/2$
  - $\mathrm{E}[X] = (a + b)/2$
  - Por ser f convexa se cumple $\mathrm{E}[f(X)] \geq f(\mathrm{E}[X])$
  - Igualdad en $\geq$: si $X = \mathrm{E}[X]$

Concavidad

- Una función f es (estrictamente) cóncava si $-f$ es (estrictamente) convexa.
- La función es cóncava entonces si $f''(x) \leq 0$ ó $\mathbf{H} \preceq 0$
- La desigualdad de Jensen establece en este caso

$$\mathrm{E}[f(X)] \leq f(\mathrm{E}[X])$$



- El logaritmo es una función cóncava
- En ejemplo:
  - Supongamos $p(X = a) = p(X = b) = 1/2$
  - $\mathrm{E}[X] = (a + b)/2$
  - Por ser f concava se cumple $\mathrm{E}[f(X)] \leq f(\mathrm{E}[X])$

Generalizando en algoritmo EM

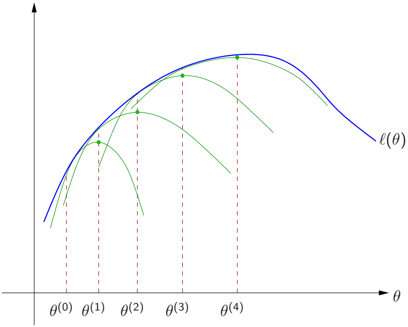
- Partamos de un problema de estimación con datos independientes $\{\underline{\mathbf{x}}^{(1)}, \ldots, \underline{\mathbf{x}}^{(m)}\}$
- Tenemos un modelo para la distribución conjunta $p(\underline{\mathbf{x}}, z; \underline{\boldsymbol{\theta}})$
- Solo podemos observar $\underline{\mathbf{x}}$ (z es latente/escondida/no observable)
- Queremos ajustar los parámetros del modelo $p(\underline{\mathbf{x}}, z; \underline{\boldsymbol{\theta}})$ a los datos, maximizando la verosimilitud logarítmica

$$\ell(\underline{\boldsymbol{\theta}}) = \sum_{i=1}^{m} \ln p(\underline{\mathbf{x}}^{(i)}; \underline{\boldsymbol{\theta}}) = \sum_{i=1}^{m} \ln \sum_{z^{(i)}} p(\underline{\mathbf{x}}^{(i)}, z^{(i)}; \underline{\boldsymbol{\theta}})$$

- Encontrar los parámetros $\underline{\boldsymbol{\theta}}$ estimados con máxima verosimilitud con los z desconocidos es tarea difícil
- Si se pudieran observar los $z^{(i)}$ estimación sería sencilla

Comprendiendo el algoritmo EM

- La estrategia de EM es entonces encontrar repetidamente una cota inferior de $\ell(\underline{\boldsymbol{\theta}})$ (Paso E), y luego optimizar esa cota (Paso M)



Generalizando el algoritmo EM (1)

- Para cada i, sea Q_i una distribución discreta sobre z

$$\sum_{z^{(i)}} Q_i(z^{(i)}) = 1, \qquad Q_i(z^{(i)}) \geq 0$$

- (Si Q_i fuese densidad probabilistica sumas $\rightarrow$ integrales)
- Se cumple entonces que

$$\begin{aligned}
\sum_i \ln p(\underline{\mathbf{x}}^{(i)}; \underline{\boldsymbol{\theta}}) &= \sum_i \ln \sum_{z^{(i)}} p(\underline{\mathbf{x}}^{(i)}, z^{(i)}; \underline{\boldsymbol{\theta}}) \\
&= \sum_i \ln \sum_{z^{(i)}} Q_i(z^{(i)}) \frac{p(\underline{\mathbf{x}}^{(i)}, z^{(i)}; \underline{\boldsymbol{\theta}})}{Q_i(z^{(i)})}
\end{aligned}$$

Generalizando el algoritmo EM (2)

- Nótese que

$$\sum_{z^{(i)}} Q_i(z^{(i)}) \frac{p(\underline{\mathbf{x}}^{(i)}, z^{(i)}; \underline{\boldsymbol{\theta}})}{Q_i(z^{(i)})}$$

es la esperanza de $p(\underline{\mathbf{x}}^{(i)}, z^{(i)}; \underline{\boldsymbol{\theta}})/Q_i(z^{(i)})$, es decir

$$\sum_i \ln p(\underline{\mathbf{x}}^{(i)}; \underline{\boldsymbol{\theta}}) = \sum_i \ln \mathrm{E}_{z^{(i)} \sim Q_i} \left[\frac{p(\underline{\mathbf{x}}^{(i)}, z^{(i)}; \underline{\boldsymbol{\theta}})}{Q_i(z^{(i)})} \right]$$

