Universidad de San Andrés

Repaso de Probabilidad para ML

Clase Tutorial 1 - Aprendizaje Automático y Aprendizaje Profundo

Ingeniería en Inteligencia Artificial
Universidad de San Andrés

2do Semestre 2024

Introducción

Motivación

- En la mayoría de las aplicaciones de ML, debemos lidiar con incertidumbre:
  - Incertidumbre sistemática/epistémica $\rightarrow$ dataset finito.
  - Incertidumbre intrínseca/aleatoria/estocástica o ruido $\rightarrow$ solo observamos información parcial.
- Probabilidad $\rightarrow$ paradigma consistente para la cuantificación y manipulación de la incertidumbre.
- Nos permite hacer predicciones óptimas dada toda la información de la que disponemos, aunque esa información pueda ser incompleta o ambigua.

Espacio Muestral

- **Espacio muestral** (Ω): conjunto de todos los posibles resultados de un experimento aleatorio.
  - Ejemplo. Experimento: lanzo una moneda dos veces, entonces: $\Omega = \{(cara, ceca), (cara, cara), (ceca, cara), (ceca, ceca)\}$
- **Evento o suceso**: cualquier subconjunto del espacio muestral.
  - Ejemplo. Evento: "Sacar cara en el 1er lanzamiento": $A = \{(cara, ceca), (cara, cara)\}$
- Dos eventos son **mutuamente excluyentes**/disjuntos si solo puede ocurrir uno de los dos.
  - Ejemplo. Primer lanzamiento siendo cara y primer lanzamiento siendo ceca.
- Decimos que un conjunto de n eventos $A_1, \ldots, A_n$ es una **partición** del evento A si $A = A_1 \cup A_2 \cup \ldots \cup A_n$ y los eventos $A_1, \ldots, A_n$ son mutuamente excluyentes.

Axiomas de Probabilidad (Kolmogórov)

Una función P es una **función de probabilidad** si satisface las siguientes condiciones:

1. La probabilidad de cualquier evento A debe ser no negativo:
$$P(A) \geq 0$$
2. La suma de las probabilidades a través de todos los posibles eventos en el espacio de resultados debe ser 1 (es decir, uno de los eventos en el espacio de resultados debe ocurrir):
$$P(\Omega) = 1$$
3. Si $A_1, \ldots, A_n$ es una partición del evento A, entonces la probabilidad de que ocurra A es la suma de las probabilidades individuales:
$$P(A) = P(A_1) + P(A_2) + \ldots + P(A_n)$$

Reglas Fundamentales de Probabilidad

Regla de la adición

$$\boxed{P(A) = \sum_{B} P(A, B)}$$

Regla del producto

$$\boxed{P(A, B) = P(B|A) \cdot P(A) = P(A|B) \cdot P(B)}$$

Probabilidad Marginal, Conjunta y Condicional

- **Probabilidad marginal**: $P(A) \rightarrow$ *"probabilidad de A"*.
- **Probabilidad conjunta**: $P(A, B) \rightarrow$ *"probabilidad de A y B"*.
  - Los eventos A y B son **independientes** si:

$$P(A, B) = P(A) \cdot P(B)$$

- **Probabilidad condicional**: $P(A|B) \rightarrow$ *"probabilidad de A dado B"*.
  - Asumiendo que $P(B) > 0$, se define como:

$$\boxed{P(A|B) = \frac{P(A, B)}{P(B)}}$$

  - Los eventos A y B son **condicionalmente independientes** dado un evento C si:

$$P(A, B|C) = P(A|C) \cdot P(B|C)$$

Probabilidad Marginal, Conjunta y Condicional

Ejercicio

60% de los alumnos de ML aprueban el final y 45% aprueban tanto el parcial como el final. ¿Que porcentaje de alumnos aprobaron el parcial dado que aprobaron el final?

Teorema de Bayes

$$\boxed{P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}}$$

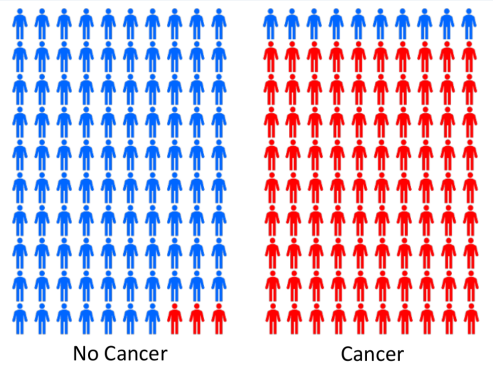
- Usando la regla de la suma, el denominador se puede expresar en términos de las cantidades del numerador:

$$P(B) = \sum_j P(B|A_j)P(A_j)$$

Teorema de Bayes

Ejemplo (cribado médico)

