

1. Fundamentos estadísticos

1.1. Teorema de Bayes

Este teorema es quien relaciona la probabilidad de un suceso dado otro a partir de cierta información:

$$p(z|x) = \frac{p(x|z) \cdot p(z)}{p(x)}$$

Donde,

$$\begin{aligned} p(z|x) &= \text{posterior} \\ p(x|z) &= \text{verosimilitud} \\ p(z) &= \text{prior} \\ p(x) &= \text{evidencia} \end{aligned}$$

Así, cada probabilidad está asociada a una distribución con cierto significado cualitativo (en el caso imágenes):

- **Distribución posterior $p(z|x)$:** Es la probabilidad de que un vector latente haya sido generado por una imagen en particular.
- **Distribución de verosimilitud $p(x|z)$:** Es la probabilidad de que dado un vector latente, una imagen x haya sido reconstruida a partir de él.
- **Distribución prior $p(z)$:** Es la creencia sobre la distribución latente inicial de la variable z sin haber visto aún los datos x . Generalmente, por simplicidad, se elige una distribución Gaussiana estándar ($\mathcal{N}(0, I)$).
- **Distribución de los datos o evidencia $p(x)$:** Es el término más complejo de tratar, es la probabilidad de los datos originales y se calcula de la siguiente manera:

$$\begin{aligned} p(x) &= \int p(x, z) dz \\ p(x, z) &= p(x|z)p(z) \\ p(x) &= \int p(x|z)p(z) dz \end{aligned}$$

Por la forma del espacio latente (alta dimensionalidad), esta integral se vuelve imposible de resolver, ya sea de forma analítica o numérica.

1.2. Inferencia y aproximaciones

Inferencia Variacional

Es una aproximación que convierte el problema de calcular la distribución posterior en un problema de optimización. Se introduce una distribución conocida (gaussiana) para aproximarla a la posterior, es decir, se busca:

$$p(z|x) \approx q(z|x) \quad (\text{término introducido})$$

Así, se busca conocer los parámetros de $q(z|x)$ que minimicen la diferencia entre estas dos distribuciones. Sin embargo, no se puede comparar directamente estas dos distribuciones, por lo que se manipula matemáticamente para llegar a la aproximación.

Divergencia de Kullback-Leibler (KL Divergence)

Esta medida indica cuánta información se pierde cuando se usa la distribución $q(z)$ para aproximar la distribución $p(z)$. También se puede decir que mide qué tanto se parece una distribución $p(z)$ a una distribución $q(z)$. Está dada por la siguiente ecuación:

$$\mathbb{D}_{\text{KL}}(q||p) = \sum_z q(z) \log \left(\frac{q(z)}{p(z)} \right)$$

Cuenta con ciertas características, como las siguientes:

- $\mathbb{D}_{\text{KL}} \geq 0$
- $\mathbb{D}_{\text{KL}} = 0$, si y solo si $p(z) = q(z)$
- $\mathbb{D}_{\text{KL}}(q||p) \neq \mathbb{D}_{\text{KL}}(p||q)$, no es simétrica.

Límite inferior de evidencia (ELBO)

Como se mencionó, no se puede comparar directamente $p(z|x)$ con $q(z|x)$, por lo que se utiliza este límite inferior para llegar a la aproximación y minimizar la diferencia entre ambas distribuciones. El tratamiento matemático es el siguiente:

Primero, se parte del término de evidencia y se aplica logaritmo a ambos lados de la ecuación:

$$p(x) = \int p(x|z)p(z)dz$$

$$\log(p(x)) = \log \left(\int p(x,z)dz \right)$$

Lo siguiente es multiplicar por $1 = \frac{q(z|x)}{q(z|x)}$, para incluirlo dentro de la integral.

$$\log(p(x)) = \log \left(\int p(x|z)p(z) \frac{q(z|x)}{q(z|x)} dz \right)$$

Ahora, recordando la definición del valor esperado:

$$\mathbb{E}_q[f(z)] = \int q(z)f(z)dz$$

Se puede reorganizar la expresión anterior para llegar a esta definición:

$$\log(p(x)) = \log \left(\int q(z|x) \frac{p(x|z)p(z)}{q(z|x)} dz \right)$$

$$f(z, x) = \frac{p(x|z)p(z)}{q(z|x)}$$

$$\log(p(x)) = \log \left(\int q(z|x)f(z, x)dz \right)$$

Así, se llega a:

$$\log(p(x)) = \log (\mathbb{E}_{q(z|x)} [f(z, x)])$$

$$\log(p(x)) = \log \left(\mathbb{E}_{q(z|x)} \left[\frac{p(x|z)p(z)}{q(z|x)} \right] \right)$$

