Modelos Generativos Profundos para Imágenes

Autoencoders variacionales

Pablo Musé
pmuse@fing.edu.uy

Instituto de Ingeniería Eléctrica
Facultad de Ingeniería

UNIVERSIDAD
DE LA REPÚBLICA
URUGUAY

Agenda

Resumen de la clase anterior

Normalizing Flows:

- Transforman distribuciones simples en complejas mediante **cambio de variables**.

Resumen de la clase anterior

Normalizing Flows:

- Transforman distribuciones simples en complejas mediante **cambio de variables**.
- **Pros:**
  - La verosimilitud marginal exacta $p(\mathbf{x})$ es fácil de calcular y optimizar.
  - La inferencia posterior exacta $p(\mathbf{z}|\mathbf{x})$ es tratable.

Resumen de la clase anterior

Normalizing Flows:

- Transforman distribuciones simples en complejas mediante **cambio de variables**.
- **Pros:**
  - La verosimilitud marginal exacta $p(\mathbf{x})$ es fácil de calcular y optimizar.
  - La inferencia posterior exacta $p(\mathbf{z}|\mathbf{x})$ es tratable.
- **Contras:**
  - La dimensión de $\mathbf{z}$ y $\mathbf{x}$ **debe ser la misma** (puede plantear desafíos computacionales).
  - Impone restricciones importantes sobre qué **familia de modelos** podemos usar.

Modelos en variables latentes: motivación

Las caras (representadas por una imagen $\mathbf{x}$) presentan una gran variabilidad debido a género, edad, color de piel, pelo, ojos, condiciones de adquisición (pose, iluminación), etc.

Modelos en variables latentes: motivación

Las caras (representadas por una imagen $\mathbf{x}$) presentan una gran variabilidad debido a género, edad, color de piel, pelo, ojos, condiciones de adquisición (pose, iluminación), etc.

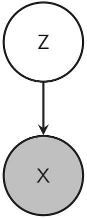
- A menos que las imágenes estén anotadas, estos factores de variación no están disponibles explícitamente (son latentes).

Modelos en variables latentes: motivación

Las caras (representadas por una imagen $\mathbf{x}$) presentan una gran variabilidad debido a género, edad, color de piel, pelo, ojos, condiciones de adquisición (pose, iluminación), etc.

- A menos que las imágenes estén anotadas, estos factores de variación no están disponibles explícitamente (son latentes).
- **Idea:** modelar explícitamente estos factores usando variables latentes z.

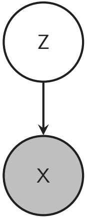
Modelos en variables latentes: definición



Un modelo en variables latentes define una densidad de probabilidad

$$p(\mathbf{x}, \mathbf{z}) = p(\mathbf{x} \mid \mathbf{z})p(\mathbf{z})$$

Modelos en variables latentes: definición

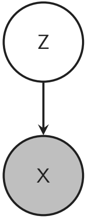


Un modelo en variables latentes define una densidad de probabilidad

$$p(\mathbf{x}, \mathbf{z}) = p(\mathbf{x} \mid \mathbf{z})p(\mathbf{z})$$

- Variables observadas $\mathbf{X}$: los objetos de alta dimensión que queremos modelar, y que están en nuestro conjunto de entrenamiento.

Modelos en variables latentes: definición



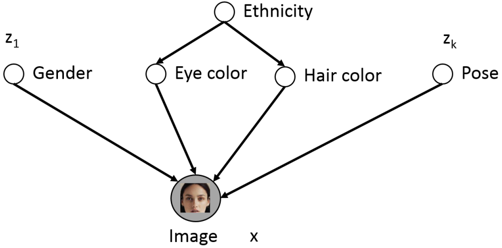
Un modelo en variables latentes define una densidad de probabilidad

$$p(\mathbf{x}, \mathbf{z}) = p(\mathbf{x} \mid \mathbf{z})p(\mathbf{z})$$

- Variables observadas **X**: los objetos de alta dimensión que queremos modelar, y que están en nuestro conjunto de entrenamiento.
- Variables latentes **Z**: no se encuentran en el conjunto de datos, pero están asociadas a **X** según lo especificado por $p(\mathbf{x}, \mathbf{z})$.

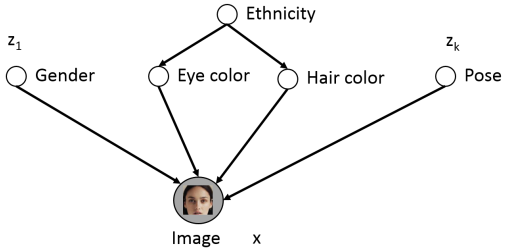
Modelos en variables latentes: ejemplo

Consideremos la siguiente densidad $p(\mathbf{x}, \mathbf{z}) = p(\mathbf{x} \mid \mathbf{z})p(\mathbf{z})$ sobre imágenes de caras.



Modelos en variables latentes: ejemplo

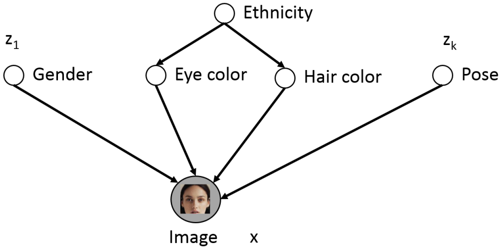
Consideremos la siguiente densidad $p(\mathbf{x}, \mathbf{z}) = p(\mathbf{x} \mid \mathbf{z})p(\mathbf{z})$ sobre imágenes de caras.



- **V.A. observables X:** imágenes de caras. **V.A. latentes Z:** características de alto nivel.

Modelos en variables latentes: ejemplo

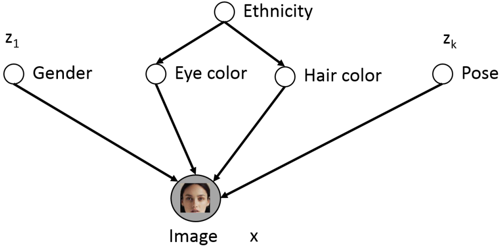
Consideremos la siguiente densidad $p(\mathbf{x}, \mathbf{z}) = p(\mathbf{x} \mid \mathbf{z})p(\mathbf{z})$ sobre imágenes de caras.



- **V.A. observables X:** imágenes de caras. **V.A. latentes Z:** características de alto nivel.
  - Si $\mathbf{z}$ se elige adecuadamente, $p(\mathbf{x} \mid \mathbf{z})$ podría ser mucho más simple que $p(\mathbf{x})$.

Modelos en variables latentes: ejemplo

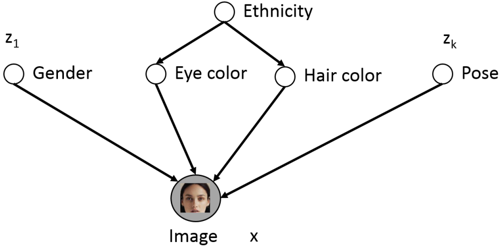
Consideremos la siguiente densidad $p(\mathbf{x}, \mathbf{z}) = p(\mathbf{x} \mid \mathbf{z})p(\mathbf{z})$ sobre imágenes de caras.



- **V.A. observables X:** imágenes de caras. **V.A. latentes Z:** características de alto nivel.
  - Si $\mathbf{z}$ se elige adecuadamente, $p(\mathbf{x} \mid \mathbf{z})$ podría ser mucho más simple que $p(\mathbf{x})$.
  - Con este modelo entrenado, podemos extraer *features* via $p(\mathbf{z} \mid \mathbf{x})$, e.g., $p(\text{Género} = \text{F} \mid \mathbf{x})$.

Modelos en variables latentes: ejemplo

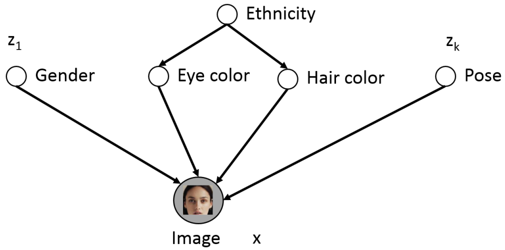
Consideremos la siguiente densidad $p(\mathbf{x}, \mathbf{z}) = p(\mathbf{x} \mid \mathbf{z})p(\mathbf{z})$ sobre imágenes de caras.



- **V.A. observables X:** imágenes de caras. **V.A. latentes Z:** características de alto nivel.
  - Si $\mathbf{z}$ se elige adecuadamente, $p(\mathbf{x} \mid \mathbf{z})$ podría ser mucho más simple que $p(\mathbf{x})$.
  - Con este modelo entrenado, podemos extraer *features* via $p(\mathbf{z} \mid \mathbf{x})$, e.g., $p(\text{Género} = \text{F} \mid \mathbf{x})$.
- **Desafíos:**

Modelos en variables latentes: ejemplo

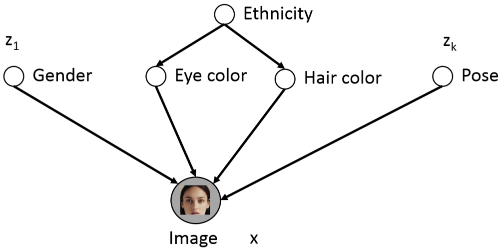
Consideremos la siguiente densidad $p(\mathbf{x}, \mathbf{z}) = p(\mathbf{x} \mid \mathbf{z})p(\mathbf{z})$ sobre imágenes de caras.



- **V.A. observables X:** imágenes de caras. **V.A. latentes Z:** características de alto nivel.
  - Si $\mathbf{z}$ se elige adecuadamente, $p(\mathbf{x} \mid \mathbf{z})$ podría ser mucho más simple que $p(\mathbf{x})$.
  - Con este modelo entrenado, podemos extraer *features* via $p(\mathbf{z} \mid \mathbf{x})$, e.g., $p(\text{Género} = \text{F} \mid \mathbf{x})$.
- **Desafíos:**
  - My difícil de especificar estas condicionales $p(\mathbf{x} \mid \mathbf{z})$ “a mano”

Modelos en variables latentes: ejemplo

Consideremos la siguiente densidad $p(\mathbf{x}, \mathbf{z}) = p(\mathbf{x} \mid \mathbf{z})p(\mathbf{z})$ sobre imágenes de caras.

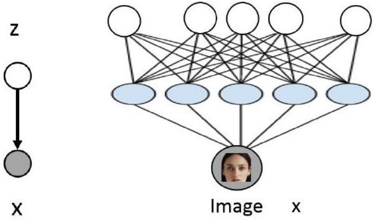


- **V.A. observables X:** imágenes de caras. **V.A. latentes Z:** características de alto nivel.
  - Si $\mathbf{z}$ se elige adecuadamente, $p(\mathbf{x} \mid \mathbf{z})$ podría ser mucho más simple que $p(\mathbf{x})$.
  - Con este modelo entrenado, podemos extraer *features* via $p(\mathbf{z} \mid \mathbf{x})$, e.g., $p(\text{Género} = \text{F} \mid \mathbf{x})$.
- **Desafíos:**
  - My difícil de especificar estas condicionales $p(\mathbf{x} \mid \mathbf{z})$ "a mano"
  - Aprendizaje no supervisado de este modelo puede ser intratable.