- *El 1% de la población tiene cáncer.*
  - $p(C = 1) = 1/100$
  - $p(C = 0) = 99/100$
- *Test: False positives: 3% y false negatives: 10%.*
  - $p(T = 1|C = 1) = 90/100$
  - $p(T = 0|C = 1) = 10/100$
  - $p(T = 1|C = 0) = 3/100$
  - $p(T = 0|C = 0) = 97/100$



Accuracy del test.
Izq: Cada 100 personas sin cáncer, 3 dan positivo.
Der: Cada 100 personas con cáncer, 10 dan negativo.

Teorema de Bayes

Ejemplo (cribado médico)

- $P(C = 1) = 1/100$
- $P(C = 0) = 99/100$
- $P(T = 1|C = 1) = 90/100$
- $P(T = 0|C = 1) = 10/100$
- $P(T = 1|C = 0) = 3/100$
- $P(T = 0|C = 0) = 97/100$

1. Si hacemos un cribado en la población, ¿cuál es la probabilidad de que alguien de positivo en la prueba?
2. Si alguien recibe un resultado positivo en la prueba, ¿cuál es la probabilidad de que realmente tenga cáncer?

Probabilidad a priori y a posteriori

Teorema de Bayes

$$P(C|T) = \frac{P(T|C) \cdot P(C)}{P(T)}$$

$$\text{Probabilidad posterior} = \frac{\text{Verosimilitud} \cdot \text{Probabilidad a priori}}{\text{Probabilidad de distribución de datos}}$$

- **Probabilidad a priori**: probabilidad disponible antes de observar el resultado.
  - En el ejemplo, antes de hacer un test la info más completa es $P(C) = 1\%$.
- **Probabilidad a posteriori**: probabilidad obtenida después de haber observado el resultado.
  - En el ejemplo, una vez que se sabe que una persona ha recibido un test positivo, por Bayes sabemos $P(C|T) = 23\%$.

Variables aleatorias

Variables aleatorias

- **Variable aleatoria** (X): es un función $X : \Omega \to \mathbb{R}$ que hace un mapeo entre el espacio de resultados y los números reales.
- X tiene asociada una **función de distribución**, definida como $F_X(x) = P(X \leq x)$
- **Variable aleatoria discreta (v.a.d.):** $X(\omega)$ toma valores en un conjunto discreto. Ejemplo. $X(\omega)$ es el número de caras que ocurren en una secuencia de lanzamientos de moneda ω.
  - Distribución definida por **función de probabilidad** $p_X(x) = P(X = x)$.
- **Variable aleatoria continua (v.a.c.):** $X(\omega)$ toma valores en un intervalo continuo. Ejemplo. $X(\omega)$ es la cantidad de tiempo que lleva una partícula radiactiva en desintegrarse.
  - Distribución definida por **función de densidad** $f_X(x) = \frac{dF_X(x)}{dx}$.

Esperanza

- Si X es una v.a.d. con función de masa de probabilidad $p_X(x)$ y $g : \mathbb{R} \to \mathbb{R}$ es una función arbitraria. En este caso, $g(X)$ puede considerarse una v.a., y definimos:

$$\mathbb{E}[g(X)] = \sum_x g(x) \cdot p_X(x)$$

- Si X es una v.a.c. con función de densidad $f_X(x)$, entonces definimos:

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) \cdot f(x)\, dx$$

- $\mathbb{E}[g(X)] \to$ *"promedio ponderado"* de los valores que $g(x)$ puede tomar para diferentes valores de x, con pesos $p_X(x)$ o $f_X(x)$.
- Si $g(x) = x \to \mathbb{E}[x]$ es la media de la variable aleatoria.

Esperanza

Propiedades:

- $\mathbb{E}[a] = a$ para cualquier constante $a \in \mathbb{R}$.
- $\mathbb{E}[af(X)] = a\mathbb{E}[f(X)]$ para cualquier constante $a \in \mathbb{R}$.
- Linealidad: $\mathbb{E}[f(X) + g(X)] = \mathbb{E}[f(X)] + \mathbb{E}[g(X)]$.
- Para un conjunto de n variables aleatorias: $\mathbb{E}[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} \mathbb{E}[X_i]$.
- Y si son independientes: $\mathbb{E}[\prod_{i=1}^{n} X_i] = \prod_{i=1}^{n} \mathbb{E}[X_i]$.

