20.1

1. Descripción del problema

Se tiene un modelo probabilístico de mezcla de modelos factorizados para observaciones vectoriales v, con la siguiente probabilidad conjunta: $p(v) = \sum_{k} p(k) \prod_{i} p(v_i | k)$ donde v_i representa el i-ésimo componente de la observación v, y k es una variable oculta que determina la mezcla.

2. Verosimilitud de datos observados

La verosimilitud para el conjunto de datos es: $\mathcal{L}(\theta) = \prod_{n=1}^{N} p(v_n | \theta)$ donde θ son los parámetros del modelo. Cada v_n es un vector, pero algunos de sus componentes pueden estar faltantes.

Suponiendo que para la n-ésima observación, el conjunto de componentes conocidos es K_n, y el conjunto de componentes faltantes es M_n. Entonces la verosimilitud de los datos observados con componentes faltantes es: $p(v_n^{obs} | \theta) = \sum_{k} p(k|\theta) \prod_{i \in K_n} p(v_n^i | k, \theta)$

3. Máxima verosimilitud con datos faltantes

Consiste en maximizar la verosimilitud total de los datos observados. Sustituyendo la expresión de la verosimilitud: $\mathcal{L}(\theta) = \prod_{n=1}^{N} \sum_{k} p(k|\theta) \prod_{i \in K_n} p(v_n^i | k, \theta)$ de modo que los componentes faltantes M_n no aparecen.

4. Conclusión

Dado que la verosimilitud solo se calcula sobre los componentes observados de cada observación, al realizar el entrenamiento por máxima verosimilitud en los datos observados, se ignoran los componentes faltantes.

20.3

1. MEZCLA DE GAUSSIANAS ISÓTROPAS

Se considera un modelo de mezcla K gaussianas isótropas, donde cada componente tiene una media μ_k y una covarianza $\Sigma_k = 2I$.
La probabilidad condicional para una observación v dada la clase k:
$p(v|k) = \frac{1}{(2\pi\sigma^2)^{D/2}} \exp\left(-\frac{1}{2\sigma^2}\|v-\mu_k\|^2\right)$ donde σ^2 es el valor escalar de la varianza y D es la dimensionalidad del espacio.

2. ALGORITMO EM

1. Paso E: Calcula la probabilidad posterior de que un punto v_n pertenezca al clúster k. Esta probabilidad es la responsabilidad r_{nk}:

$$r_{nk} = \frac{p(k|v_n)}{\sum_{j=1}^{K} p(j|v_n)} = \frac{p(v_n|k)p(k)}{\sum_{j=1}^{K} p(v_n|j)p(j)}$$

, como todas las covarianzas son iguales y las probabilidades a priori p(k) también:

$$r_{nk} = \frac{\exp\left(-\frac{1}{2\sigma^2}\|v_n-\mu_k\|^2\right)}{\sum_{j=1}^{K} \exp\left(-\frac{1}{2\sigma^2}\|v_n-\mu_j\|^2\right)}$$

2. Paso M: Actualiza las medias μ_k en función de las responsabilidades r_{nk}:

$$\mu_k = \frac{\sum_{n=1}^{N} r_{nk} v_n}{\sum_{n=1}^{N} r_{nk}}$$

3. LÍMITE CUANDO $2 \to 0$

Cuando $\sigma^2 \to 0$, las distribuciones gaussianas se concentran alrededor de sus medias μ_k. De modo que cada punto de datos v_n se asigna por completo al clúster cuya media esté más cercana, lo que equivale a la etapa de asignación de clústers en el algoritmo K-means.

4. CONCLUSIÓN

En resumen, cuando la covarianza de las gaussianas isótropas tiende a cero, el algoritmo EM asigna de manera determinista cada punto al clúster más cercano, y las actualizaciones de las medias se hacen de la misma forma que en K-means. Por tanto, en este límite, el algoritmo EM tiende al algoritmo de K-means.

T