Modelos estadísticos con métodos variacionales

Doble Grado en Ingeniería Informática y Matemáticas

Luis Antonio Ortega Andrés

15 de septiembre de 2020

Trabajo Fin de Grado

E.T.S. de Ingenierías Informática y de Telecomunicación
Facultad de Ciencias

UNIVERSIDAD
DE GRANADA

Índice

Introducción

Los métodos variacionales permiten la resolución de problemas de inferencia con variables ocultas, tanto con enfoque Bayesiano como basado en verosimilitud. Han aumentado notablemente los modelos estadísticos en los que se puede hacer inferencia.

Las variables ocultas pueden ser parámetros bajo un modelo Bayesiano.

Modelos concretos como la mixtura de Gaussianas, asignación latente de Dirichlet y análisis de componentes principales pueden ser estudiados mediante la utilización de lenguajes de programación especializados.

Divergencia Kullback-Leibler

Sean P y Q dos distribuciones de probabilidad sobre el mismo espacio probabilístico, su *divergencia de Kullback-Leibler* $KL\Big(Q \mid P\Big)$ mide la "diferencia" de Q a P

$$KL\Big(Q \mid P\Big) = \mathbb{E}_Q\Big[\log Q(x) - \log P(x)\Big].$$

La divergencia de Kullback-Leibler es siempre no negativa.

El problema

- Un conjunto de variables observadas i.i.d $\boldsymbol{X} = (X_1, \ldots, X_N)$.
- Variables ocultas globales $\boldsymbol{\theta}$ y variables ocultas locales $\boldsymbol{Z} = (Z_1, \ldots, Z_N)$.
- La distribución conjunta factoriza como

$$P(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\theta}) = P(\boldsymbol{\theta}) \prod_{n=1}^{N} P(x_n, z_n \mid \boldsymbol{\theta}).$$

Las variables ocultas pueden ser tratadas como variables o como parámetros.

Objetivo: calcular

$$P(\boldsymbol{\theta}, \boldsymbol{z} \mid \boldsymbol{x}) = \frac{P(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{x})}{\int_{\boldsymbol{\theta}, \boldsymbol{z}} P(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{x})}.$$

Inferencia variacional

Inferencia variacional

Resuelve el problema de inferencia mediante uno de *optimización*:

Dada una familia de distribuciones $\mathcal{Q}$ sobre el conjunto de variables ocultas $\boldsymbol{Z}$, encontrar

$$Q^{opt} = \arg\min_{Q \in \mathcal{Q}} KL\Big(Q(\boldsymbol{z}) \;\Big|\; P(\boldsymbol{z} \mid \boldsymbol{x})\Big).$$

El problema se afronta mediante técnicas de aprendizaje automático como *gradiente descendente* o *descenso coordinado*.

Cota inferior para la evidencia

Utilizando la positividad de la divergencia de Kullback-Leibler, se obtiene una cota inferior para la evidencia:

$$KL\Big(Q(\boldsymbol{z}) \mid P(\boldsymbol{z} \mid \boldsymbol{x})\Big) \geq 0$$

$$\Updownarrow$$

$$\log P(\boldsymbol{x}) \geq \underbrace{-\mathbb{E}_{Q(\boldsymbol{z})}\Big[\log Q(\boldsymbol{z})\Big]}_{\text{Entropía}} + \underbrace{\mathbb{E}_{Q(\boldsymbol{z})}\Big[\log P(\boldsymbol{x}, \boldsymbol{z})\Big]}_{\text{Energía}} = ELBO(Q).$$

Algoritmo Esperanza-Maximización

No considera distribución sobre los parámetros.

$$ELBO(Q, \boldsymbol{\theta}) = \underbrace{-\mathbb{E}_{Q(\boldsymbol{z})}\Big[\log Q(\boldsymbol{z})\Big]}_{\text{Entropía}} + \underbrace{\mathbb{E}_{Q(\boldsymbol{z})}\Big[\log P(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{\theta})\Big]}_{\text{Energía}}.$$

Dos pasos iterativos:

- **Paso E**: Fijado el parámetro $\boldsymbol{\theta}$,

$$Q^{new}(\boldsymbol{z}) = \arg\max_{Q} ELBO(Q, \boldsymbol{\theta}) = P(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{\theta}).$$