Varianza

- La varianza de una variable aleatoria X es una medida de qué tan concentrada está la distribución de X alrededor de su media $\mathbb{E}[X] = \mu$. Se define como:

$$Var[X] = \mathbb{E}[(X - \mathbb{E}[X])^2]$$

$$Var[X] = \mathbb{E}[X^2 - 2X\mathbb{E}[X] + (\mathbb{E}[X])^2]$$

$$Var[X] = \mathbb{E}[X^2 - (\mathbb{E}[X])^2]$$

$$\boxed{Var[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2}$$

- Propiedades:
  - $\text{Var}[a] = 0$ para cualquier constante $a \in \mathbb{R}$.
  - $\text{Var}[af(X)] = a^2\text{Var}[f(X)]$ para cualquier constante $a \in \mathbb{R}$.

Esperanza y varianza

Universidad de SanAndrés

Ejercicio

Calcule la media y la varianza de la variable aleatoria uniforme X con función de densidad de probabilidad $f_X(x) = 1$, *para todo x en el intervalo [0, 1], y 0 en otro lugar.*

Distribuciones discretas útiles

- $X \sim Bernoulli(p)$: Asociada a la ocurrencia o no de un éxito p. Ejemplo. Probabilidad de lanzar cara de una moneda (p=1).
  - $p_X(x) = p^x(1-p)^{1-x}$
  - $\mathbb{E}[X] = p$
  - $Var[X] = p(1-p)$

Distribuciones discretas útiles

- $X \sim Bernoulli(p)$: Asociada a la ocurrencia o no de un éxito p.
  Ejemplo. Probabilidad de lanzar cara de una moneda (p=1).
  - $p_X(x) = p^x(1-p)^{1-x}$
  - $\mathbb{E}[X] = p$
  - $Var[X] = p(1-p)$
- $X \sim Binomial(n, p)$: Asociada a la cantidad de éxitos en n ensayos.
  Ejemplo. Probabilidad de obtener cara en n lanzamientos.
  - $p_X(x) = \binom{n}{x} p^x(1-p)^{n-x}$
  - $\mathbb{E}[X] = np$
  - $Var[X] = np(1-p)$

Distribuciones discretas útiles

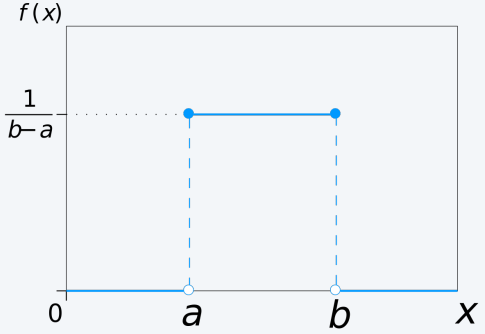
- $X \sim Bernoulli(p)$: Asociada a la ocurrencia o no de un éxito p.
  Ejemplo. Probabilidad de lanzar cara de una moneda (p=1).
  - $p_X(x) = p^x(1-p)^{1-x}$
  - $\mathbb{E}[X] = p$
  - $Var[X] = p(1-p)$
- $X \sim Binomial(n, p)$: Asociada a la cantidad de éxitos en n ensayos.
  Ejemplo. Probabilidad de obtener cara en n lanzamientos.
  - $p_X(x) = \binom{n}{x} p^x(1-p)^{n-x}$
  - $\mathbb{E}[X] = np$
  - $Var[X] = np(1-p)$
- $X \sim Geometrica(p)$: Asociada a la cantidad de ensayos que debo realizar hasta observar el 1^{er} éxito.
  Ejemplo. Cantidad de lanzamientos hasta observar la 1er cara.
  - $p_X(x) = (1-p)^{x-1}p$
  - $\mathbb{E}[X] = 1/p$
  - $Var[X] = (1-p)/p^2$

Distribuciones continuas útiles

Uniforme

$X \sim \mathcal{U}[a, b]$: Todos los puntos son equiprobables.