Modelos en variables latentes profundos

Se utilizan redes neuronales para modelar las condicionales:

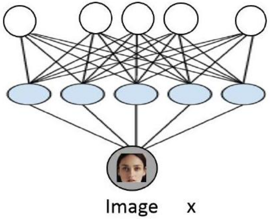
1. $\mathrm{z} \sim \mathcal{N}(0, I)$
2. $p(\mathbf{x} \mid \mathbf{z}) = \mathcal{N}\left(\mu_\theta(\mathbf{z}), \Sigma_\theta(\mathbf{z})\right)$, $\mu_\theta, \Sigma_\theta$ son redes neuronales.



Modelos en variables latentes profundos

Se utilizan redes neuronales para modelar las condicionales:

1. $\mathrm{z} \sim \mathcal{N}(0, I)$
2. $p(\mathbf{x} \mid \mathbf{z}) = \mathcal{N}\left(\mu_\theta(\mathbf{z}), \Sigma_\theta(\mathbf{z})\right)$, $\mu_\theta, \Sigma_\theta$ son redes neuronales.

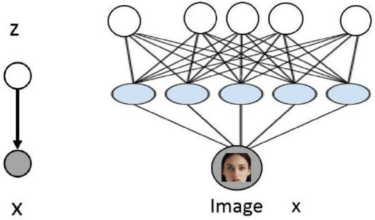


Aprendizaje no supervisado de representaciones:

Modelos en variables latentes profundos

Se utilizan redes neuronales para modelar las condicionales:

1. $\mathrm{z} \sim \mathcal{N}(0, I)$
2. $p(\mathbf{x} \mid \mathbf{z}) = \mathcal{N}\left(\mu_\theta(\mathbf{z}), \Sigma_\theta(\mathbf{z})\right)$, $\mu_\theta, \Sigma_\theta$ son redes neuronales.

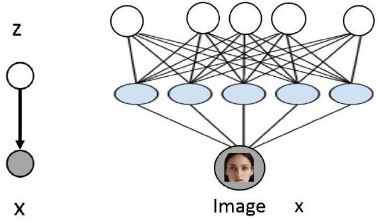


Aprendizaje no supervisado de representaciones:

- Una vez entrenado, **Z** codifica factores latentes de variación significativos (*features*).

Modelos en variables latentes profundos

Se utilizan redes neuronales para modelar las condicionales:



1. $\mathrm{z} \sim \mathcal{N}(0, I)$
2. $p(\mathbf{x} \mid \mathbf{z}) = \mathcal{N}\left(\mu_\theta(\mathbf{z}), \Sigma_\theta(\mathbf{z})\right)$, $\mu_\theta, \Sigma_\theta$ son redes neuronales.

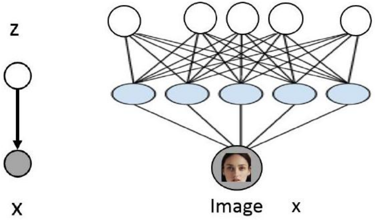
Aprendizaje no supervisado de representaciones:

- Una vez entrenado, $\mathbf{Z}$ codifica factores latentes de variación significativos (*features*).
- Como antes, los *features* se pueden calcular a través de $p(\mathbf{z} \mid \mathbf{x})$.

Modelos en variables latentes profundos

Se utilizan redes neuronales para modelar las condicionales:

1. $\mathrm{z} \sim \mathcal{N}(0, I)$
2. $p(\mathbf{x} \mid \mathbf{z}) = \mathcal{N}\left(\mu_\theta(\mathbf{z}), \Sigma_\theta(\mathbf{z})\right)$, $\mu_\theta, \Sigma_\theta$ son redes neuronales.



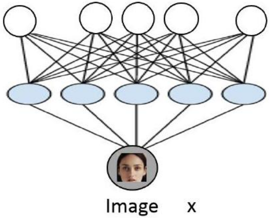
Aprendizaje no supervisado de representaciones:

- Una vez entrenado, $\mathbf{Z}$ codifica factores latentes de variación significativos (*features*).
- Como antes, los *features* se pueden calcular a través de $p(\mathbf{z} \mid \mathbf{x})$.
- La V.A. continua $\mathbf{Z}$ da una parametrización suave de $\mathbf{X}$. E.g.: caras $\mathbf{x}$ similares (misma edad, género, etc.) tienen $\mathbf{z}$ similares.

Modelos en variables latentes profundos

Se utilizan redes neuronales para modelar las condicionales:

1. $z \sim \mathcal{N}(0, I)$
2. $p(\mathbf{x} \mid \mathbf{z}) = \mathcal{N}\left(\mu_\theta(\mathbf{z}), \Sigma_\theta(\mathbf{z})\right)$, $\mu_\theta, \Sigma_\theta$ son redes neuronales.

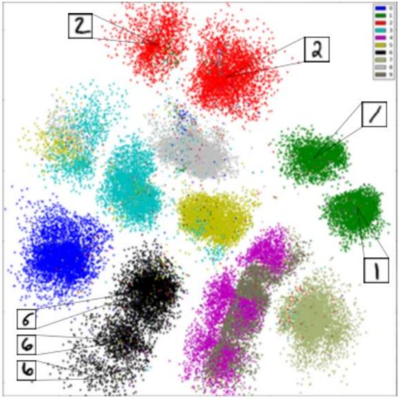


Aprendizaje no supervisado de representaciones:

- Una vez entrenado, $\mathbf{Z}$ codifica factores latentes de variación significativos (*features*).
- Como antes, los *features* se pueden calcular a través de $p(\mathbf{z} \mid \mathbf{x})$.
- La V.A. continua $\mathbf{Z}$ da una parametrización suave de $\mathbf{X}$. E.g.: caras $\mathbf{x}$ similares (misma edad, género, etc.) tienen $\mathbf{z}$ similares.

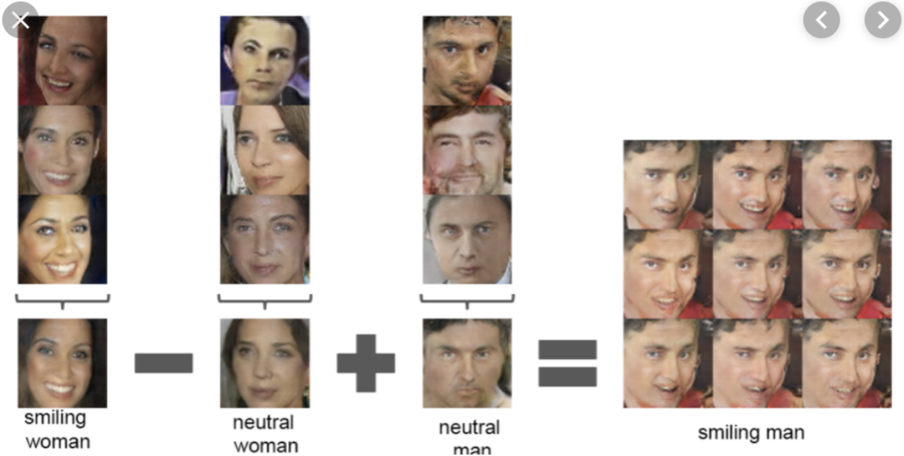
  $\Rightarrow$ Podemos hacer aritmética en $\mathbf{z}$: e.g., dados $\mathbf{x}_0$ y $\mathbf{x}_1$, inferimos $\mathbf{z}_0$ y $\mathbf{z}_1$ usando $p(\mathbf{z} \mid \mathbf{x}_i)$, interpolamos $\mathbf{z}_\lambda = \lambda \mathbf{z}_1 + (1-\lambda)\mathbf{z}_0$, y generamos $\mathbf{x}_\lambda \sim p(\mathbf{x} \mid \mathbf{z}_\lambda)$.

Ejemplo: aprendizaje no supervisado sobre dígitos manuscritos



Agrupamiento no supervisado de dígitos de MNIST

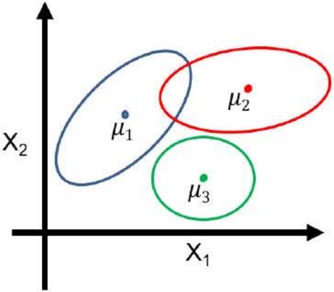
Ejemplo: aprendizaje no supervisado sobre imágenes de caras



Mezcla de gaussianas: modelo de variables latentes superficial

Mezcla de Gaussianas. Red Bayesiana: $Z \to \mathbf{X}$.

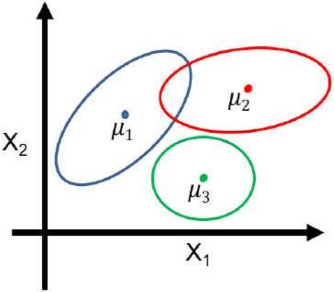
1. $Z \sim \mathrm{Cat}(1, \cdots, K)$
2. $p(\mathbf{x} \mid z = k) = \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$



Mezcla de gaussianas: modelo de variables latentes superficial

Mezcla de Gaussianas. Red Bayesiana: $Z \to \mathbf{X}$.

1. $Z \sim \text{Cat}(1, \cdots, K)$
2. $p(\mathbf{x} \mid z = k) = \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$



Proceso generativo:

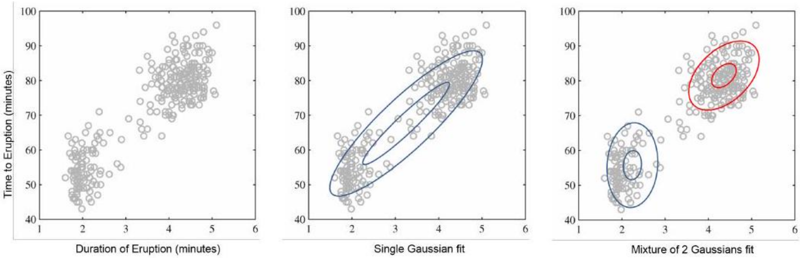
1. Elegir un componente de mezcla k muestreando z
2. Generar un punto de datos muestreando de esa gaussiana

Mezcla de gaussianas: modelo de variables latentes superficial

1. $Z \sim \text{Cat}(1, \cdots, K)$
2. $p(\mathbf{x} \mid z = k) = \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$

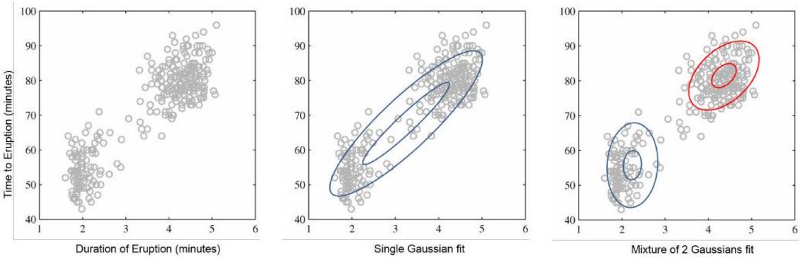
Mezcla de gaussianas: modelo de variables latentes superficial

1. $Z \sim \mathrm{Cat}(1, \cdots, K)$
2. $p(\mathbf{x} \mid z = k) = \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$



Mezcla de gaussianas: modelo de variables latentes superficial

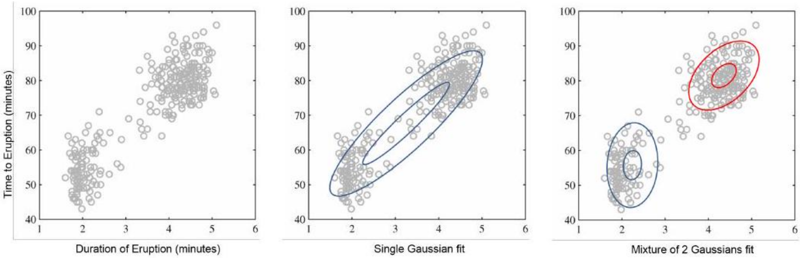
1. $Z \sim \mathrm{Cat}(1, \cdots, K)$
2. $p(\mathbf{x} \mid z = k) = \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$



- **Clustering:** La densidad a posteriori $p(z \mid \mathbf{x})$ identifica el componente de mezcla.