- **Paso M**: Fijada la distribución Q,

$$\boldsymbol{\theta}^{new} = \arg\max_{\boldsymbol{\theta}} ELBO(Q, \boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \mathbb{E}_{Q(\boldsymbol{z})}\Big[\log P(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{\theta})\Big].$$

El algoritmo EM incrementa la verosimilitud.

$$\log P(\boldsymbol{x} \mid \boldsymbol{\theta}) \geq -\mathbb{E}_{Q(\boldsymbol{z})}\Big[\log Q(\boldsymbol{z})\Big] + \mathbb{E}_{Q(\boldsymbol{z})}\Big[\log P(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{\theta})\Big].$$

$$\Downarrow \text{ Paso E}$$

$$\log P(\boldsymbol{x} \mid \boldsymbol{\theta}) = -\mathbb{E}_{P(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{\theta})}\Big[\log P(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{\theta})\Big] + \mathbb{E}_{P(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{\theta})}\Big[\log P(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{\theta})\Big].$$

$$\Downarrow \text{ Paso M}$$

$$\log P(\boldsymbol{x} \mid \boldsymbol{\theta}^{new}) \geq -\mathbb{E}_{P(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{\theta})}\Big[\log P(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{\theta})\Big] + \mathbb{E}_{P(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{\theta})}\Big[\log P(\boldsymbol{x}, \boldsymbol{z} \mid \boldsymbol{\theta}^{new})\Big].$$

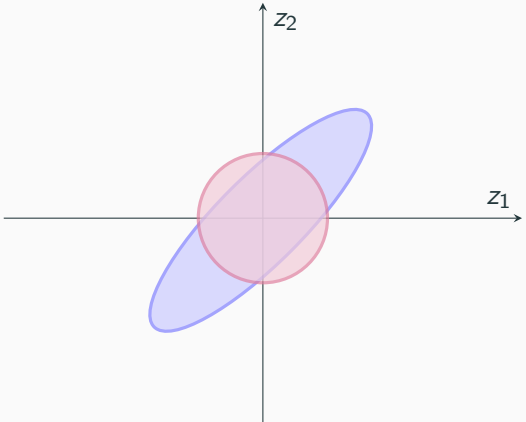
Converge a un máximo local.

Algoritmo de ascenso coordinado en inferencia variacional

La familia de distribuciones de *campo medio* factoriza como producto de marginales:

$$Q(\boldsymbol{z}) = \prod_{m=1}^{M} Q(z_m).$$

- Pueden capturar cualquier distribución marginal.
- No pueden capturar correlación entre variables.



El algoritmo de ascenso coordinado se basa en:

- Los parámetros se consideran variables ocultas $\boldsymbol{\theta} \subset \boldsymbol{Z}$.
- La familia de distribuciones $\mathcal{Q}$ del problema de optimización es la familia de campo medio.
- La distribución marginal de cada variable se actualiza cada vez.

$$Q^{new}(z_m) = \arg\max_{Q \in \mathcal{Q}} ELBO(Q)$$

$$\Downarrow$$

$$Q^{new}(z_m) \propto \exp \mathbb{E}_{Q_{\setminus m}} \Big[\log P(z_m \mid \boldsymbol{z}_{\setminus m}, \boldsymbol{x}) \Big] \propto \exp \mathbb{E}_{Q_{\setminus m}} \Big[\log P(\boldsymbol{z}, \boldsymbol{x}) \Big]$$

Familia exponencial

Una variable aleatoria X sigue una distribución en la *familia exponencial* con parámetros $\boldsymbol{\theta}$ si y solo si existen h, $\boldsymbol{T}, \boldsymbol{\eta}$ y ψ tales que

$$P(x \mid \boldsymbol{\theta}) = h(x) \exp\Big(\boldsymbol{\eta}(\boldsymbol{\theta})^T \boldsymbol{T}(x) - \psi(\boldsymbol{\theta})\Big).$$

- $\boldsymbol{T}$ es el estadístico natural de X. *Teorema de Fisher-Neyman*.
- $\boldsymbol{\eta}$ se denomina la función paramétrica de la distribución.
- h se denomina medida base.
- ψ asegura normalización logarítmica

$$\psi(\boldsymbol{\theta}) = \log \int_x h(x) \exp\Big(\boldsymbol{\eta}(\boldsymbol{\theta})^T \boldsymbol{T}(x)\Big).$$

Modelos condicionalmente conjugados

Se consideran distribuciones conjugadas en la familia exponencial.

