Modelos Generativos

Nahuel Costa

Grado en Ciencia e Ingeniería de Datos

Universidad de Oviedo
Universidá d'Uviéu
University of Oviedo

Probabilidad

Existen dos interpretaciones diferentes de la probabilidad

- Frecuentista: las probabilidades representan frecuencias (a largo plazo) de sucesos que pueden ocurrir varias veces
- Bayesiana: la probabilidad se utiliza para cuantificar la incertidumbre/ignorancia sobre algo

Probabilidad

La interpretación bayesiana puede utilizarse para modelar la incertidumbre sobre sucesos puntuales que no tienen frecuencias a largo plazo.
Por ejemplo, podríamos querer calcular la probabilidad de que el casquete polar se derrita en 2030. Este suceso ocurrirá cero o una vez, pero no puede ocurrir repetidamente. No obstante, deberíamos ser capaces de cuantificar la incertidumbre sobre este suceso; en función de lo probable que creamos que es este suceso, podremos decidir cómo tomar la acción óptima.

Probabilidad

De aquí en adelante, adoptaremos la interpretación bayesiana. Afortunadamente, las reglas básicas de la teoría de la probabilidad son las mismas, independientemente de la interpretación que se adopte.

Probabilidad

La incertidumbre en las predicciones puede surgir fundamentalmente por dos razones:

- Epistemic uncertainty: se debe a la ignorancia sobre las causas subyacentes de los datos o sobre el mecanismo que los genera.
- Aleatoric uncertainty: surge de la variabilidad intrínseca de los datos (no puede reducirse aunque se recopilen más muestras)

Probabilidad de un evento

Definimos un evento A, como un estado que se cumple o no. Por ejemplo, A puede ser el suceso "lloverá mañana", o "llovió ayer". La expresión $Pr(A)$ denota la probabilidad con la que se cree que el suceso A es cierto. Exigimos que $0 \leq Pr(A) \leq 1$, donde $Pr(A) = 0$ significa que el suceso no ocurrirá, y $Pr(A) = 1$ significa que el suceso sí ocurrirá.

Variables random

Supongamos que X representa alguna cantidad desconocida, como la cara en la que caerá un dado al lanzarlo. Si el valor de X es desconocido y/o puede cambiar, lo llamamos variable aleatoria. El conjunto de valores posibles, denotado X, se conoce como espacio muestral o espacio de estados. Un suceso es un conjunto de resultados de un espacio muestral determinado.
Por ejemplo, si X representa la cara de un dado, por lo que $X = 1, 2, ..., 6$, el suceso de "ver un 1" se denota $X = 1$, el suceso de "ver un número impar" se denota $X \in 1, 3, 6$, el suceso de "ver un número entre 1 y 3" se denota $1 \leq X \leq 3$, etc.

Variables random discretas

Si el espacio muestral X es finito o contablemente infinito, X se denomina variable aleatoria discreta.

Definimos la **probability mass function** o **pmf** como una función que calcula la probabilidad de sucesos que corresponden a poner la variable random en cada valor posible: $p(x) = Pr(X = x)$

Variables random continuas

Si $X \in \mathbb{R}$ (es un valor real), se denomina variable aleatoria continua. En este caso, ya no podemos crear un conjunto finito de los distintos valores posibles que puede tomar. Sin embargo, existe un número contable de intervalos en los que se puede dividir.
En general, definimos la **cumulative distribution function** o **cdf** de X de la siguiente manera: $P(x) = Pr(X \leq x)$
Nótese que se usa una P mayúscula para representar la cdf
Usando esto, se puede calcular la probabilidad de estar en cualquier intervalo de la siguiente manera: $Pr(a < X \leq b) = P(b) - P(a)$

Variables random continuas

Definimos la **probability density function** o **pdf** como la derivada de la **cdf**:
$p(x) = \frac{d}{dx}P(x)$

Dada una pdf, podemos calcular la probabilidad de que una variable continua se encuentre en un intervalo finito como: $Pr(a < X \leq b) = \int_a^b p(x)d(x) = P(b) - P(a)$

Estadísticos

- Media (o valor esperado): es la propiedad más conocida de una distribución, a menudo denotado por μ
- Varianza: es una medida de la dispersión de una distribución, a menudo denotada por σ^2
- Desviación estándar: es la raíz de la varianza, σ

Inferencia Bayesiana

El término “inferencia“ hace referencia al acto de generalizar a partir de datos de muestra, normalmente con cierto grado de confianza.

El término “bayesiano“ se utiliza para referirse a los métodos de inferencia que representan esa confianza utilizando la teoría de la probabilidad y el teorema de Bayes.