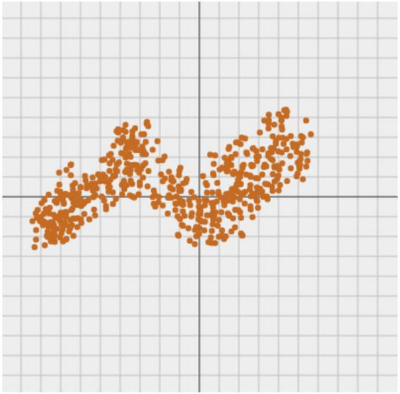
Mezcla de gaussianas: modelo de variables latentes superficial

1. $Z \sim \mathrm{Cat}(1, \cdots, K)$
2. $p(\mathbf{x} \mid z = k) = \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$

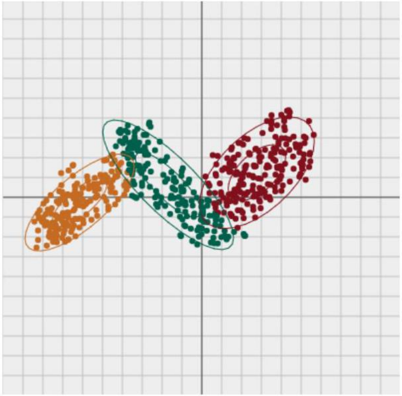


- **Clustering:** La densidad a posteriori $p(z \mid \mathbf{x})$ identifica el componente de mezcla.
- **Aprendizaje no supervisado:** esperamos aprender de datos no etiquetados.

Aprendizaje no supervisado



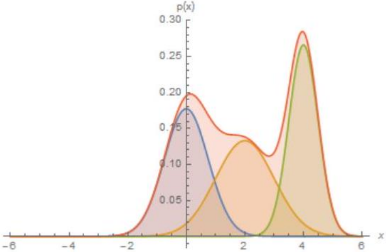
Aprendizaje no supervisado



Se muestra la densidad a posteriori de que un dato sea generado por el i-ésimo componente de la mezcla, $\mathrm{P}(Z = i \mid \mathbf{x})$

Modelos de mezcla

Motivación alternativa: combinar modelos simples para crear uno más complejo y más expresivo.



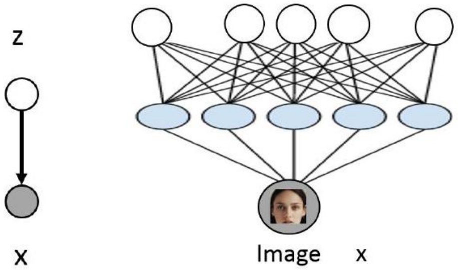
$$p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z}) = \sum_{\mathbf{z}} p(\mathbf{z}) p(\mathbf{x} \mid \mathbf{z}) = \sum_{k=1}^{K} p(\mathbf{z} = k) \underbrace{\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}_{\text{componente } k\text{-ésimo}}$$

Modelos gaussianos profundos

Podemos extender los GMM a una mezcla de un número infinito de gaussianas:

1. $p(\mathbf{z}) = \mathcal{N}(0, I)$
2. $p(\mathbf{x} \mid \mathbf{z}) = \mathcal{N}\left(\boldsymbol{\mu}_\theta(\mathbf{z}), \boldsymbol{\Sigma}_\theta(\mathbf{z})\right)$

$\mu_\theta, \Sigma_\theta$ son redes neuronales.

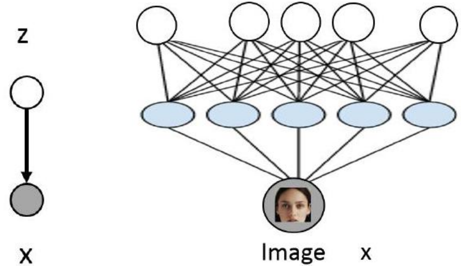


Modelos gaussianos profundos

Podemos extender los GMM a una mezcla de un número infinito de gaussianas:

1. $p(\mathbf{z}) = \mathcal{N}(0, I)$
2. $p(\mathbf{x} \mid \mathbf{z}) = \mathcal{N}\left(\boldsymbol{\mu}_\theta(\mathbf{z}), \boldsymbol{\Sigma}_\theta(\mathbf{z})\right)$

$\mu_\theta, \Sigma_\theta$ son redes neuronales.



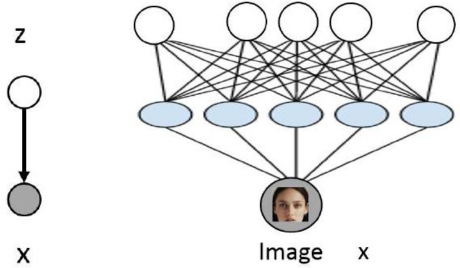
- Ejemplo:
  - $\boldsymbol{\mu}_\theta(\mathbf{z}) = \sigma(A\mathbf{z} + \mathbf{c}) = \left(\sigma\left(\mathbf{a}_1^T\mathbf{z} + c_1\right), \sigma\left(\mathbf{a}_2^T\mathbf{z} + c_2\right)\right) = (\mu_1(\mathbf{z}), \mu_2(\mathbf{z}))$
  - $\boldsymbol{\Sigma}_\theta(\mathbf{z}) = \mathrm{diag}(\exp(\sigma(B\mathbf{z} + \mathbf{d}))) = \begin{pmatrix} \exp\left(\sigma\left(\mathbf{b}_1^T\mathbf{z} + d_1\right)\right) & 0 \\ 0 & \exp\left(\sigma\left(\mathbf{b}_2^T\mathbf{z} + d_2\right)\right) \end{pmatrix}$
  - $\boldsymbol{\theta} = (A, B, \mathbf{c}, \mathbf{d})$

Modelos gaussianos profundos

Podemos extender los GMM a una mezcla de un número infinito de gaussianas:

1. $p(\mathbf{z}) = \mathcal{N}(0, I)$
2. $p(\mathbf{x} \mid \mathbf{z}) = \mathcal{N}\left(\boldsymbol{\mu}_\theta(\mathbf{z}), \boldsymbol{\Sigma}_\theta(\mathbf{z})\right)$

$\mu_\theta, \Sigma_\theta$ son redes neuronales.



- Ejemplo:
  - $\boldsymbol{\mu}_\theta(\mathbf{z}) = \sigma(A\mathbf{z} + \mathbf{c}) = \left(\sigma\left(\mathbf{a}_1^T\mathbf{z} + c_1\right), \sigma\left(\mathbf{a}_2^T\mathbf{z} + c_2\right)\right) = (\mu_1(\mathbf{z}), \mu_2(\mathbf{z}))$
  - $\boldsymbol{\Sigma}_\theta(\mathbf{z}) = \mathrm{diag}(\exp(\sigma(B\mathbf{z} + \mathbf{d}))) = \begin{pmatrix} \exp\left(\sigma\left(\mathbf{b}_1^T\mathbf{z} + d_1\right)\right) & 0 \\ 0 & \exp\left(\sigma\left(\mathbf{b}_2^T\mathbf{z} + d_2\right)\right) \end{pmatrix}$
  - $\boldsymbol{\theta} = (A, B, \mathbf{c}, \mathbf{d})$
- A pesar de que $p(\mathbf{x} \mid \mathbf{z})$ es simple, la marginal $p(\mathbf{x})$ es muy compleja/flexible.

Recap

Modelos en variables latentes:

Recap

Modelos en variables latentes:

- Permiten definir modelos complejos $p(\mathbf{x})$ combinando modelos más simples $p(\mathbf{x} \mid \mathbf{z})$.

Recap

Modelos en variables latentes:

- Permiten definir modelos complejos $p(\mathbf{x})$ combinando modelos más simples $p(\mathbf{x} \mid \mathbf{z})$.
- Modelo natural para tareas de aprendizaje no supervisado (clustering, aprendizaje no supervisado de representaciones, etc.).

Recap

Modelos en variables latentes:

- Permiten definir modelos complejos $p(\mathbf{x})$ combinando modelos más simples $p(\mathbf{x} \mid \mathbf{z})$.
- Modelo natural para tareas de aprendizaje no supervisado (clustering, aprendizaje no supervisado de representaciones, etc.).
- Pero: son en general más difíciles de entrenar que los modelos autorregresivos.

Recap

Modelos en variables latentes:

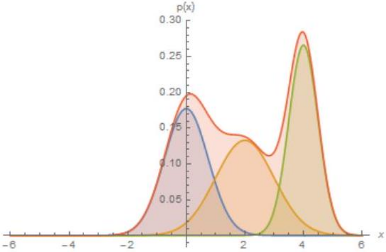
- Permiten definir modelos complejos $p(\mathbf{x})$ combinando modelos más simples $p(\mathbf{x} \mid \mathbf{z})$.
- Modelo natural para tareas de aprendizaje no supervisado (clustering, aprendizaje no supervisado de representaciones, etc.).
- Pero: son en general más difíciles de entrenar que los modelos autorregresivos.

Obs.: los Normalizing Flows son también un tipo de modelo en variables latentes ($\mathbf{Z} \mapsto \mathbf{X}$).

Verosimilitud marginal para modelos en variables latentes

Queremos entrenar un modelo en variables latentes maximizando la verosimilitud marginal de los datos $p(\mathbf{x})$. Ejemplo: GMMs,

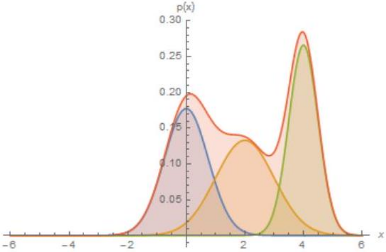
$$p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z}) = \sum_{\mathbf{z}} p(\mathbf{z}) p(\mathbf{x} \mid \mathbf{z}) = \sum_{k=1}^{K} p(\mathbf{z} = k) \underbrace{\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}_{\text{componente } k\text{-ésimo}}$$



Verosimilitud marginal para modelos en variables latentes

Queremos entrenar un modelo en variables latentes maximizando la verosimilitud marginal de los datos $p(\mathbf{x})$. Ejemplo: GMMs,

$$p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z}) = \sum_{\mathbf{z}} p(\mathbf{z}) p(\mathbf{x} \mid \mathbf{z}) = \sum_{k=1}^{K} p(\mathbf{z} = k) \underbrace{\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}_{\text{componente } k\text{-ésimo}}$$



Es difícil: función objetivo **no convexa** en los parámetros $\boldsymbol{\theta}$, y en general **intratable para de calcular en altas dimensiones (se integra en z).**

Desafíos en la maximización de la verosimilitud marginal

Recordemos la maximización de la verosimilitud de los datos:

$$\log \prod_{\mathbf{x}\in\mathcal{D}} p(\mathbf{x};\theta) = \sum_{\mathbf{x}\in\mathcal{D}} \log p(\mathbf{x};\theta) = \sum_{\mathbf{x}\in\mathcal{D}} \log \sum_{\mathbf{z}} p(\mathbf{x},\mathbf{z};\theta)$$

Desafíos en la maximización de la verosimilitud marginal

Recordemos la maximización de la verosimilitud de los datos:

$$\log \prod_{\mathbf{x} \in \mathcal{D}} p(\mathbf{x}; \theta) = \sum_{\mathbf{x} \in \mathcal{D}} \log p(\mathbf{x}; \theta) = \sum_{\mathbf{x} \in \mathcal{D}} \log \sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z}; \theta)$$

- Evaluar $\log \sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z}; \theta)$ puede ser inmanejable:

Desafíos en la maximización de la verosimilitud marginal

Recordemos la maximización de la verosimilitud de los datos:

$$\log \prod_{\mathbf{x}\in\mathcal{D}} p(\mathbf{x};\theta) = \sum_{\mathbf{x}\in\mathcal{D}} \log p(\mathbf{x};\theta) = \sum_{\mathbf{x}\in\mathcal{D}} \log \sum_{\mathbf{z}} p(\mathbf{x},\mathbf{z};\theta)$$

- Evaluar $\log \sum_{\mathbf{z}} p(\mathbf{x},\mathbf{z};\theta)$ puede ser inmanejable:
  - Ejemplo: 30 variables latentes binarias, $\mathbf{z} \in \{0,1\}^{30} \Rightarrow$ La suma en $\mathbf{z}$ tiene 2^{30} términos.