Las actualizaciones de las distribuciones variacionales $Q^{new}(z_n)$ y $Q^{new}(\boldsymbol{\theta})$ consisten en actualizar la *función paramétrica de la distribución* $\boldsymbol{\eta}$.

Estas actualizaciones puede calcularse de forma eficiente, calculando esperanzas de los estadísticos suficientes.

Redes Bayesianas

Red Bayesiana

Una *red de creencia o red Bayesiana* es una pareja (G, P) formada por un grafo dirigido acíclico G y una distribución de probabilidad P tal que existe una correspondencia entre variables y nodos verificando:

$$P(x_1, \ldots, x_N) = \prod_{n=1}^{N} P(x_n \mid pa(x_n)).$$

$$P(x_1, x_2, x_3) = P(x_1 \mid x_3)P(x_2 \mid x_3)P(x_3).$$

Algoritmo de paso de mensajes

Proceso de paso de mensajes entre los nodos de la red.

- Familia variacional de campo medio.

$$Q(\boldsymbol{z}) = \prod_{m=1}^{M} Q(z_m).$$

- Modelo condicionalmente conjugado en la familia exponencial.
- Ascenso coordinado en redes Bayesianas.

La actualización de la distribución dada la red es:

$$\log Q^{new}(z) = \mathbb{E}_{Q_{\setminus Z}}\Big[\sum_{n=1}^{N} \log P(x_n \mid pa(x_n))\Big] + \text{const.}$$

La aportación de Z:

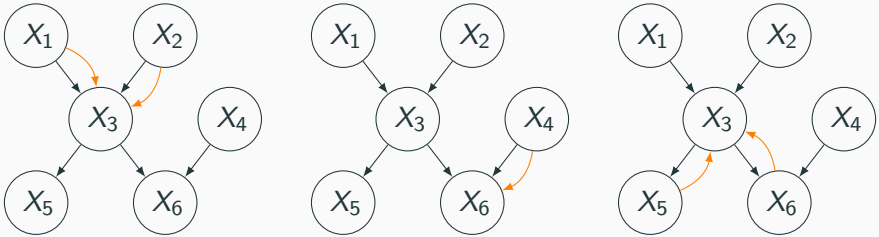
$$\log Q^{new}(z) = \mathbb{E}_{Q_{\setminus Z}}\Big[\log P(z \mid pa(z))\Big] + \sum_{X \in ch(Z)} \mathbb{E}_{Q_{\setminus Z}}\Big[\log P(x \mid z, cp(z,x))\Big] + \text{const.}$$

El mensaje de un nodo padre X a uno hijo Z:

$$\boldsymbol{m}_{X\to Z} = \mathbb{E}_Q\Big[\boldsymbol{T}_X(x)\Big].$$

El mensaje de un nodo hijo X a un nodo padre Z:

$$\boldsymbol{m}_{X\to Z} = \bar{\eta}_{X,Z}\Big(\mathbb{E}_Q\Big[\boldsymbol{T}_X(x)\Big], \{\boldsymbol{m}_{Y\to X}\}_{Y\in cp_{Z,X}}\Big),$$