Lo siguiente es aplicar la desigualdad de Jensen, la cual dice que:

$$f(\mathbb{E}[x]) \geq \mathbb{E}[f(x)]$$

Por tanto,

$$\log \left(\mathbb{E}_{q(z|x)} \left[\frac{p(x|z)p(z)}{q(z|x)} \right] \right) \geq \mathbb{E}_{q(z|x)} \left[\log \left(\frac{p(x|z)p(z)}{q(z|x)} \right) \right]$$

Reemplazando en la anterior ecuación:

$$\log(p(x)) \geq \mathbb{E}_{q(z|x)} \left[\log \left(\frac{p(x|z)p(z)}{q(z|x)} \right) \right]$$

Aplicando propiedades del logaritmo:

$$\log(p(x)) \geq \mathbb{E}_{q(z|x)} [\log(p(x|z))] + \mathbb{E}_{q(z|x)} \left[\log \left(\frac{p(z)}{q(z|x)} \right) \right] = \mathcal{L}_{ELBO}$$

Así, se llega a:

$$\mathcal{L}_{ELBO} = \mathbb{E}_{q(z|x)} [\log(p(x|z))] + \mathbb{E}_{q(z|x)} \left[\log \left(\frac{p(z)}{q(z|x)} \right) \right]$$

Tratando el segundo término,

$$\mathbb{E}_{q(z|x)} \left[\log \left(\frac{p(z)}{q(z|x)} \right) \right] = -\mathbb{E}_{q(z|x)} \left[\log \left(\frac{q(z|x)}{p(z)} \right) \right] = -\mathbb{D}_{KL}((q(z|x)||p(z)))$$

Se deja el signo negativo en la divergencia KL porque se busca maximizar el ELBO reduciendo la diferencia entre las distribuciones $q(z|x)$ y $p(z)$, entonces, entre más parecidas sean estas dos distribuciones, el término va a tender a 0, maximizando así el ELBO. Así, se queda con:

$$\mathcal{L}_{ELBO} = \mathbb{E}_{q(z|x)} [\log(p(x|z))] - \mathbb{D}_{KL}((q(z|x)||p(z)))$$

Donde,

$$\begin{aligned} \mathbb{E}_{q(z|x)} [\log(p(x|z))]: & \text{ Verosimilitud esperada / término de reconstrucción} \\ \mathbb{D}_{KL}((q(z|x)||p(z))): & \text{ Divergencia KL / término de regularización} \end{aligned}$$

Ya con esto, a la hora de implementarlo al modelo para el entrenamiento, maximizar el ELBO es equivalente a minimizar su negativo, siendo este último la función de pérdida del modelo:

$$\mathcal{L}_{VAE} = -\mathcal{L}_{ELBO} = -\mathbb{E}_{q(z|x)} [\log(p(x|z))] + \mathbb{D}_{KL}((q(z|x)||p(z)))$$

Aquí, cada término cumple cierta función:

- **Término de reconstrucción:** Mide que tan bien el decodificador reconstruye x a partir de z . Se busca que este término disminuya.
- **Término de regularización:** Fuerza a la distribución latente a seguir la distribución prior que se asumió (gaussiana), generando así un espacio latente continuo y uniforme.

En algunos casos, de ser necesario, se puede modificar el peso o importancia de cada término añadiendo un hiperparámetro β que escala la divergencia KL:

$$\mathcal{L}_{VAE} = -\mathbb{E}_{q(z|x)} [\log(p(x|z))] + \beta \cdot \mathbb{D}_{KL}((q(z|x)||p(z)))$$

Así,

- Para $\beta > 1$, se le da más importancia a la KL divergence.
- Para $\beta < 1$, se le da más importancia al término de reconstrucción.

1.3. Aclaración de conceptos

Estimación: Puntual vs Variacional

Puntual	Variacional
Tiene como objetivo encontrar un único valor (un punto en el espacio latente) que sea la mejor estimación del dato desconocido.	Tiene como objetivo encontrar una distribución de probabilidad completa que se asemeje a la distribución de los datos originales

Cuadro 1: Estimación Puntual vs Variacional

Método: Clásico vs Variacional

	Clásico	Variacional
Forma de z	Determinista: z es punto en el espacio latente.	Probabilística: z es una muestra de la distribución de probabilidad $q(z)$.
Función del codificador	Mapear la entrada x hacia un punto latente z .	Mapear la entrada x a los parámetros (media y desv. estándar) de la distribución de probabilidad $q(z)$.
Función de pérdida	Pérdida de reconstrucción	-ELBO
Capacidad generativa	Mala. Al ser determinista, el espacio latente es poco uniforme, haciendo que, al interpolar puntos, se llegue a resultados incoherentes.	Buena. El término de regularización permite tener un espacio latente continuo y suave, produciendo así nuevas muestras coherentes.

Cuadro 2: Método: Clásico vs Variacional

Regularización

La regularización es una término de penalización que se incluye para evitar sobreajuste, espacios latentes incoherentes y asegurar la generalización del modelo con espacios latentes continuos y suaves. Así, se logra que el modelo sea capaz de generar nuevas muestras coherentes.