Desafíos en la maximización de la verosimilitud marginal

Recordemos la maximización de la verosimilitud de los datos:

$$\log \prod_{\mathbf{x}\in\mathcal{D}} p(\mathbf{x};\theta) = \sum_{\mathbf{x}\in\mathcal{D}} \log p(\mathbf{x};\theta) = \sum_{\mathbf{x}\in\mathcal{D}} \log \sum_{\mathbf{z}} p(\mathbf{x},\mathbf{z};\theta)$$

- Evaluar $\log \sum_{\mathbf{z}} p(\mathbf{x},\mathbf{z};\theta)$ puede ser inmanejable:
  - Ejemplo: 30 variables latentes binarias, $\mathbf{z} \in \{0,1\}^{30} \Rightarrow$ La suma en $\mathbf{z}$ tiene 2^{30} términos.
  - Caso continuo: $\log \int_{\mathbf{z}} p(\mathbf{x},\mathbf{z};\theta)d\mathbf{z}$ es intratable

Desafíos en la maximización de la verosimilitud marginal

Recordemos la maximización de la verosimilitud de los datos:

$$\log \prod_{\mathbf{x} \in \mathcal{D}} p(\mathbf{x}; \theta) = \sum_{\mathbf{x} \in \mathcal{D}} \log p(\mathbf{x}; \theta) = \sum_{\mathbf{x} \in \mathcal{D}} \log \sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z}; \theta)$$

- Evaluar $\log \sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z}; \theta)$ puede ser inmanejable:
  - Ejemplo: 30 variables latentes binarias, $\mathbf{z} \in \{0, 1\}^{30} \Rightarrow$ La suma en $\mathbf{z}$ tiene 2^{30} términos.
  - Caso continuo: $\log \int_{\mathbf{z}} p(\mathbf{x}, \mathbf{z}; \theta) d\mathbf{z}$ es intratable
  - Los gradientes $\nabla_{\boldsymbol{\theta}} \log \mathcal{D}$ son también difíciles de calcular.

Desafíos en la maximización de la verosimilitud marginal

Recordemos la maximización de la verosimilitud de los datos:

$$\log \prod_{\mathbf{x}\in\mathcal{D}} p(\mathbf{x};\theta) = \sum_{\mathbf{x}\in\mathcal{D}} \log p(\mathbf{x};\theta) = \sum_{\mathbf{x}\in\mathcal{D}} \log \sum_{\mathbf{z}} p(\mathbf{x},\mathbf{z};\theta)$$

- Evaluar $\log \sum_{\mathbf{z}} p(\mathbf{x},\mathbf{z};\theta)$ puede ser inmanejable:
  - Ejemplo: 30 variables latentes binarias, $\mathbf{z} \in \{0,1\}^{30} \Rightarrow$ La suma en $\mathbf{z}$ tiene 2^{30} términos.
  - Caso continuo: $\log \int_{\mathbf{z}} p(\mathbf{x},\mathbf{z};\theta) d\mathbf{z}$ es intratable
  - Los gradientes $\nabla_{\boldsymbol{\theta}} \log \mathcal{D}$ son también difíciles de calcular.
- **Entrenamiento:** en cada iteración, una evaluación de gradiente por cada dato de entrenamiento $\Rightarrow$ Es necesario hacer aproximaciones para que la evaluación de los gradientes sea más eficiente.

Aproximación de la verosimilitud marginal: 1er intento - Monte Carlo ingenuo

$$p_\theta(\mathbf{x}) = \sum_{\mathbf{z}} p_\theta(\mathbf{x}, \mathbf{z}) = |\mathcal{Z}| \sum_{\mathbf{z} \in \mathcal{Z}} \frac{1}{|\mathcal{Z}|} p_\theta(\mathbf{x}, \mathbf{z}) = |\mathcal{Z}| \mathbb{E}_{\mathbf{z} \sim \mathrm{Uniform}(\mathcal{Z})} \left[p_\theta(\mathbf{x}, \mathbf{z}) \right]$$

Aproximación de la verosimilitud marginal: 1er intento - Monte Carlo ingenuo

$$p_\theta(\mathbf{x}) = \sum_{\mathbf{z}} p_\theta(\mathbf{x}, \mathbf{z}) = |\mathcal{Z}| \sum_{\mathbf{z} \in \mathcal{Z}} \frac{1}{|\mathcal{Z}|} p_\theta(\mathbf{x}, \mathbf{z}) = |\mathcal{Z}| \mathbb{E}_{\mathbf{z} \sim \mathsf{Uniform}(\mathcal{Z})} \left[p_\theta(\mathbf{x}, \mathbf{z}) \right]$$

Podemos pensarlo como una esperanza (intratable) $\Rightarrow$ Monte Carlo:

Aproximación de la verosimilitud marginal: 1er intento - Monte Carlo ingenuo

$$p_\theta(\mathbf{x}) = \sum_{\mathbf{z}} p_\theta(\mathbf{x}, \mathbf{z}) = |\mathcal{Z}| \sum_{\mathbf{z} \in \mathcal{Z}} \frac{1}{|\mathcal{Z}|} p_\theta(\mathbf{x}, \mathbf{z}) = |\mathcal{Z}| \mathbb{E}_{\mathbf{z} \sim \text{Uniform}(\mathcal{Z})} \left[p_\theta(\mathbf{x}, \mathbf{z}) \right]$$

Podemos pensarlo como una esperanza (intratable) $\Rightarrow$ Monte Carlo:

1. Muestrear $\mathbf{z}^{(1)}, \cdots, \mathbf{z}^{(k)} \sim \mathcal{U}(\mathcal{Z})$
2. Aproximar la esperanza por el promedio: $\sum_{\mathbf{z}} p_\theta(\mathbf{x}, \mathbf{z}) \approx |\mathcal{Z}| \frac{1}{k} \sum_{j=1}^{k} p_\theta \left(\mathbf{x}, \mathbf{z}^{(j)} \right)$

Aproximación de la verosimilitud marginal: 1er intento - Monte Carlo ingenuo

$$p_\theta(\mathbf{x}) = \sum_{\mathbf{z}} p_\theta(\mathbf{x}, \mathbf{z}) = |\mathcal{Z}| \sum_{\mathbf{z} \in \mathcal{Z}} \frac{1}{|\mathcal{Z}|} p_\theta(\mathbf{x}, \mathbf{z}) = |\mathcal{Z}| \mathbb{E}_{\mathbf{z} \sim \text{Uniform}(\mathcal{Z})} \left[p_\theta(\mathbf{x}, \mathbf{z}) \right]$$

Podemos pensarlo como una esperanza (intratable) $\Rightarrow$ Monte Carlo:

1. Muestrear $\mathbf{z}^{(1)}, \cdots, \mathbf{z}^{(k)} \sim \mathcal{U}(\mathcal{Z})$
2. Aproximar la esperanza por el promedio: $\sum_{\mathbf{z}} p_\theta(\mathbf{x}, \mathbf{z}) \approx |\mathcal{Z}| \frac{1}{k} \sum_{j=1}^{k} p_\theta\left(\mathbf{x}, \mathbf{z}^{(j)}\right)$

No funciona en la práctica en espacios de muy alta dimensionalidad:

Aproximación de la verosimilitud marginal: 1$^{\text{er}}$ intento - Monte Carlo ingenuo

$$p_\theta(\mathbf{x}) = \sum_{\mathbf{z}} p_\theta(\mathbf{x}, \mathbf{z}) = |\mathcal{Z}| \sum_{\mathbf{z}\in\mathcal{Z}} \frac{1}{|\mathcal{Z}|} p_\theta(\mathbf{x}, \mathbf{z}) = |\mathcal{Z}| \mathbb{E}_{\mathbf{z}\sim\text{Uniform}(\mathcal{Z})} [p_\theta(\mathbf{x}, \mathbf{z})]$$

Podemos pensarlo como una esperanza (intratable) $\Rightarrow$ Monte Carlo:

1. Muestrear $\mathbf{z}^{(1)}, \cdots, \mathbf{z}^{(k)} \sim \mathcal{U}(\mathcal{Z})$
2. Aproximar la esperanza por el promedio: $\sum_{\mathbf{z}} p_\theta(\mathbf{x}, \mathbf{z}) \approx |\mathcal{Z}| \frac{1}{k} \sum_{j=1}^{k} p_\theta\left(\mathbf{x}, \mathbf{z}^{(j)}\right)$

No funciona en la práctica en espacios de muy alta dimensionalidad:

- $p_\theta(\mathbf{x}, \mathbf{z})$ es muy bajo para la mayoría de los $\mathbf{z}$.