- $f_X(x) = \frac{1}{b-a}\mathbf{I}\{a \leq x \leq b\}$
- $\mathbb{E}[X] = (a+b)/2$
- $Var(X) = (b-a)^2/12$

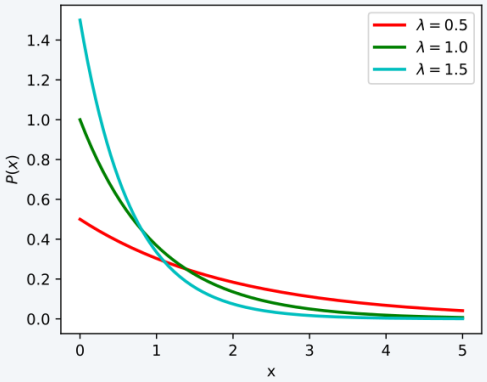


Distribuciones continuas útiles

Exponencial

$X \sim \mathcal{E}(\lambda)$: Modela tiempos hasta eventos que no tienen memoria.
Ejemplo. Cantidad de tiempo (que comienza ahora) hasta que se produzca un terremoto.

- $f_X(x) = \lambda e^{-\lambda x} \mathbf{I}\{x > 0\}$
- $\mathbb{E}[X] = 1/\lambda$
- $Var(X) = 1/\lambda^2$



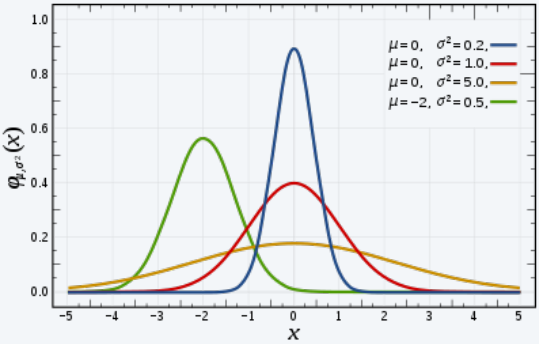
Distribuciones continuas útiles

Normal/Gaussiana

$X \sim \mathcal{N}(\mu, \sigma^2)$:

- $f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
- $\mathbb{E}[X] = \mu$
- $Var(X) = \sigma^2$
- Desvío estándar $= \sqrt{Var(X)} = \sigma$
- Precisión $= \beta = 1/Var(X) = 1/\sigma^2$

¿Por qué se utiliza tanto la distribución gaussiana?

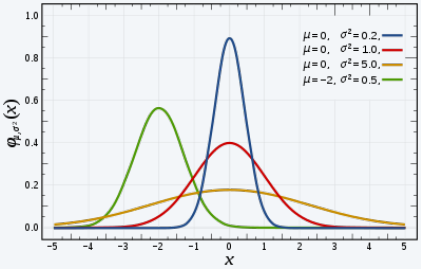


Distribuciones continuas útiles

Normal/Gaussiana

¿Por qué se utiliza tanto la distribución gaussiana?

- Sus parámetros son fáciles de interpretar.
- Teorema del Límite Central: suma de v.a. independientes tienen distribución aprox. gaussiana $\rightarrow$ modela "ruido".
- Hace el menor número de suposiciones (tiene la entropía máxima entre todas las distribuciones con la misma media y varianza).
- Forma matemática simple $\rightarrow$ métodos fáciles de implementar y altamente efectivos.



Función de verosimilitud

- Dataset de N observaciones de la variable escalar x: $\mathbf{x} = \{x_1, \ldots, x_n\}$.
- Observaciones se extraen de forma independiente (i.i.d.) de una distribución Gaussiana con μ y σ^2 desconocidas.
- Queremos determinar estos parámetros a partir del conjunto de datos.
  - El problema de estimar una distribución, dado un conjunto finito de observaciones, se conoce como **estimación de densidad**.
- Como $\mathbf{x}$ es i.i.d, podemos escribir la probabilidad del conjunto de datos, dado μ y σ^2 como:

$$p(\mathbf{x}|\mu, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu, \sigma^2).$$

- Cuando se ve como una función de μ y σ^2, esto se llama la **función de verosimilitud** para la Gaussiana.

Máxima verosimilitud

- **Máxima verosimilitud**: encontrar los valores de los parámetros que maximizan la función de verosimilitud.
- En la práctica, maximizamos el logaritmo de la función de verosimilitud.
- Maximizar $log(f(x))$ = Maximizar $f(x) \rightarrow$ logaritmo es una función monotónicamente creciente de su argumento.
  - ✓ Simplifica el análisis matemático subsiguiente.
  - ✓ Ayuda numéricamente $\rightarrow$ producto de muchas probabilidades chicas puede desbordar la precisión numérica, entonces computamos la suma de los logaritmos de las probabilidades.

$$\ln p(\mathbf{x}|\mu, \sigma^2) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi)$$

Máxima verosimilitud

Ejercicio

Demostrar que el estimador de máxima verosimilitud (MLE) para la media μ de una distribución Gaussiana es la media muestral de los datos observados.

