Distancias y divergencias entre medidas de probabilidad

Emilien Joly

References

[1] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. OUP Oxford, 2013.

[2] Luc Devroye, Abbas Mehrabian, and Tommy Reddad. The total variation distance between high-dimensional gaussians with the same mean. *arXiv preprint arXiv:1810.08693*, 2018.

[3] DC Dowson and BV Landau. The fréchet distance between multivariate normal distributions. *Journal of multivariate analysis*, 12(3):450–455, 1982.

[4] Elizabeth S Meckes and Mark W Meckes. On the equivalence of modes of convergence for log-concave measures. In *Geometric Aspects of Functional Analysis: Israel Seminar (GAFA) 2011-2013*, pages 385–394. Springer, 2014.

[5] C. Villani. *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 2008.

1 Definiciones y propiedades básicas

En nuestro contexto, consideramos medidas de probabilidad $\mu, \nu \in \mathcal{P}(\mathbb{R}^n)$.

Definition 1. *Sean $\mu, \nu \in \mathcal{P}(\mathbb{R}^n)$. Definimos las distancias*

- *La distancia en **variación total**:*

$$d_{TV}(\mu, \nu) = \sup_{A \in \mathcal{B}(\mathbb{R}^n)} |\mu(A) - \nu(A)|$$

- *La **distancia de Wasserstein**:*
  Para una distancia d sobre $\mathbb{R}^n$

$$W_p(\mu, \nu) = \inf_{\pi : \in \Gamma(\mu, \nu)} \left(\int_{\mathbb{R}^n \times \mathbb{R}^n} d^p(x, y) d\pi(x, y) \right)^{1/p}$$

  donde $\Gamma(\mu, \nu)$ corresponde a las medidas de probabilidad sobre $\mathbb{R}^n \times \mathbb{R}^n$ y de marginales dadas por μ y ν.

- *La **divergencia de Kullback y Leibler** para medidas tal que $\mu << \nu$:*

$$D(\mu||\nu) = \int_{\mathbb{R}^n} \log\left(\frac{d\mu}{d\nu}(x)\right) d\mu(x).$$

1.1 Propiedades de la variación total

Proposition 1. *Para dos medidas de probabilidad μ, ν,*

- *Si μ y ν son discretas sobre $\mathcal{X}$,*

$$\begin{aligned} d_{TV}(\mu,\nu) &= \frac{1}{2}\sum_{x\in\mathcal{X}} |\mu(\{x\}) - \nu(\{x\})| \\ &= 1 - \sum_{x\in\mathcal{X}} \mu(\{x\}) \wedge \nu(\{x\}) \end{aligned}$$

  Adicionalmente, si denotamos $A^* = \{x \in \mathcal{X} : \mu(\{x\}) \geq \nu(\{x\})\}$, entonces $d_{TV}(\mu,\nu) = \mu(A^*) - \nu(A^*)$.

- *Si μ y ν tienen densidades f y g (al respecto de la medida de Lebesgue λ),*

$$\begin{aligned} d_{TV}(\mu,\nu) &= \frac{1}{2}\int_{\mathbb{R}^n} |f(x) - g(x)| d\lambda(x) \\ &= 1 - \int_{\mathbb{R}^n} f(x) \wedge g(x) d\lambda(x) \end{aligned}$$

  Adicionalmente, si denotamos $A^* = \{x \in \mathcal{X} : f(x) \geq g(x)\}$, entonces $d_{TV}(\mu,\nu) = \mu(A^*) - \nu(A^*)$.

Proof. Hacemos la prueba en el caso discreto, el caso con densidad se trata de manera analoga. Para $A \in \mathcal{B}(\mathbb{R}^n)$,

$$\begin{aligned} \mu(A) - \nu(A) &= \underbrace{\mu(A\cap A^*) - \nu(A\cap A^*)}_{\geq 0} + \underbrace{(\mu(A\cap \overline{A^*}) - \nu(A\cap \overline{A^*}))}_{<0} \\ &\leq \mu(A\cap A^*) - \nu(A\cap A^*) \\ &= \sum_{x\in A\cap A^*} \mu(\{x\}) - \nu(\{x\}) \\ &\leq \sum_{x\in A^*} \mu(\{x\}) - \nu(\{x\}) \\ &\leq \mu(A^*) - \nu(A^*). \end{aligned}$$

De la misma manera, tenemos

$$\nu(A) - \mu(A) \leq \nu(\overline{A^*}) - \mu(\overline{A^*}) = \mu(A^*) - \nu(A^*)$$

lo que prueba $|\nu(A) - \mu(A)| \leq \mu(A^*) - \nu(A^*)$ con caso de igualdad si $A = A^*$ y eso demuestra la segunda parte. Para la primera, notamos que

$$\sum_{x\in\mathcal{X}} |\mu(\{x\}) - \nu(\{x\})| = \underbrace{\sum_{x\in A^*} \mu(\{x\}) - \nu(\{x\})}_{I} + \underbrace{\sum_{x\in\overline{A^*}} \nu(\{x\}) - \mu(\{x\})}_{II}$$

Tenemos $I - II = 1 - 1 = 0$ y $I = \mu(A^*) - \nu(A^*) = d_{TV}(\mu, \nu)$. Entonces,

$$\sum_{x\in\mathcal{X}} |\mu(\{x\}) - \nu(\{x\})| = I + II = 2d_{TV}(\mu, \nu)$$