Aproximación de la verosimilitud marginal: 1er intento - Monte Carlo ingenuo

$$p_\theta(\mathbf{x}) = \sum_{\mathbf{z}} p_\theta(\mathbf{x}, \mathbf{z}) = |\mathcal{Z}| \sum_{\mathbf{z} \in \mathcal{Z}} \frac{1}{|\mathcal{Z}|} p_\theta(\mathbf{x}, \mathbf{z}) = |\mathcal{Z}| \mathbb{E}_{\mathbf{z} \sim \text{Uniform}(\mathcal{Z})} \left[p_\theta(\mathbf{x}, \mathbf{z}) \right]$$

Podemos pensarlo como una esperanza (intratable) $\Rightarrow$ Monte Carlo:

1. Muestrear $\mathbf{z}^{(1)}, \cdots, \mathbf{z}^{(k)} \sim \mathcal{U}(\mathcal{Z})$
2. Aproximar la esperanza por el promedio: $\sum_{\mathbf{z}} p_\theta(\mathbf{x}, \mathbf{z}) \approx |\mathcal{Z}| \frac{1}{k} \sum_{j=1}^{k} p_\theta \left(\mathbf{x}, \mathbf{z}^{(j)} \right)$

No funciona en la práctica en espacios de muy alta dimensionalidad:

- $p_\theta(\mathbf{x}, \mathbf{z})$ es muy bajo para la mayoría de los $\mathbf{z}$.
- Para los $\mathbf{z}$ en los que $p_\theta(\mathbf{x}, \mathbf{z})$ no es chica, las chances de muestrearlos con una uniforme son bajas

Aproximación de la verosimilitud marginal: 1er intento - Monte Carlo ingenuo

$$p_\theta(\mathbf{x}) = \sum_{\mathbf{z}} p_\theta(\mathbf{x},\mathbf{z}) = |\mathcal{Z}| \sum_{\mathbf{z}\in\mathcal{Z}} \frac{1}{|\mathcal{Z}|} p_\theta(\mathbf{x},\mathbf{z}) = |\mathcal{Z}| \mathbb{E}_{\mathbf{z}\sim \text{Uniform}(\mathcal{Z})} \left[p_\theta(\mathbf{x},\mathbf{z})\right]$$

Podemos pensarlo como una esperanza (intratable) $\Rightarrow$ Monte Carlo:

1. Muestrear $\mathbf{z}^{(1)}, \cdots, \mathbf{z}^{(k)} \sim \mathcal{U}(\mathcal{Z})$
2. Aproximar la esperanza por el promedio: $\sum_{\mathbf{z}} p_\theta(\mathbf{x},\mathbf{z}) \approx |\mathcal{Z}| \frac{1}{k} \sum_{j=1}^{k} p_\theta\left(\mathbf{x},\mathbf{z}^{(j)}\right)$

No funciona en la práctica en espacios de muy alta dimensionalidad:

- $p_\theta(\mathbf{x},\mathbf{z})$ es muy bajo para la mayoría de los $\mathbf{z}$.
- Para los $\mathbf{z}$ en los que $p_\theta(\mathbf{x},\mathbf{z})$ no es chica, las chances de muestrearlos con una uniforme son bajas $\Rightarrow$ Necesitamos una forma inteligente de seleccionar $\mathbf{z}^{(j)}$ para reducir la varianza del estimador.

Aproximación de la verosimilitud marginal: 2° intento - Muestreo de importancia

$$\sum_{\mathbf{z} \in \mathcal{Z}} p_\theta(\mathbf{x}, \mathbf{z}) = \sum_{\mathbf{z} \in \mathcal{Z}} \frac{q(\mathbf{z})}{q(\mathbf{z})} p_\theta(\mathbf{x}, \mathbf{z}) = \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z})} \left[\frac{p_\theta(\mathbf{x}, \mathbf{z})}{q(\mathbf{z})} \right]$$

1. Muestrear $\mathbf{z}^{(1)}, \cdots, \mathbf{z}^{(k)}$ de $q(\mathbf{z})$.
2. Aproximar la esperanza con el promedio de la muestra: $p_\theta(\mathbf{x}) \approx \frac{1}{k} \sum_{j=1}^{k} \frac{p_\theta\left(\mathbf{x}, \mathbf{z}^{(j)}\right)}{q\left(\mathbf{z}^{(j)}\right)}$

Aproximación de la verosimilitud marginal: 2° intento - Muestreo de importancia

$$\sum_{\mathbf{z}\in\mathcal{Z}} p_\theta(\mathbf{x},\mathbf{z}) = \sum_{\mathbf{z}\in\mathcal{Z}} \frac{q(\mathbf{z})}{q(\mathbf{z})} p_\theta(\mathbf{x},\mathbf{z}) = \mathbb{E}_{\mathbf{z}\sim q(\mathbf{z})}\left[\frac{p_\theta(\mathbf{x},\mathbf{z})}{q(\mathbf{z})}\right]$$

1. Muestrear $\mathbf{z}^{(1)}, \cdots, \mathbf{z}^{(k)}$ de $q(\mathbf{z})$.
2. Aproximar la esperanza con el promedio de la muestra: $p_\theta(\mathbf{x}) \approx \frac{1}{k}\sum_{j=1}^{k} \frac{p_\theta\left(\mathbf{x},\mathbf{z}^{(j)}\right)}{q\left(\mathbf{z}^{(j)}\right)}$

- Recordar que para entrenar necesitamos la log-verosimilitud $\log p_\theta(\mathbf{x})$

Aproximación de la verosimilitud marginal: 2° intento - Muestreo de importancia

$$\sum_{\mathbf{z}\in\mathcal{Z}} p_\theta(\mathbf{x},\mathbf{z}) = \sum_{\mathbf{z}\in\mathcal{Z}} \frac{q(\mathbf{z})}{q(\mathbf{z})} p_\theta(\mathbf{x},\mathbf{z}) = \mathbb{E}_{\mathbf{z}\sim q(\mathbf{z})}\left[\frac{p_\theta(\mathbf{x},\mathbf{z})}{q(\mathbf{z})}\right]$$

1. Muestrear $\mathbf{z}^{(1)},\cdots,\mathbf{z}^{(k)}$ de $q(\mathbf{z})$.
2. Aproximar la esperanza con el promedio de la muestra: $p_\theta(\mathbf{x}) \approx \frac{1}{k}\sum_{j=1}^{k} \frac{p_\theta\left(\mathbf{x},\mathbf{z}^{(j)}\right)}{q\left(\mathbf{z}^{(j)}\right)}$

- Recordar que para entrenar necesitamos la log-verosimilitud $\log p_\theta(\mathbf{x})$
- Nos veríamos tentados a intercambiar el log y la esperanza y aplicar Monte Carlo.

Aproximación de la verosimilitud marginal: 2° intento - Muestreo de importancia

$$\sum_{\mathbf{z}\in\mathcal{Z}} p_\theta(\mathbf{x},\mathbf{z}) = \sum_{\mathbf{z}\in\mathcal{Z}} \frac{q(\mathbf{z})}{q(\mathbf{z})} p_\theta(\mathbf{x},\mathbf{z}) = \mathbb{E}_{\mathbf{z}\sim q(\mathbf{z})}\left[\frac{p_\theta(\mathbf{x},\mathbf{z})}{q(\mathbf{z})}\right]$$

1. Muestrear $\mathbf{z}^{(1)}, \cdots, \mathbf{z}^{(k)}$ de $q(\mathbf{z})$.
2. Aproximar la esperanza con el promedio de la muestra: $p_\theta(\mathbf{x}) \approx \frac{1}{k}\sum_{j=1}^{k} \frac{p_\theta\left(\mathbf{x},\mathbf{z}^{(j)}\right)}{q\left(\mathbf{z}^{(j)}\right)}$

- Recordar que para entrenar necesitamos la log-verosimilitud $\log p_\theta(\mathbf{x})$
- Nos veríamos tentados a intercambiar el log y la esperanza y aplicar Monte Carlo.
- Sin embargo, sabemos que

$$\mathbb{E}_{\mathbf{z}\sim q(\mathbf{z})}\left[\log\left(\frac{p_\theta(\mathbf{x},\mathbf{z})}{q(\mathbf{z})}\right)\right] \neq \log\left(\mathbb{E}_{\mathbf{z}\sim q(\mathbf{z})}\left[\frac{p_\theta(\mathbf{x},\mathbf{z})}{q(\mathbf{z})}\right]\right)$$

Aproximación de la verosimilitud marginal: 2° intento - Muestreo de importancia

$$\sum_{\mathbf{z}\in\mathcal{Z}} p_\theta(\mathbf{x},\mathbf{z}) = \sum_{\mathbf{z}\in\mathcal{Z}} \frac{q(\mathbf{z})}{q(\mathbf{z})} p_\theta(\mathbf{x},\mathbf{z}) = \mathbb{E}_{\mathbf{z}\sim q(\mathbf{z})}\left[\frac{p_\theta(\mathbf{x},\mathbf{z})}{q(\mathbf{z})}\right]$$

1. Muestrear $\mathbf{z}^{(1)}, \cdots, \mathbf{z}^{(k)}$ de $q(\mathbf{z})$.
2. Aproximar la esperanza con el promedio de la muestra: $p_\theta(\mathbf{x}) \approx \frac{1}{k}\sum_{j=1}^{k} \frac{p_\theta\left(\mathbf{x},\mathbf{z}^{(j)}\right)}{q\left(\mathbf{z}^{(j)}\right)}$

- Recordar que para entrenar necesitamos la log-verosimilitud $\log p_\theta(\mathbf{x})$
- Nos veríamos tentados a intercambiar el log y la esperanza y aplicar Monte Carlo.
- Sin embargo, sabemos que

$$\mathbb{E}_{\mathbf{z}\sim q(\mathbf{z})}\left[\log\left(\frac{p_\theta(\mathbf{x},\mathbf{z})}{q(\mathbf{z})}\right)\right] \neq \log\left(\mathbb{E}_{\mathbf{z}\sim q(\mathbf{z})}\left[\frac{p_\theta(\mathbf{x},\mathbf{z})}{q(\mathbf{z})}\right]\right)$$

$\Rightarrow$ Necesitamos otro enfoque.

Desigualdad de Jensen

Queremos aproximar la log-verosimilitud marginal:

$$\log\left(\sum_{\mathbf{z}\in\mathcal{Z}} p_\theta(\mathbf{x},\mathbf{z})\right) = \log\left(\sum_{\mathbf{z}\in\mathcal{Z}} \frac{q(\mathbf{z})}{q(\mathbf{z})} p_\theta(\mathbf{x},\mathbf{z})\right) = \log\left(\mathbb{E}_{\mathbf{z}\sim q(\mathbf{z})}\left[\frac{p_\theta(\mathbf{x},\mathbf{z})}{q(\mathbf{z})}\right]\right)$$

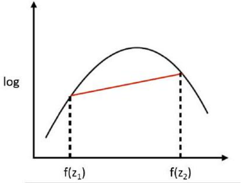
Desigualdad de Jensen

Queremos aproximar la log-verosimilitud marginal:

$$\log\left(\sum_{\mathbf{z}\in\mathcal{Z}} p_\theta(\mathbf{x},\mathbf{z})\right) = \log\left(\sum_{\mathbf{z}\in\mathcal{Z}} \frac{q(\mathbf{z})}{q(\mathbf{z})} p_\theta(\mathbf{x},\mathbf{z})\right) = \log\left(\mathbb{E}_{\mathbf{z}\sim q(\mathbf{z})}\left[\frac{p_\theta(\mathbf{x},\mathbf{z})}{q(\mathbf{z})}\right]\right)$$

Idea: usar la desigualdad de Jensen (para funciones cóncavas como el log)

$$\log\left(\mathbb{E}_{\mathbf{z}\sim q(\mathbf{z})}[f(\mathbf{z})]\right) = \log\left(\sum_{\mathbf{z}} q(\mathbf{z}) f(\mathbf{z})\right) \geqslant \sum_{\mathbf{z}} q(\mathbf{z}) \log f(\mathbf{z}).$$



Evidence Lower Bound (ELBO) via desigualdad de Jensen

Queremos aproximar la log-verosimilitud marginal:

$$\log\left(\sum_{\mathbf{z}\in\mathcal{Z}} p_\theta(\mathbf{x},\mathbf{z})\right) = \log\left(\sum_{\mathbf{z}\in\mathcal{Z}} \frac{q(\mathbf{z})}{q(\mathbf{z})} p_\theta(\mathbf{x},\mathbf{z})\right) = \log\left(\mathbb{E}_{\mathbf{z}\sim q(\mathbf{z})}\left[\frac{p_\theta(\mathbf{x},\mathbf{z})}{q(\mathbf{z})}\right]\right).$$