Máxima verosimilitud

Ejercicio

Demostrar que el estimador de máxima verosimilitud (MLE) para la varianza σ^2 de una distribución Gaussiana es la varianza muestral de los datos observados.

Máxima verosimilitud

Universidad de San Andrés

En resumen

- El estimador de la varianza σ^{2*} depende del estimador de la media μ^* porque se realiza una maximización conjunta de la función de verosimilitud con respecto a μ y σ^2.
- En una distribución Gaussiana, esta maximización conjunta se descompone en dos pasos:
  - Primero, se estima μ.
  - Luego, se utiliza este estimador para determinar σ^2.

Sesgo del estimador de máxima verosimilitud

La técnica de máxima verosimilitud se utiliza ampliamente en DL y forma la base de la mayoría de los algoritmos de ML. Sin embargo, tiene ciertas limitaciones:

- Las soluciones de máxima verosimilitud μ^* y σ^{2*} son funciones de los valores del conjunto de datos $x_1, \ldots, x_N$.
- Supongamos que estos dos se generaron a partir de una distribución gaussiana con parámetros verdaderos μ y σ^2.
- Es sencillo demostrar que las esperanzas de μ^* y σ^{2*} con respecto a los valores del conjunto de datos son:

$$\mathbb{E}[\mu^*] = \mu \quad \text{y} \quad \mathbb{E}[\sigma^{2*}] = \left(\frac{N-1}{N}\right)\sigma^2$$

- La estimación de máxima verosimilitud de la varianza subestimará la varianza verdadera por un factor de $(N-1)/N \rightarrow$ **sesgo**.

Sesgo del estimador de máxima verosimilitud

Sesgo: el estimador de una cantidad aleatoria es sistemáticamente diferente del valor verdadero.

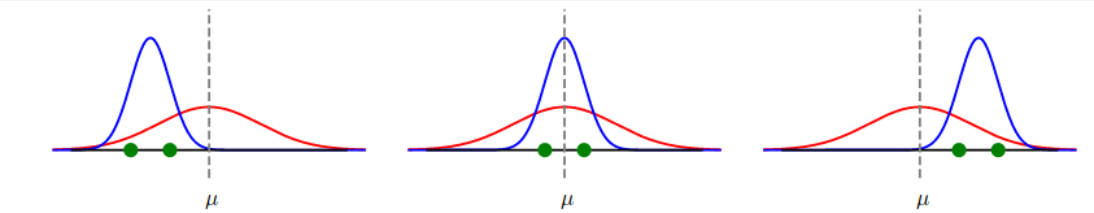


Figure 2.10 Illustration of how bias arises when using maximum likelihood to determine the mean and variance of a Gaussian. The red curves show the true Gaussian distribution from which data is generated, and the three blue curves show the Gaussian distributions obtained by fitting to three data sets, each consisting of two data points shown in green, using the maximum likelihood results (2.57) and (2.58). Averaged across the three data sets, the mean is correct, but the variance is systematically underestimated because it is measured relative to the sample mean and not relative to the true mean.

Dos variables aleatorias

Distribución conjunta y marginales

- Tenemos dos variables aleatorias X e Y. Si las consideramos por separado, necesitamos $F_X(x)$ y $F_Y(y)$.
- Si deseamos conocer sus valores simultáneos durante un experimento aleatorio $\rightarrow$ **función de distribución conjunta**: $F_{XY}(x, y) = P(X \leq x, Y \leq y)$
- Caso discreto:
  - **Función de probabilidad conjunta**: $p_{XY}(x, y) = P(X = x, Y = y)$
  - **Funciones de probabilidad marginales**: $p_X(x) = \sum_y P_{XY}(x, y) \mid p_Y(y) = \sum_x P_{XY}(x, y)$.
- Caso continuo:
  - **Función de densidad conjunta**: $f_{XY}(x, y) = \frac{\partial^2 F_{XY}(x,y)}{\partial x \partial y}$
  - **Funciones de densidad marginales**: $f_X(x) = \int_{\mathbf{R}} f_{XY}(x, y)\, dy \mid f_Y(y) = \int_{\mathbf{R}} f_{XY}(x, y)\, dx$