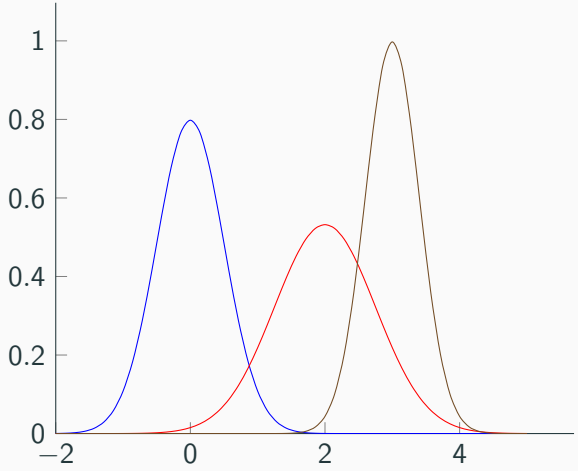


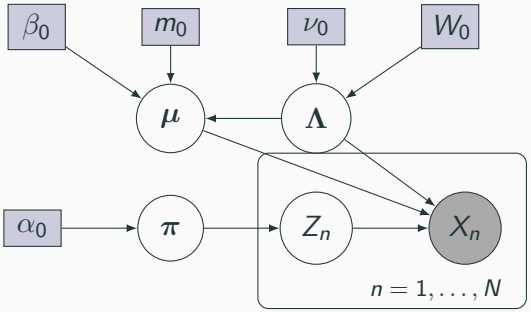
Modelos

Mixtura de Gaussianas

Mixtura de Gaussianas $\mathcal{N}(0, 0.5)$, $\mathcal{N}(2, 0.75)$ y $\mathcal{N}(3, 0.4)$.
Dados unos pesos $0.2, 0.3$ y 0.5, la probabilidad de un punto sería

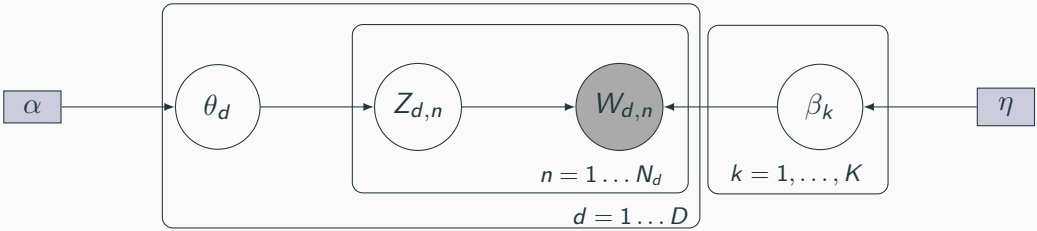
$$\begin{aligned} P(x) = & 0.2 \times \mathcal{N}(0, 0.5)(x) \\ & + 0.3 \times \mathcal{N}(2, 0.75)(x) \\ & + 0.5 \times \mathcal{N}(3, 0.4)(x). \end{aligned}$$





- **Pesos**: $\boldsymbol{\pi} \sim \text{Dirichlet}(\alpha_0)$.
- **Medias y precisiones:** $(\boldsymbol{\mu}, \boldsymbol{\Lambda}) \sim \text{Gaussian-Wishart}(\beta_0, m_0, \nu_0, W_0)$.
- **Componentes**: $Z_n \mid \boldsymbol{\pi} \sim \textit{Categorica}(\boldsymbol{\pi})$.
- **Observaciones**: $X_n \mid z_n, \boldsymbol{\mu}, \boldsymbol{\Lambda} \sim \mathcal{N}(\mu_{z_n}, \Lambda_{z_n})$.

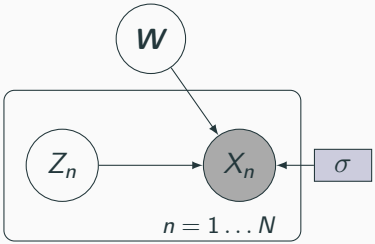
Asignación latente de Dirichlet



- **Palabras por tema:** $\beta_k \sim \text{Dirichlet}(\eta)$.
- **Temas por documento:** $\theta_d \sim \text{Dirichlet}(\alpha)$.
- **Temas por documento y palabra:** $Z_{d,n} \mid \theta_d \sim \text{Categorica}(\theta_d)$.
- **Palabras:** $W_{d,n} \mid z_{d,n}, \boldsymbol{\beta} \sim \text{Categorica}(\beta_{z_{d,n}})$.

Reducción de dimensionalidad

Reducción de espacio D dimensional a uno K dimensional.



- **Transformación lineal:** $\boldsymbol{W} \sim \mathcal{N}_{D \times K}(0, I)$.
- **Representación oculta:** $Z_n \sim \mathcal{N}_K(0, I)$.
- **Representación observada:** $X_n \mid \boldsymbol{W}, z_n \sim \mathcal{N}_d(\boldsymbol{W} z_n, \sigma I)$.