Evidence Lower Bound (ELBO) via desigualdad de Jensen

Queremos aproximar la log-verosimilitud marginal:

$$\log\left(\sum_{\mathbf{z}\in\mathcal{Z}} p_\theta(\mathbf{x},\mathbf{z})\right) = \log\left(\sum_{\mathbf{z}\in\mathcal{Z}} \frac{q(\mathbf{z})}{q(\mathbf{z})} p_\theta(\mathbf{x},\mathbf{z})\right) = \log\left(\mathbb{E}_{\mathbf{z}\sim q(\mathbf{z})}\left[\frac{p_\theta(\mathbf{x},\mathbf{z})}{q(\mathbf{z})}\right]\right).$$

Usando la desigualdad de Jensen tenemos una cota inferior:

$$\log\left(\mathbb{E}_{\mathbf{z}\sim q(\mathbf{z})}\left[\frac{p_\theta(\mathbf{x},\mathbf{z})}{q(\mathbf{z})}\right]\right) \geqslant \mathbb{E}_{\mathbf{z}\sim q(\mathbf{z})}\left[\log\left(\frac{p_\theta(\mathbf{x},\mathbf{z})}{q(\mathbf{z})}\right)\right].$$

Evidence Lower Bound (ELBO) via desigualdad de Jensen

Queremos aproximar la log-verosimilitud marginal:

$$\log\left(\sum_{\mathbf{z}\in\mathcal{Z}} p_\theta(\mathbf{x},\mathbf{z})\right) = \log\left(\sum_{\mathbf{z}\in\mathcal{Z}} \frac{q(\mathbf{z})}{q(\mathbf{z})} p_\theta(\mathbf{x},\mathbf{z})\right) = \log\left(\mathbb{E}_{\mathbf{z}\sim q(\mathbf{z})}\left[\frac{p_\theta(\mathbf{x},\mathbf{z})}{q(\mathbf{z})}\right]\right).$$

Usando la desigualdad de Jensen tenemos una cota inferior:

$$\log\left(\mathbb{E}_{\mathbf{z}\sim q(\mathbf{z})}\left[\frac{p_\theta(\mathbf{x},\mathbf{z})}{q(\mathbf{z})}\right]\right) \geqslant \mathbb{E}_{\mathbf{z}\sim q(\mathbf{z})}\left[\log\left(\frac{p_\theta(\mathbf{x},\mathbf{z})}{q(\mathbf{z})}\right)\right].$$

Llamada Cota Inferior de Evidencia (ELBO).

Inferencia variacional

- La cota inferior de evidencia (ELBO), que vale para cualquier densidad $q(\mathbf{z})$, es:

$$
\begin{aligned}
\log p(\mathbf{x};\theta) &\geqslant \sum_{\mathbf{z}} q(\mathbf{z}) \log\left(\frac{p_\theta(\mathbf{x},\mathbf{z})}{q(\mathbf{z})}\right) \\
&= \sum_{\mathbf{z}} q(\mathbf{z}) \log p_\theta(\mathbf{x},\mathbf{z}) \underbrace{- \sum_{\mathbf{z}} q(\mathbf{z}) \log q(\mathbf{z})}_{\text{Entropía } H(q)} = \sum_{\mathbf{z}} q(\mathbf{z}) \log p_\theta(\mathbf{x},\mathbf{z}) + H(q)
\end{aligned}
$$

Esta cota es la que optimizamos también en el algoritmo EM

Inferencia variacional

- La cota inferior de evidencia (ELBO), que vale para cualquier densidad $q(\mathbf{z})$, es:

$$
\begin{aligned}
\log p(\mathbf{x};\theta) &\geqslant \sum_{\mathbf{z}} q(\mathbf{z}) \log \left(\frac{p_\theta(\mathbf{x},\mathbf{z})}{q(\mathbf{z})} \right) \\
&= \sum_{\mathbf{z}} q(\mathbf{z}) \log p_\theta(\mathbf{x},\mathbf{z}) \underbrace{- \sum_{\mathbf{z}} q(\mathbf{z}) \log q(\mathbf{z})}_{\text{Entropía } H(q)} = \sum_{\mathbf{z}} q(\mathbf{z}) \log p_\theta(\mathbf{x},\mathbf{z}) + H(q)
\end{aligned}
$$

- **Optimizar el ELBO es una forma de inferencia variacional:**

Esta cota es la que optimizamos también en el algoritmo EM

Inferencia variacional

- La cota inferior de evidencia (ELBO), que vale para cualquier densidad $q(\mathbf{z})$, es:

$$\begin{aligned}\log p(\mathbf{x};\theta) &\geqslant \sum_{\mathbf{z}} q(\mathbf{z}) \log\left(\frac{p_\theta(\mathbf{x},\mathbf{z})}{q(\mathbf{z})}\right)\\ &= \sum_{\mathbf{z}} q(\mathbf{z}) \log p_\theta(\mathbf{x},\mathbf{z}) \underbrace{- \sum_{\mathbf{z}} q(\mathbf{z}) \log q(\mathbf{z})}_{\text{Entropía } H(q)} = \sum_{\mathbf{z}} q(\mathbf{z}) \log p_\theta(\mathbf{x},\mathbf{z}) + H(q)\end{aligned}$$

- **Optimizar el ELBO es una forma de inferencia variacional:**
  1. Se elige un $q(\mathbf{z})$ que haga que la cota sea ajustada

Esta cota es la que optimizamos también en el algoritmo EM

Inferencia variacional

- La cota inferior de evidencia (ELBO), que vale para cualquier densidad $q(\mathbf{z})$, es:

$$\log p(\mathbf{x};\theta) \geqslant \sum_{\mathbf{z}} q(\mathbf{z}) \log\left(\frac{p_\theta(\mathbf{x},\mathbf{z})}{q(\mathbf{z})}\right)$$

$$= \sum_{\mathbf{z}} q(\mathbf{z}) \log p_\theta(\mathbf{x},\mathbf{z}) \underbrace{- \sum_{\mathbf{z}} q(\mathbf{z}) \log q(\mathbf{z})}_{\text{Entropía } H(q)} = \sum_{\mathbf{z}} q(\mathbf{z}) \log p_\theta(\mathbf{x},\mathbf{z}) + H(q)$$

- **Optimizar el ELBO es una forma de inferencia variacional:**
  1. Se elige un $q(\mathbf{z})$ que haga que la cota sea ajustada
  2. Se maximiza la log-verosimilitud maximizando su cota inferior.

Esta cota es la que optimizamos también en el algoritmo EM

¿Cómo hacer la cota ajustada?

$$\log p(\mathbf{x};\theta) \geqslant \sum_{\mathbf{z}} q(\mathbf{z}) \log \frac{p(\mathbf{x},\mathbf{z};\theta)}{q(\mathbf{z})}$$

¿Cómo hacer la cota ajustada?

$$\log p(\mathbf{x};\theta) \geqslant \sum_{\mathbf{z}} q(\mathbf{z}) \log \frac{p(\mathbf{x},\mathbf{z};\theta)}{q(\mathbf{z})}$$

- **La igualdad se cumple si $q(\mathbf{z}) = p(\mathbf{z} \mid \mathbf{x};\theta)$:**

¿Cómo hacer la cota ajustada?

$$\log p(\mathbf{x};\theta) \geqslant \sum_{\mathbf{z}} q(\mathbf{z}) \log \frac{p(\mathbf{x},\mathbf{z};\theta)}{q(\mathbf{z})}$$

- **La igualdad se cumple si $q(\mathbf{z}) = p(\mathbf{z} \mid \mathbf{x};\theta)$:**

$$\sum_{\mathbf{z}} p(\mathbf{z} \mid \mathbf{x};\theta) \log \frac{p(\mathbf{x},\mathbf{z};\theta)}{p(\mathbf{z} \mid \mathbf{x};\theta)} = \sum_{\mathbf{z}} p(\mathbf{z} \mid \mathbf{x};\theta) \log \frac{p(\mathbf{z} \mid \mathbf{x};\theta) p(\mathbf{x};\theta)}{p(\mathbf{z} \mid \mathbf{x};\theta)}$$

¿Cómo hacer la cota ajustada?

$$\log p(\mathbf{x};\theta) \geqslant \sum_{\mathbf{z}} q(\mathbf{z}) \log \frac{p(\mathbf{x},\mathbf{z};\theta)}{q(\mathbf{z})}$$

- **La igualdad se cumple si $q(\mathbf{z}) = p(\mathbf{z} \mid \mathbf{x};\theta)$:**

$$\begin{aligned}\sum_{\mathbf{z}} p(\mathbf{z} \mid \mathbf{x};\theta) \log \frac{p(\mathbf{x},\mathbf{z};\theta)}{p(\mathbf{z} \mid \mathbf{x};\theta)} &= \sum_{\mathbf{z}} p(\mathbf{z} \mid \mathbf{x};\theta) \log \frac{p(\mathbf{z} \mid \mathbf{x};\theta)p(\mathbf{x};\theta)}{p(\mathbf{z} \mid \mathbf{x};\theta)} \\ &= \sum_{\mathbf{z}} p(\mathbf{z} \mid \mathbf{x};\theta) \log p(\mathbf{x};\theta)\end{aligned}$$

¿Cómo hacer la cota ajustada?

$$\log p(\mathbf{x};\theta) \geqslant \sum_{\mathbf{z}} q(\mathbf{z}) \log \frac{p(\mathbf{x},\mathbf{z};\theta)}{q(\mathbf{z})}$$

- **La igualdad se cumple si $q(\mathbf{z}) = p(\mathbf{z} \mid \mathbf{x};\theta)$:**

$$\begin{aligned}
\sum_{\mathbf{z}} p(\mathbf{z} \mid \mathbf{x};\theta) \log \frac{p(\mathbf{x},\mathbf{z};\theta)}{p(\mathbf{z} \mid \mathbf{x};\theta)} &= \sum_{\mathbf{z}} p(\mathbf{z} \mid \mathbf{x};\theta) \log \frac{p(\mathbf{z} \mid \mathbf{x};\theta) p(\mathbf{x};\theta)}{p(\mathbf{z} \mid \mathbf{x};\theta)} \\
&= \sum_{\mathbf{z}} p(\mathbf{z} \mid \mathbf{x};\theta) \log p(\mathbf{x};\theta) \\
&= \log p(\mathbf{x};\theta) \sum_{\mathbf{z}} p(\mathbf{z} \mid \mathbf{x};\theta) = \log p(\mathbf{x};\theta)
\end{aligned}$$

¿Cómo hacer la cota ajustada?

$$\log p(\mathbf{x};\theta) \geqslant \sum_{\mathbf{z}} q(\mathbf{z}) \log \frac{p(\mathbf{x},\mathbf{z};\theta)}{q(\mathbf{z})}$$

- **La igualdad se cumple si $q(\mathbf{z}) = p(\mathbf{z} \mid \mathbf{x};\theta)$:**

$$\begin{aligned}
\sum_{\mathbf{z}} p(\mathbf{z} \mid \mathbf{x};\theta) \log \frac{p(\mathbf{x},\mathbf{z};\theta)}{p(\mathbf{z} \mid \mathbf{x};\theta)} &= \sum_{\mathbf{z}} p(\mathbf{z} \mid \mathbf{x};\theta) \log \frac{p(\mathbf{z} \mid \mathbf{x};\theta) p(\mathbf{x};\theta)}{p(\mathbf{z} \mid \mathbf{x};\theta)} \\
&= \sum_{\mathbf{z}} p(\mathbf{z} \mid \mathbf{x};\theta) \log p(\mathbf{x};\theta) \\
&= \log p(\mathbf{x};\theta) \sum_{\mathbf{z}} p(\mathbf{z} \mid \mathbf{x};\theta) = \log p(\mathbf{x};\theta)
\end{aligned}$$

- Confirma intuición del muestreo por importancia: debemos muestrear $\mathbf{z}$'s probables.