Distribuciones condicionales

- Buscan responder: ¿cuál es la distribución de probabilidad sobre Y cuando sabemos que X debe tomar un cierto valor x?
- Caso discreto:
  - **Función de probabilidad condicional**:

$$p_{Y|X}(y|x) = \frac{p_{XY}(x,y)}{p_X(x)} \text{ suponiendo } p_X(x) \neq 0$$

- Caso continuo:
  - **Función de densidad condicional**:

$$f_{Y|X}(y|x) = \frac{f_{XY}(x,y)}{f_X(x)} \text{ suponiendo } f_X(x) \neq 0$$

Teorema de Bayes

- Caso discreto:

$$P_{Y|X}(y|x) = \frac{P_{XY}(x,y)}{P_X(x)} = \frac{P_{X|Y}(x|y)P_Y(y)}{\sum_{y_0 \in \mathcal{Y}} P_{X|Y}(x|y_0)P_Y(y_0)}$$

- Caso continuo:

$$f_{Y|X}(y|x) = \frac{f_{XY}(x,y)}{f_X(x)} = \frac{f_{X|Y}(x|y)f_Y(y)}{\int_{-\infty}^{\infty} f_{X|Y}(x|y_0)f_Y(y_0)\,dy_0}$$

Independencia de variables aleatorias

- Diremos que dos v.a. X e Y son **independientes** si vale que

$$F_{XY}(x, y) = F_X(x)F_Y(y) \ \ \forall x, y \in \mathbb{R}$$

- Informalmente: X e Y son independientes si conoces toda la información sobre el par (X, Y) solo conociendo $f(x)$ y $f(y)$.
- Caso discreto:

$$p_{XY}(x, y) = p_X(x)p_Y(y) \ \ \forall x, y \in \mathbb{R}$$

- Caso continuo:

$$f_{XY}(x, y) = f_X(x)f_Y(y) \ \ \forall x, y \in \mathbb{R}$$

Esperanza

- Si X e Y son v.a.d. y $g : \mathbb{R}^2 \to \mathbb{R}$ es una función de esas dos v.a. Entonces definimos al valor esperado de g:

$$\mathbb{E}[g(X,Y)] = \sum_x \sum_y g(x,y) \cdot p_{XY}(x,y)$$

- Si X e Y son v.a.c. con función de densidad conjunta $f_{XY}(x,y)$, entonces definimos:

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) \cdot f_{XY}(x,y)\, dxdy$$

Covarianza

- Se puede usar el concepto de esperanza para estudiar la relación entre dos v.a. X e Y
- **Covarianza:** mide el grado en que X e Y están (linealmente) relacionadas. Mide el grado de variación conjunta de X e Y respecto a sus medias.

$$\begin{aligned} Cov[X,Y] &= E[(X - E[X])(Y - E[Y])] \\ &= E[XY - XE[Y] - YE[X] + E[X]E[Y]] \\ &= E[XY] - E[X]E[Y] - E[Y]E[X] + E[X]E[Y] \\ &= \boxed{E[XY] - E[X]E[Y]} \end{aligned}$$

- Interpretación:
  - $Cov[X,Y] > 0$: Dependencia directa/positiva. $(\uparrow X \rightarrow \uparrow Y)$
  - $Cov[X,Y] = 0$: No existencia de una relación lineal entre las dos variables.
  - $Cov[X,Y] < 0$: Dependencia inversa/negativa. $(\uparrow X \rightarrow \downarrow Y)$

Correlación

- A las v.a. cuya $Cov[X, Y] = 0$ se dicen que son **no correlacionadas**.

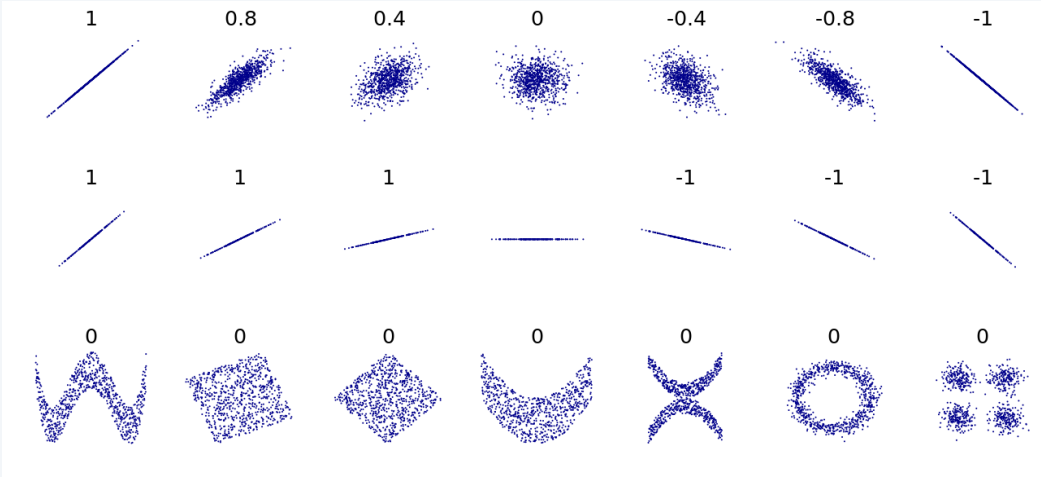
$$\boxed{\rho = corr[X, Y] = \frac{Cov[X, Y]}{\sigma_X \sigma_Y}}$$