Inferencia Bayesiana

El teorema de Bayes es una fórmula para calcular la distribución de probabilidad sobre posibles valores de una cantidad desconocida H dados unos datos observados $Y = y$:

$$p(H = h|Y = y) = \frac{P(Y = y|H = h) \cdot P(H = h)}{P(Y = y)}$$

Inferencia Bayesiana

El término $p(H)$ representa lo que se conoce sobre los posibles valores de H antes de ver ningún dato, esto es la **distribución a priori**.

El término $P(Y|H = h)$ representa la distribución sobre los posibles resultados Y que esperamos ver si $H = h$; esto es la **distribución de observación**.

Si la evaluamos en un punto correspondiente a las observaciones reales, y, obtenemos la función $p(Y = y|H = h)$, que se denomina **likelihood**.

Inferencia Bayesiana

Multiplicando la distribución a priori $p(H = h)$ por la función de likelihood $p(Y = y|H = h)$ para cada h se obtiene la **unnormalized joint distribution**, $p(H = h; Y = y)$. Se puede convertir en una distribución normalizada dividiendo por $p(Y = y)$, es lo que se conoce como marginal likelihood (veremos esto más adelante).

Al normalizar la joint distribution calculando $p(H = h|Y = y) = p(Y = y)$ para cada h, se obtiene la **posterior distribution**, $p(H = h|Y = y)$, que representa lo que se sabe de la distribución después de ver evidencia. En otras palabras:
$posterior \propto prior \times likelihood$

Inferencia Bayesiana

Un reto del enfoque bayesiano es que requiere especificar una distribución a priori, lo que puede resultar difícil, o incluso una limitación en modelos de gran tamaño, como las redes neuronales.

Veremos ejemplos de selección de distribuciones a priori y diferentes aproximaciones como **conjugate priors**, **uninformative priors**, **hierarchical priors**, útiles cuando hay varios datasets relacionados o las **empirical priors**, que pueden aprenderse a partir de los datos.

Distribuciones de probabilidad

Existe una gran variedad de distribuciones de probabilidad que se utilizan para distintos tipos de modelos. En este enlace se pueden observar algunas de las más utilizadas de forma interactiva.

Un modelo generativo es una distribución de probabilidad conjunta $p(x)$, para $x \in X$.

Modelos generativos

Existen muchos tipos de modelos generativos. A alto nivel podemos distinguir entre modelos generativos profundos, DGM (Deep Generative Models), basados en redes neuronales profundas que aprenden a mapear los datos observados x a una representación en un vector latente z; y modelos gráficos probabilísticos, PGM (Probabilistic Graphical Models) más "clásicos", que asignan un conjunto de variables latentes interconectadas $z_1, ..., z_L$ a las variables observadas $x_1, ..., x_D$ mediante mapeos más sencillos, a menudo lineales. Por supuesto son posibles muchos híbridos. De aquí en adelante nos centraremos en los DGM.

Modelos generativos

Los principales tipos de DGM son: los Variational Autoencoders (VAE), los AutoRegressive Models, los normalizing flows, los diffusion models, los energy based models (EBM) y las redes generativas adversariales (GAN).

Se puede clasificar estos modelos en función de los siguientes criterios:

- Density: modelos que pueden estimar la función de densidad de probabilidad $p(x)$. Por ejemplo, las redes GAN modelan la distribución de los datos de forma implícita, por lo que no pueden.

Modelos generativos

- Sampling: modelos que pueden generar nuevas muestras a partir de la distribución modelada. Modelos como VAEs y GANs admiten un muestreo rápido, sin embargo, ARMs, modelos diffusion y normalizing flows son lentos para el muestreo.
- Training: ¿qué tipo de método se utiliza para estimar los parámetros? Para algunos modelos (como AR y flows), se puede realizar una estimación exacta de la maximum likelihood estimation (MLE). Para otros modelos, no es tan sencillo. Por ejemplo, en el caso de los VAE, se maximiza un límite inferior de la likelihood o en el caso de las GAN, que se utiliza un entrenamiento min-max, que puede ser inestable, no existe una función objetivo clara que controlar.

Modelos generativos

- Latents: ¿usa el modelo un vector latente z para generar x o no, y si es así, tiene el mismo tamaño que x o es una representación potencialmente comprimida? Por ejemplo, los ARM no utilizan representaciones latentes; los flows y diffusion sí, pero no son representaciones comprimidas.
- Architectura: ¿Qué tipo de red neuronal se puede utilizar? ¿Existen restricciones? En el caso de los flows, por ejemplo, sólo se puede utilizar redes neuronales invertibles en las que cada capa tenga un jacobiano manejable.

Modelos generativos