¿Cómo hacer la cota ajustada?

$$\log p(\mathbf{x};\theta) \geqslant \sum_{\mathbf{z}} q(\mathbf{z}) \log \frac{p(\mathbf{x},\mathbf{z};\theta)}{q(\mathbf{z})}$$

- **La igualdad se cumple si** $q(\mathbf{z}) = p(\mathbf{z} \mid \mathbf{x};\theta)$**:**

$$\begin{aligned}
\sum_{\mathbf{z}} p(\mathbf{z} \mid \mathbf{x};\theta) \log \frac{p(\mathbf{x},\mathbf{z};\theta)}{p(\mathbf{z} \mid \mathbf{x};\theta)} &= \sum_{\mathbf{z}} p(\mathbf{z} \mid \mathbf{x};\theta) \log \frac{p(\mathbf{z} \mid \mathbf{x};\theta)p(\mathbf{x};\theta)}{p(\mathbf{z} \mid \mathbf{x};\theta)} \\
&= \sum_{\mathbf{z}} p(\mathbf{z} \mid \mathbf{x};\theta) \log p(\mathbf{x};\theta) \\
&= \log p(\mathbf{x};\theta) \sum_{\mathbf{z}} p(\mathbf{z} \mid \mathbf{x};\theta) = \log p(\mathbf{x};\theta)
\end{aligned}$$

- Confirma intuición del muestreo por importancia: debemos muestrear $\mathbf{z}$'s probables.
- ¿Qué pasa si la posterior $p(\mathbf{z} \mid \mathbf{x};\theta)$ no es caclulable? ¿Qué tan relajada es la cota?

Inferencia variacional (cont.)

- ¿Cuánto difiere una pdf $q(\mathbf{z})$ cualquiera de la pdf que realiza la igualdad?

Inferencia variacional (cont.)

- ¿Cuánto difiere una pdf $q(\mathbf{z})$ cualquiera de la pdf que realiza la igualdad?

$$0 \leqslant KL(q(\mathbf{z}) \| p(\mathbf{z} \mid \mathbf{x}; \theta)) = -\sum_{\mathbf{z}} q(\mathbf{z}) \log p(\mathbf{z}, \mathbf{x}; \theta) + \log p(\mathbf{x}; \theta) - H(q).$$

Inferencia variacional (cont.)

- ¿Cuánto difiere una pdf $q(\mathbf{z})$ cualquiera de la pdf que realiza la igualdad?

$$0 \leqslant KL(q(\mathbf{z}) \| p(\mathbf{z} \mid \mathbf{x}; \theta)) = -\sum_{\mathbf{z}} q(\mathbf{z}) \log p(\mathbf{z}, \mathbf{x}; \theta) + \log p(\mathbf{x}; \theta) - H(q).$$

- Reordenando, volvemos a obtener el ELBO:

$$\log p(\mathbf{x}; \theta) \geqslant \sum_{\mathbf{z}} q(\mathbf{z}) \log p(\mathbf{z}, \mathbf{x}; \theta) + H(q).$$

Inferencia variacional (cont.)

- ¿Cuánto difiere una pdf $q(\mathbf{z})$ cualquiera de la pdf que realiza la igualdad?

$$0 \leqslant KL(q(\mathbf{z}) \| p(\mathbf{z} \mid \mathbf{x}; \theta)) = -\sum_{\mathbf{z}} q(\mathbf{z}) \log p(\mathbf{z}, \mathbf{x}; \theta) + \log p(\mathbf{x}; \theta) - H(q).$$

- Reordenando, volvemos a obtener el ELBO:

$$\log p(\mathbf{x}; \theta) \geqslant \sum_{\mathbf{z}} q(\mathbf{z}) \log p(\mathbf{z}, \mathbf{x}; \theta) + H(q).$$

- La igualdad se cumple si $q(\mathbf{z}) = p(\mathbf{z} \mid \mathbf{x}; \theta)$ porque $D_{KL}(q(\mathbf{z}) \| p(\mathbf{z} \mid \mathbf{x}; \theta)) = 0$:

$$\log p(\mathbf{x}; \theta) = \sum_{\mathbf{z}} q(\mathbf{z}) \log p(\mathbf{z}, \mathbf{x}; \theta) + H(q).$$

Inferencia variacional (cont.)

- ¿Cuánto difiere una pdf $q(\mathbf{z})$ cualquiera de la pdf que realiza la igualdad?

$$0 \leqslant KL(q(\mathbf{z}) \| p(\mathbf{z} \mid \mathbf{x}; \theta)) = -\sum_{\mathbf{z}} q(\mathbf{z}) \log p(\mathbf{z}, \mathbf{x}; \theta) + \log p(\mathbf{x}; \theta) - H(q).$$

- Reordenando, volvemos a obtener el ELBO:

$$\log p(\mathbf{x}; \theta) \geqslant \sum_{\mathbf{z}} q(\mathbf{z}) \log p(\mathbf{z}, \mathbf{x}; \theta) + H(q).$$

- La igualdad se cumple si $q(\mathbf{z}) = p(\mathbf{z} \mid \mathbf{x}; \theta)$ porque $D_{KL}(q(\mathbf{z}) \| p(\mathbf{z} \mid \mathbf{x}; \theta)) = 0$:

$$\log p(\mathbf{x}; \theta) = \sum_{\mathbf{z}} q(\mathbf{z}) \log p(\mathbf{z}, \mathbf{x}; \theta) + H(q).$$

- En suma, $\log p(\mathbf{x}; \theta) = \mathrm{ELBO} + D_{KL}(q(\mathbf{z}) \| p(\mathbf{z} \mid \mathbf{x}; \theta))$

Inferencia variacional (cont.)

- ¿Cuánto difiere una pdf $q(\mathbf{z})$ cualquiera de la pdf que realiza la igualdad?

$$0 \leqslant KL(q(\mathbf{z}) \| p(\mathbf{z} \mid \mathbf{x}; \theta)) = -\sum_{\mathbf{z}} q(\mathbf{z}) \log p(\mathbf{z}, \mathbf{x}; \theta) + \log p(\mathbf{x}; \theta) - H(q).$$

- Reordenando, volvemos a obtener el ELBO:

$$\log p(\mathbf{x}; \theta) \geqslant \sum_{\mathbf{z}} q(\mathbf{z}) \log p(\mathbf{z}, \mathbf{x}; \theta) + H(q).$$

- La igualdad se cumple si $q(\mathbf{z}) = p(\mathbf{z} \mid \mathbf{x}; \theta)$ porque $D_{KL}(q(\mathbf{z}) \| p(\mathbf{z} \mid \mathbf{x}; \theta)) = 0$:

$$\log p(\mathbf{x}; \theta) = \sum_{\mathbf{z}} q(\mathbf{z}) \log p(\mathbf{z}, \mathbf{x}; \theta) + H(q).$$

- En suma, $\log p(\mathbf{x}; \theta) = \mathrm{ELBO} + D_{KL}(q(\mathbf{z}) \| p(\mathbf{z} \mid \mathbf{x}; \theta))$ $\Rightarrow$ Cuanto más cerca esté $q(\mathbf{z})$ de $p(\mathbf{z} \mid \mathbf{x}; \theta)$, más cerca estará la ELBO de la verdadera log-verosimilitud.

Inferencia variacional: ¿Cómo elegir $q(\mathbf{z})$?

¿Qué pasa si la posterior $p(\mathbf{z} \mid \mathbf{x}; \theta)$ es intratable?

Inferencia variacional: ¿Cómo elegir $q(\mathbf{z})$?

¿Qué pasa si la posterior $p(\mathbf{z} \mid \mathbf{x}; \theta)$ es intratable?

- Sea $q(\mathbf{z}; \phi)$ una pdf (tratable) parametrizada por ϕ (parámetros variacionales):

Inferencia variacional: ¿Cómo elegir $q(\mathbf{z})$?

¿Qué pasa si la posterior $p(\mathbf{z} \mid \mathbf{x}; \theta)$ es intratable?

- Sea $q(\mathbf{z}; \phi)$ una pdf (tratable) parametrizada por ϕ (parámetros variacionales):
  E.g., gaussiana con media y covarianza especificadas por ϕ, $q(\mathbf{z}; \phi) = \mathcal{N}(\mathbf{z}; \phi_1, \phi_2)$

Inferencia variacional: ¿Cómo elegir $q(\mathbf{z})$?

¿Qué pasa si la posterior $p(\mathbf{z} \mid \mathbf{x}; \theta)$ es intratable?

- Sea $q(\mathbf{z}; \phi)$ una pdf (tratable) parametrizada por ϕ (parámetros variacionales):
  E.g., gaussiana con media y covarianza especificadas por ϕ, $q(\mathbf{z}; \phi) = \mathcal{N}(\mathbf{z}; \phi_1, \phi_2)$

- **Inferencia variacional:**
  Optimizar ϕ para que $q(\mathbf{z}; \phi)$ se acerque lo más posible a $p(\mathbf{z} \mid \mathbf{x}; \theta)$.

Inferencia variacional: ¿Cómo elegir $q(\mathbf{z})$?

¿Qué pasa si la posterior $p(\mathbf{z} \mid \mathbf{x}; \theta)$ es intratable?

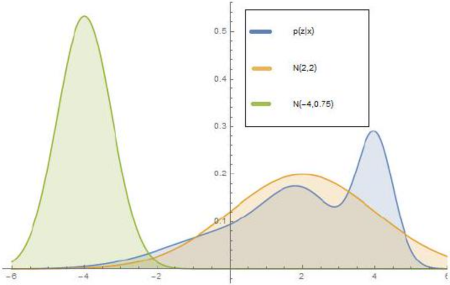
- Sea $q(\mathbf{z}; \phi)$ una pdf (tratable) parametrizada por ϕ (parámetros variacionales):
  E.g., gaussiana con media y covarianza especificadas por ϕ, $q(\mathbf{z}; \phi) = \mathcal{N}(\mathbf{z}; \phi_1, \phi_2)$

- **Inferencia variacional:**

  Optimizar ϕ para que $q(\mathbf{z}; \phi)$ se acerque lo más posible a $p(\mathbf{z} \mid \mathbf{x}; \theta)$.

- Ejemplo:

  La posterior $p(\mathbf{z} \mid \mathbf{x}; \theta)$ (azul) es mejor aproximada por $\mathcal{N}(\mathbf{z}; 2, 2)$ (naranja) que por $\mathcal{N}(\mathbf{z}; -4, 0.75)$ (verde).