- Si X e Y son v.a. **independientes** $\rightarrow Cov[X, Y] = 0$ dado que:
  - Si son independientes: $E[XY] = E[X]E[Y]$
  - Reemplazando en la fórmula de $Cov[X, Y]$:

$$\begin{aligned} Cov[X, Y] &= E[XY] - E[X]E[Y] \\ &= E[X]E[Y] - E[X]E[Y] \\ &= 0 \end{aligned}$$

- IMPORTANTE: lo opuesto (generalmente) no es cierto.
  Independencia implica incorrelación, pero incorrelación no implica independencia.

Correlación e independencia



Correlación e independencia

Ejercicio

Dado $X \sim$ Uniforme$(-1, 1)$ e $Y = X^2$, demuestre que X e Y no están correlacionadas si bien no son independientes.

Multiples variables aleatorias

Vectores aleatorios

- Podemos trabajar con n v.a. juntas en un **vector aleatorio** $\mathbf{X} = [X_1, X_2, \ldots, X_n]^T$ que mapea de $\Omega \to \mathbb{R}^n$.

Esperanza

- Si consideramos una función arbitraria $g : \mathbb{R}^n \to \mathbb{R}$, su esperanza es:

$$E[g(X)] = \int_{\mathbb{R}^n} g(x_1, \ldots, x_n) f_{X_1,\ldots X_n}(x_1, \ldots, x_n)\, dx_1 \ldots dx_n$$

- Si $g : \mathbb{R}^n \to \mathbb{R}^m$:

$$g(x) = \begin{pmatrix} g_1(x) \\ g_2(x) \\ \vdots \\ g_m(x) \end{pmatrix}, \text{ entonces } E[g(x)] = \begin{pmatrix} E[g_1(x)] \\ E[g_2(x)] \\ \vdots \\ E[g_m(x)] \end{pmatrix}$$

Vectores aleatorios

Matriz de covarianza

- Dado un vector aleatorio $X : \Omega \to \mathbb{R}^n$, su matríz de covarianza Σ es una matríz cuadrada de *nxn* cuyas entradas están dadas por $\Sigma_{ij} = Cov[X_i, X_j]$

$$\Sigma = \begin{bmatrix} Cov[X_1, X_1] & \cdots & Cov[X_1, X_n] \\ \vdots & \ddots & \vdots \\ Cov[X_n, X_1] & \cdots & Cov[X_n, X_n] \end{bmatrix} = \begin{bmatrix} E[X_1^2] - E[X_1]E[X_1] & \cdots & E[X_1X_n] - E[X_1]E[X_n] \\ \vdots & \ddots & \vdots \\ E[X_nX_1] - E[X_n]E[X_1] & \cdots & E[X_n^2] - E[X_n]E[X_n] \end{bmatrix}$$

$$= \begin{bmatrix} E[X_1^2] & \cdots & E[X_1X_n] \\ \vdots & \ddots & \vdots \\ E[X_nX_1] & \cdots & E[X_n^2] \end{bmatrix} - \begin{bmatrix} E[X_1]E[X_1] & \cdots & E[X_1]E[X_n] \\ \vdots & \ddots & \vdots \\ E[X_n]E[X_1] & \cdots & E[X_n]E[X_n] \end{bmatrix}$$

$$= E[XX^T] - E[X]E[X]^T = \ldots = E[(X - E[X])(X - E[X])^T]$$

- Propiedades:
  - $\Sigma \geq 0 \to \Sigma$ es semidefinida positiva.
  - $\Sigma = \Sigma^T \to \Sigma$ es simétrica.