Modelo no lineal: Red neuronal totalmente conectada de 2 capas.

$$X_n \mid z_n \sim \mathcal{N}_D(f(z_n), \sigma I).$$

Codificadores variacionales automáticos

Modelo paramétrico

Representación oculta

$$Z_n \sim \mathcal{N}_K(0, I).$$

Representación observada

$$X_n \mid z_n \sim \mathcal{N}_D(f(z_n), \sigma I).$$

Modelo variacional

Representación observada:

$$X_n \sim \mathcal{N}_D(0, I)$$

Representación oculta

$$Z_n \mid x_n \sim \mathcal{N}_K(\mu, \sigma I) \quad (\mu, \sigma) = g(x_n).$$

Caso práctico

Frameworks utilizados

`InferPy`, `BayesPy` y `BayesianGaussianMixture` de `Scikit-Learn`.

Promovieron la utilización de métodos variacionales.

Especificación de modelos con alta abstracción.

PCA no lineal con InferPy:

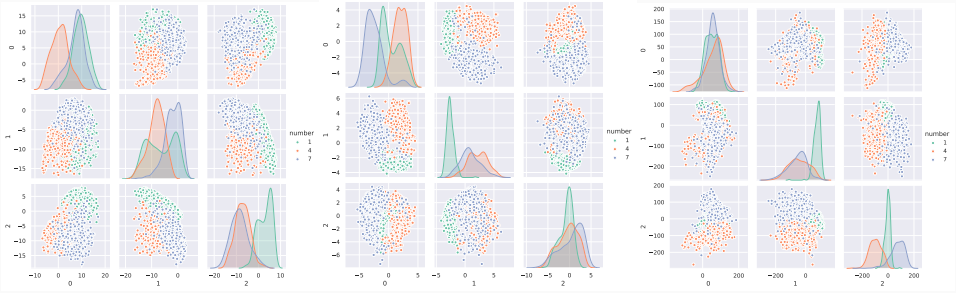
```
nn = decoder(k, l, d)
with inf.datamodel():
    z = inf.Normal(loc=tf.zeros([k]), scale=1, name="z")
    output = nn(z)
    x = inf.Normal(loc=output, scale=1.0, name="x")
```

InferPy

Integración con Keras.

No puede aprender variables categóricas.

Algoritmo gradiente descendente.

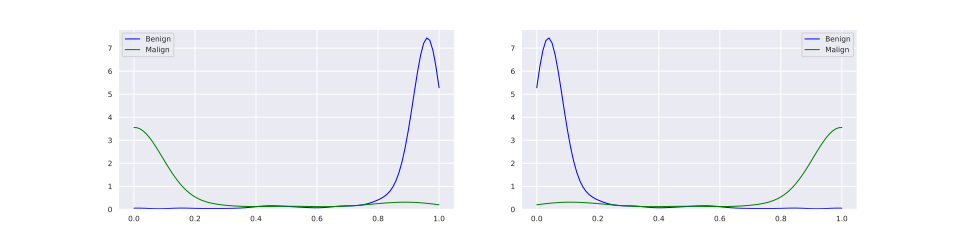


BayesPy

Flexibilidad en diseño de modelos.

Altos requisitos de memoria.

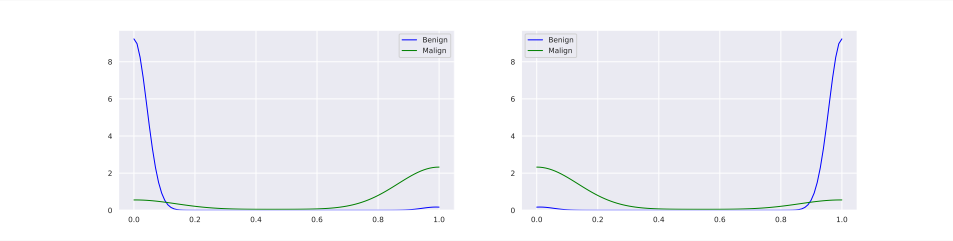
Algoritmo de paso de mensajes variacional.



Scikit-Learn

Modelo de mixtura pre-definido.

Algoritmo EM.



Conclusiones

- La utilización de modelos gráficos como redes Bayesianas y modelos conjugados en la familia exponencial simplifican la inferencia con variables ocultas hasta su automatización.
- La integración de redes neuronales, permite aplicar inferencia a modelos fuera de los modelos conjugados.
- Cada software utilizado presenta ventajas e inconvenientes sobre los demás, haciendo que su elección dependa de la tarea que se desee realizar.

Gracias por su atención