- Considerando que $\ln(x)$ es cóncava y la desigualdad de Jensen $\ln(\mathrm{E}[X]) \geq \mathrm{E}[\ln(X)]$ entonces

$$\sum_i \ln p(\underline{\mathbf{x}}^{(i)}; \underline{\boldsymbol{\theta}}) \geq \sum_i \mathrm{E}_{z^{(i)} \sim Q_i} \left[\ln \left(\frac{p(\underline{\mathbf{x}}^{(i)}, z^{(i)}; \underline{\boldsymbol{\theta}})}{Q_i(z^{(i)})} \right) \right]$$

Generalizando el algoritmo EM (3)

- Para **cualquier** conjunto de distribuciones Q_i la ecuación anterior da una cota inferior de $\ell(\underline{\boldsymbol{\theta}})$

$$\ell(\underline{\boldsymbol{\theta}}) \geq \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \ln \frac{p(\underline{\mathbf{x}}^{(i)}, z^{(i)}; \underline{\boldsymbol{\theta}})}{Q_i(z^{(i)})}$$

- Lo que haremos es elegir una distribución $Q_i(z^{(i)})$ que fuerce la igualdad de la cota inferior para poder así alcanzar a $\ell(\underline{\boldsymbol{\theta}})$

Forzando la cota inferior estricta

- Queremos que en $\underline{\boldsymbol{\theta}}$ la cota inferior alcance a $\ell(\underline{\boldsymbol{\theta}})$
- Vimos que $\mathrm{E}[f(X)] = f(E[X])$ si $X = E[X]$
- Entonces, requerimos que

$$\frac{p(\underline{\mathbf{x}}^{(i)}, z^{(i)}; \underline{\boldsymbol{\theta}})}{Q_i(z^{(i)})} = c$$

  con c constante respecto a $z^{(i)}$
- Eso lo logramos haciendo

$$Q_i(z^{(i)}) \propto p(\underline{\mathbf{x}}^{(i)}, z^{(i)}; \underline{\boldsymbol{\theta}})$$

Encontrando la distribución $Q_i(z^{(i)})$

- Puesto que $\sum_{z^{(i)}} Q_i(z^{(i)}) = 1$ entonces

$$Q_i(z^{(i)}) = \frac{p(\underline{\mathbf{x}}^{(i)}, z^{(i)}; \underline{\boldsymbol{\theta}})}{\sum_{z^{(i)}} p(\underline{\mathbf{x}}^{(i)}, z^{(i)}; \underline{\boldsymbol{\theta}})} = \frac{p(\underline{\mathbf{x}}^{(i)}, z^{(i)}; \underline{\boldsymbol{\theta}})}{p(\underline{\mathbf{x}}^{(i)}; \underline{\boldsymbol{\theta}})} = p(z^{(i)} | \underline{\mathbf{x}}^{(i)}; \underline{\boldsymbol{\theta}})$$

  así, símplemente usamos la distribución a posteriori de $z^{(i)}$ dado $\underline{\mathbf{x}}^{(i)}$ para Q_i, considerando la configuración $\underline{\boldsymbol{\theta}}$
- Con este Q_i tenemos la cota inferior estricta de $\ell(\underline{\boldsymbol{\theta}})$. Esto es el **paso E**
- En el paso M maximizamos la cota con respecto a $\underline{\boldsymbol{\theta}}$.