Distribución Normal Multivariada

- Es una generalización de la distribución normal unidimensional a dimensiones superiores.
- Para un vector aleatorio X de dimensión D, $X \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$:

$$f_{X_1,X_2,\ldots,X_N}(x_1, x_2, \ldots, x_n|\mu, \Sigma) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

  - $\boldsymbol{\mu}$ es el vector de media de D dimensiones
  - $\boldsymbol{\Sigma}$ es la matriz de covarianza de $D \times D$
  - $|\boldsymbol{\Sigma}|$ es el determinante de $\boldsymbol{\Sigma}$

Distribución Normal Multivariada

Marginales y condicionales

- Supongamos $y = (y_1, y_2)$ es un vector aleatorio que sigue una distribución gaussiana multivariada con parámetros:

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}, \Lambda = \Sigma^{-1} = \begin{pmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{pmatrix}$$

  donde Λ es la matríz de precisión.
- Las marginales están dadas por:

$$p(y_1) = \mathcal{N}(y_1|\mu_1, \Sigma_{11})$$
$$p(y_2) = \mathcal{N}(y_2|\mu_2, \Sigma_{22})$$

- La probabilidad a posteriori está dada por:

$$p(y_1|y_2) = \mathcal{N}(y_1|\mu_{1|2}, \Sigma_{1|2})$$
$$\mu_1|2 = \mu_1 + \Sigma_{12}\Sigma_2^{-1}2(y_2 - \mu_2)$$
$$\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

Teoría de la información

Entropía

- Tenemos una v.a. X. ¿Cuánta información recibo al saber su valor específico?
- Cantidad de información $\rightarrow$ "grado de sorpresa" de conocer el valor de X.
  - Evento improbable (sorprendente): $\uparrow$ información.
  - Evento probable (no sorprendente): $\downarrow$ información.
- La "sorpresa" de X es una función que aumenta a medida que su probabilidad $p(X)$ disminuye:

$$\log\left(\frac{1}{p(X)}\right),$$

  que da 0 de sorpresa cuando la $p(X) = 1$.
- Por lo tanto, podemos definir la información, o sorpresa, de X como:

$$I(X) = -\log_2(p(X)) = \log_2\left(\frac{1}{p(X)}\right) [bits]$$

Entropía

- **Entropía:** valor medio de información transmitida por un evento, al considerar todos los resultados posibles.

$$H(X) = \mathbb{E}[I(X)] = \mathbb{E}[-\log p(x)] = -\sum p(x) \log_b p(x)$$

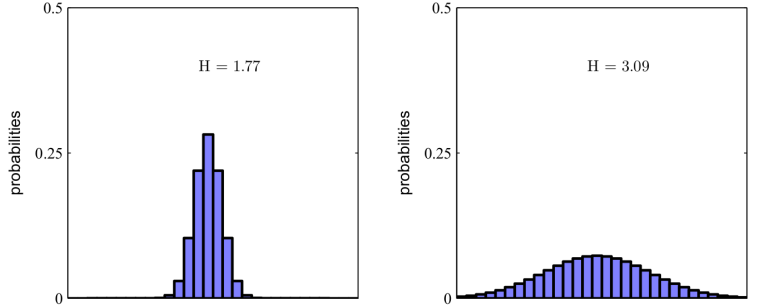


Figure 2.14 Histograms of two probability distributions over 30 bins illustrating the higher value of the entropy H for the broader distribution. The largest entropy would arise from a uniform distribution which would give $\mathrm{H} = -\ln(1/30) = 3.40$.

Cross-Entropy

- **Cross-entropy:** mide la diferencia entre dos distribuciones de probabilidad de un conjunto de eventos.
- El objetivo de muchas tareas de ML es construir una función de probabilidad sobre la variable de interés que se aproxime al máximo a la distribución real.
- Para saber la diferencia entre la distribución inferida q y la real p (conjunto de entrenamiento) se utiliza la medida de cross-entropy.
- A nivel matemático:

$$H(p, q) = -\sum_{x \in \mathcal{X}} p(x) \log q(x)$$

- Si la predicción es perfecta, $H(p) = H(p, q)$.
- Si difieren, la diferencia se llama **entropía relativa** o **Divergencia de Kullback-Leibler**:

$$KL(p||q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}$$

Referencias


Christopher M. Bishop y Hugh Bishop (2023)
Deep Learning - Foundations and Concepts
Springer Capítulos 2 y 3.


Kevin P. Murphy (2022)
Probabilistic Machine Learning
MIT Press Capítulo I, Secciones 2, 3 y 6.


Christopher M. Bishop (2006)
Pattern Recognition and Machine Learning
Springer Capítulos 1 y 2.

Universidad de San Andrés

Gracias por su atención

Contacto:
tmonreal@udesa.edu.ar