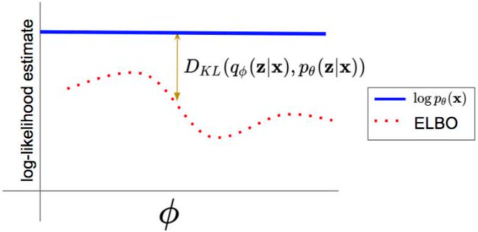


La cota inferior de evidencia

$$\underbrace{\mathcal{L}(\mathbf{x};\theta,\phi)}_{\text{ELBO}} := \sum_{\mathbf{z}} q(\mathbf{z};\phi)\log p(\mathbf{z},\mathbf{x};\theta) + H(q(\mathbf{z};\phi))$$

$$\log p(\mathbf{x};\theta) = \mathcal{L}(\mathbf{x};\theta,\phi) + D_{KL}(q(\mathbf{z};\phi)\|p(\mathbf{z}\mid\mathbf{x};\theta))$$

$$\log p(\mathbf{x};\theta) \geqslant \mathcal{L}(\mathbf{x};\theta,\phi)$$

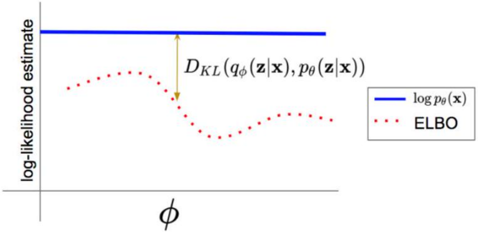


La cota inferior de evidencia

$$\underbrace{\mathcal{L}(\mathbf{x};\theta,\phi)}_{\text{ELBO}} := \sum_{\mathbf{z}} q(\mathbf{z};\phi)\log p(\mathbf{z},\mathbf{x};\theta) + H(q(\mathbf{z};\phi))$$

$$\log p(\mathbf{x};\theta) = \mathcal{L}(\mathbf{x};\theta,\phi) + D_{KL}(q(\mathbf{z};\phi)\|p(\mathbf{z}\mid\mathbf{x};\theta))$$

$$\log p(\mathbf{x};\theta) \geqslant \mathcal{L}(\mathbf{x};\theta,\phi)$$



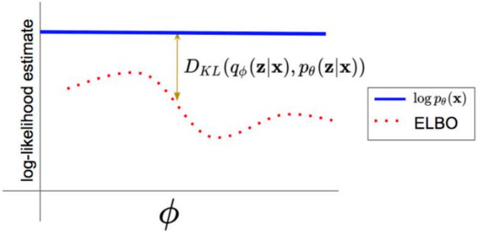
- Cuanto mejor $q(\mathbf{z};\phi)$ aproxime el *posterior* $p(\mathbf{z}\mid\mathbf{x};\theta)$:

La cota inferior de evidencia

$$\underbrace{\mathcal{L}(\mathbf{x};\theta,\phi)}_{\text{ELBO}} := \sum_{\mathbf{z}} q(\mathbf{z};\phi)\log p(\mathbf{z},\mathbf{x};\theta) + H(q(\mathbf{z};\phi))$$

$$\log p(\mathbf{x};\theta) = \mathcal{L}(\mathbf{x};\theta,\phi) + D_{KL}(q(\mathbf{z};\phi)\|p(\mathbf{z}\mid\mathbf{x};\theta))$$

$$\log p(\mathbf{x};\theta) \geqslant \mathcal{L}(\mathbf{x};\theta,\phi)$$



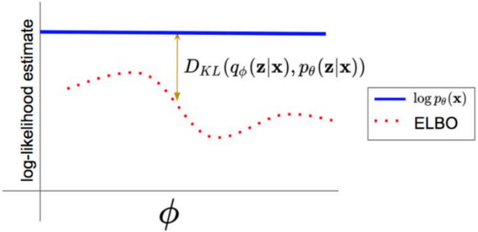
- Cuanto mejor $q(\mathbf{z};\phi)$ aproxime el *posterior* $p(\mathbf{z}\mid\mathbf{x};\theta)$:
  - Menor será $D_{KL}(q(\mathbf{z};\phi)\|p(\mathbf{z}\mid\mathbf{x};\theta))$,

La cota inferior de evidencia

$$\underbrace{\mathcal{L}(\mathbf{x};\theta,\phi)}_{\text{ELBO}} := \sum_{\mathbf{z}} q(\mathbf{z};\phi)\log p(\mathbf{z},\mathbf{x};\theta) + H(q(\mathbf{z};\phi))$$

$$\log p(\mathbf{x};\theta) = \mathcal{L}(\mathbf{x};\theta,\phi) + D_{KL}(q(\mathbf{z};\phi)\|p(\mathbf{z}\mid\mathbf{x};\theta))$$

$$\log p(\mathbf{x};\theta) \geqslant \mathcal{L}(\mathbf{x};\theta,\phi)$$



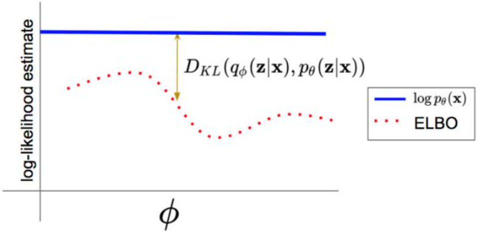
- Cuanto mejor $q(\mathbf{z};\phi)$ aproxime el *posterior* $p(\mathbf{z}\mid\mathbf{x};\theta)$:
  - Menor será $D_{KL}(q(\mathbf{z};\phi)\|p(\mathbf{z}\mid\mathbf{x};\theta))$,
  - Más cerca estará la ELBO de $\log p(\mathbf{x};\theta)$.

La cota inferior de evidencia

$$\underbrace{\mathcal{L}(\mathbf{x};\theta,\phi)}_{\text{ELBO}} := \sum_{\mathbf{z}} q(\mathbf{z};\phi)\log p(\mathbf{z},\mathbf{x};\theta) + H(q(\mathbf{z};\phi))$$

$$\log p(\mathbf{x};\theta) = \mathcal{L}(\mathbf{x};\theta,\phi) + D_{KL}(q(\mathbf{z};\phi)\|p(\mathbf{z}\mid\mathbf{x};\theta))$$

$$\log p(\mathbf{x};\theta) \geqslant \mathcal{L}(\mathbf{x};\theta,\phi)$$



- Cuanto mejor $q(\mathbf{z};\phi)$ aproxime el *posterior* $p(\mathbf{z}\mid\mathbf{x};\theta)$:
  - Menor será $D_{KL}(q(\mathbf{z};\phi)\|p(\mathbf{z}\mid\mathbf{x};\theta))$,
  - Más cerca estará la ELBO de $\log p(\mathbf{x};\theta)$.
- **Siguiente paso:** optimizar conjuntamente sobre θ y ϕ para maximizar la ELBO sobre un conjunto de datos, i.e.:

$$\max_{\theta,\phi}\sum_{n=1}^{N}\mathcal{L}(\mathbf{x}_n;\theta,\phi).$$

Resumen

Resumen

- **Ventajas de los modelos en variables latentes:**
  - Fáciles de construir modelos flexibles
  - Adecuados para aprendizaje no supervisado

Resumen

- **Ventajas de los modelos en variables latentes:**
  - Fáciles de construir modelos flexibles
  - Adecuados para aprendizaje no supervisado
- **Desventajas de los modelos en variables latentes:**

Resumen

- **Ventajas de los modelos en variables latentes:**
  - Fáciles de construir modelos flexibles
  - Adecuados para aprendizaje no supervisado
- **Desventajas de los modelos en variables latentes:**
  - No son tratables para evaluar densidades exactas

Resumen

- **Ventajas de los modelos en variables latentes:**
  - Fáciles de construir modelos flexibles
  - Adecuados para aprendizaje no supervisado
- **Desventajas de los modelos en variables latentes:**
  - No son tratables para evaluar densidades exactas
  - No son tratables para entrenar maxmizando directamente la verosimilitud marginal

Resumen

- **Ventajas de los modelos en variables latentes:**
  - Fáciles de construir modelos flexibles
  - Adecuados para aprendizaje no supervisado
- **Desventajas de los modelos en variables latentes:**
  - No son tratables para evaluar densidades exactas
  - No son tratables para entrenar maxmizando directamente la verosimilitud marginal
- **En modelos en variables latentes, calcular $p(\mathbf{x}; \theta)$ para un x dado es complejo:**

Resumen

- **Ventajas de los modelos en variables latentes:**
  - Fáciles de construir modelos flexibles
  - Adecuados para aprendizaje no supervisado

- **Desventajas de los modelos en variables latentes:**
  - No son tratables para evaluar densidades exactas
  - No son tratables para entrenar maxmizando directamente la verosimilitud marginal

- **En modelos en variables latentes, calcular $p(\mathbf{x};\theta)$ para un x dado es complejo:**
  - La inferencia variacional produce una cota ajustada en $\log p(\mathbf{x};\theta)$

$$\log p(\mathbf{x};\theta) \geqslant \sum_{\mathbf{z}} q(\mathbf{z}) \log p_\theta(\mathbf{x},\mathbf{z}) + H(q)$$

  para algún $q(\mathbf{z}) \approx p(\mathbf{z} \mid \mathbf{x})$ bueno encontrado por optimización.

Resumen

- **Ventajas de los modelos en variables latentes:**
  - Fáciles de construir modelos flexibles
  - Adecuados para aprendizaje no supervisado
- **Desventajas de los modelos en variables latentes:**
  - No son tratables para evaluar densidades exactas
  - No son tratables para entrenar maxmizando directamente la verosimilitud marginal
- **En modelos en variables latentes, calcular $p(\mathbf{x};\theta)$ para un $\mathbf{x}$ dado es complejo:**
  - La inferencia variacional produce una cota ajustada en $\log p(\mathbf{x};\theta)$

$$\log p(\mathbf{x};\theta) \geqslant \sum_{\mathbf{z}} q(\mathbf{z}) \log p_\theta(\mathbf{x},\mathbf{z}) + H(q)$$

    para algún $q(\mathbf{z}) \approx p(\mathbf{z} \mid \mathbf{x})$ bueno encontrado por optimización.
  - Esta es una aproximación compleja que solo vale para un $\mathbf{x}$.

Resumen

- **Ventajas de los modelos en variables latentes:**
  - Fáciles de construir modelos flexibles
  - Adecuados para aprendizaje no supervisado
- **Desventajas de los modelos en variables latentes:**
  - No son tratables para evaluar densidades exactas
  - No son tratables para entrenar maxmizando directamente la verosimilitud marginal
- **En modelos en variables latentes, calcular $p(\mathbf{x};\theta)$ para un $\mathbf{x}$ dado es complejo:**
  - La inferencia variacional produce una cota ajustada en $\log p(\mathbf{x};\theta)$

$$\log p(\mathbf{x};\theta) \geqslant \sum_{\mathbf{z}} q(\mathbf{z}) \log p_\theta(\mathbf{x},\mathbf{z}) + H(q)$$

    para algún $q(\mathbf{z}) \approx p(\mathbf{z} \mid \mathbf{x})$ bueno encontrado por optimización.
  - Esta es una aproximación compleja que solo vale para un $\mathbf{x}$.
- **Próximo paso:** como escalar este procedimiento a grandes datasets de $\mathbf{x}$.

Referencias

C. M. Bishop, *Pattern Recognition and Machine Learning*.
Springer, 2006.

Stanford, "CS236 Deep Generative Models." https://deepgenerativemodels.github.ioLecture, 2024.

J. M. Tomczak, *Deep Generative Modeling*.
Springer Cham, 2024.