Algoritmo EM generalizado

repeat

`// Paso E`

foreach i **do**

$$Q_i(z^{(i)}) := p(z^{(i)}|\underline{\mathbf{x}}^{(i)};\underline{\boldsymbol{\theta}})$$

end

`// Paso M`

Actualice los parámetros:

$$\underline{\boldsymbol{\theta}} := \arg\max_{\underline{\boldsymbol{\theta}}} \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \ln \frac{p(\underline{\mathbf{x}}^{(i)}, z^{(i)}; \underline{\boldsymbol{\theta}})}{Q_i(z^{(i)})}$$

until *convergencia*

Convergencia del algoritmo EM

- Se puede demostrar que este algoritmo incrementa monotónicamente $\ell(\underline{\boldsymbol{\theta}})$
- Para detectar convergencia basta entonces con detectar la tasa de cambio de $\ell(\underline{\boldsymbol{\theta}})$ y detener el algoritmo si dicha tasa es suficientemente pequeña
- Se puede además demostrar que EM realiza un ascenso de coordenadas maximizando respecto a Q primero (paso E) y luego respecto a $\underline{\boldsymbol{\theta}}$ (paso M)

Retomando el EM con mezcla de gaussianos

- Si aplicamos lo anterior a la mezcla de gaussianas llegamos a los resultados ya presentados (ver notas de Andrew Ng `cs229-notes8.pdf`).
- Partimos símplemente para el paso E que

$$w_j^{(i)} = Q_i(z^{(i)} = j) = p(z^{(i)} = j|\underline{\mathbf{x}}^{(i)}; \underline{\phi}, \underline{\boldsymbol{\mu}}, \boldsymbol{\Sigma})$$

- Luego para el paso M usaríamos el hecho de que $p(\underline{\mathbf{x}}^{(i)}|z^{(i)} = j; \underline{\boldsymbol{\mu}}, \boldsymbol{\Sigma})$ son gaussianas.

Ejemplo con modelos bayesianos ingenuos

Aglomeración de textos

- Supongamos que queremos hacer aglomeración de textos
- La idea es agrupar textos similares
- Aplicaremos el algoritmo EM con entradas discretas
- Mezcla de modelos bayesianos ingenuos (*mixture of naïve Bayes models*)
- Cuando hablamos de modelos bayesianos ingenuos vimos dos modelos:
  1. Modelo de eventos multivariados de Bernoulli
  2. Modelo de eventos multinomial
- Usaremos ahora el modelo multivariado

Mezcla de modelos bayesianos ingenuos

- Dado el conjunto de entrada $\{\underline{\mathbf{x}}^{(1)}, \ldots, \underline{\mathbf{x}}^{(m)}\}$, con $\underline{\mathbf{x}}^{(i)} \in \{0,1\}^n$ un documento de texto
- Cada componente $x_j^{(i)}$ denota la presencia de la j-ésima palabra en el documento
- Supongamos $z^{(i)} \in \{0,1\}$ (2 conglomerados)
- En la mezcla de modelos bayesianos ingenuos suponemos que:
  - $z^{(i)} \sim \mathrm{Ber}(\phi)$
  - $p(\underline{\mathbf{x}}^{(i)}|z^{(i)}) = \prod_{j=1}^{n} p(x_j^{(i)}|z^{(i)})$ específicamente $p(x_j^{(i)} = 1|z^{(i)} = 0) = \phi_{j|z=0}$
  - $w^{(i)} = p(z^{(i)} = 1|\underline{\mathbf{x}}^{(i)}) = \dfrac{p(\underline{\mathbf{x}}^{(i)}|z^{(i)} = 1)p(z^{(i)} = 1)}{\sum_{l=0}^{1} p(\underline{\mathbf{x}}^{(i)}|z^{(i)} = l)p(z^{(i)} = l)}$
- Realizando todas las manipulaciones algebraicas con el algoritmo EM obtenemos el siguiente algoritmo

Algoritmo EM para mezcla de modelos bayesianos ingenuos

repeat

// Paso E

foreach i **do**

$$w^{(i)} := p(z^{(i)} = 1 | \underline{\mathbf{x}}^{(i)}; \phi_{j|z}, \phi_z)$$

end

// Paso M

Actualice los parámetros

$$\phi_{j|z=1} := \frac{\sum_{i=1}^{m} w^{(i)} 1\left\{x_j^{(i)} = 1\right\}}{\sum_{i=1}^{m} w^{(i)}} \qquad \phi_z = \frac{\sum_{i=1}^{m} w^{(i)}}{m}$$

$$\phi_{j|z=0} := \frac{\sum_{i=1}^{m} (1 - w^{(i)}) 1\left\{x_j^{(i)} = 1\right\}}{\sum_{i=1}^{m} (1 - w^{(i)})}$$

until *convergencia*

Resumen

Este documento ha sido elaborado con software libre incluyendo LaTeX, Beamer, GNUPlot, GNU/Octave, XFig, Inkscape, GNU-Make y Subversion en GNU/Linux