En fin,

$$\begin{aligned}\sum_{x\in\mathcal{X}} \mu(\{x\}) \wedge \nu(\{x\}) &= \sum_{x\in\mathcal{X}} \frac{1}{2}(\mu(\{x\}) + \nu(\{x\}) - |\mu(\{x\}) - \nu(\{x\})|) \\ &= \frac{1}{2} + \frac{1}{2} - d_{TV}(\mu, \nu)\end{aligned}$$

□

Example 1. *Sean $\mu \sim U[0,1]$ y $\nu \sim U[0,2]$, $d_{TV}(\mu, \nu) = 1/2$*

Sean $\mu \sim \mathcal{N}(m, \Sigma_1)$ y $\nu \sim \mathcal{N}(m, \Sigma_2)$ con $m \in \mathbb{R}^d$ y $\Sigma_1, \Sigma_2 \in \mathcal{M}_{d\times d}(\mathbb{R})$. De manera interesante, no hay formulas cerradas para la distancia en variación total entre μ y ν. Hay cotas del estilo:

Example 2. *Sean $\lambda_1, \ldots, \lambda_d$ los valores propios de la matriz $\Sigma_1^{-1}\Sigma_2 - I_d$. Entonces,*

$$\frac{1}{100} \leq \frac{d_{TV}(\mu, \nu)}{\min\left(1, \sqrt{\sum_{i=1}^d \lambda_i^2}\right)} \leq \frac{3}{2}$$

Ver [2]

La topologia de la distancia en variación total es la misma que la topologia de la convergencia en distribución. Para ver eso:

Exercice 1. *Sea $\mathcal{F}$ el conjunto de funciones medibles $f : \mathbb{R}^n \to [-1, 1]$. Mostrar que*

$$d_{TV}(\mu, \nu) = \frac{1}{2} \sup_{f\in\mathcal{F}} \left| \int f d\mu - \int f d\nu \right|.$$

Mostrar que el supremo se alcanza para $f = \mathbb{1}_{A^*} - \mathbb{1}_{\overline{A^*}}$.

1.2 Ejemplos de distancia Wasserstein-2

Para simplificar el problema, es facil ver que es suficiente considerar distancias de Wasserstein de translaciones de X y Y de tal manera que las dos variables son centradas. En efecto,

$$\inf_{\pi} \mathbb{E}\left[\|X-Y\|^2\right] = \inf_{\pi} \mathbb{E}\left[\|X-\mathbb{E}[X]-(Y-\mathbb{E}[Y])\|^2\right] + \|\mathbb{E}[X]-\mathbb{E}[Y]\|_2^2$$

Para dos distribuciones gaussianas, hay una manera de calcular explicitamente la distancia de Wasserstein entre ellas.

Proposition 2 (Ver [3]). *Sean $\mu \sim \mathcal{N}(m_1, \Sigma_1)$ y $\nu \sim \mathcal{N}(m_2, \Sigma_2)$ dos medidas gaussianas de $\mathcal{P}(\mathbb{R}^n)$. Entonces,*

$$W_2(\mu,\nu)^2 = \|m_1-m_2\|_2^2 + \mathrm{Tr}(\Sigma_1+\Sigma_2-2(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2})$$

Denotamos $d(\Sigma_1,\Sigma_2)^2 = \mathrm{Tr}(\Sigma_1+\Sigma_2-2(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2})$ de tal manera que d es una distancia entre matrices symetricas.

Proof. El resultado es de hecho un poco más general. Podemos probar el

Lemma 1. *Sean dos variables X, Y de $\mathbb{R}^n$ centradas y de varianza Σ_X y Σ_Y. Entonces,*

$$d(\Sigma_X,\Sigma_Y)^2 \leq \mathbb{E}\left[\|X-Y\|^2\right] \leq \mathrm{Tr}(\Sigma_X+\Sigma_Y+2(\Sigma_X\Sigma_Y)^{1/2}).$$

Para demostrar el lema, consideramos el vector aleatorio

$$W = \begin{pmatrix} X \\ Y \end{pmatrix} \qquad \text{con} \quad \Sigma_W = \begin{pmatrix} \Sigma_X & V \\ V^T & \Sigma_Y \end{pmatrix}.$$

Entonces,

$$\begin{aligned}\mathbb{E}\left[\|X-Y\|^2\right] &= \mathbb{E}\left[\mathrm{Tr}\,(X-Y)^T(X-Y)\right] = \mathrm{Tr}\,\mathbb{E}\left[(X-Y)^T(X-Y)\right] \\ &= \mathrm{Tr}(\Sigma_X+\Sigma_Y-V-V^T)\end{aligned}$$

Por la descomposición en matrices de rango 1, podemos ver que $\Sigma_W = \sum_{i=1}^{2n} w_i w_i^T$ donde los w_i son vectores de $\mathbb{R}^n$. Escribiendo

$$w_i = \begin{pmatrix} a_i \\ b_i \end{pmatrix}$$

con vectores $a_i, b_i \in \mathbb{R}^n$ caracterizamos las matrices

$$\Sigma_X = \sum_{i=1}^{2n} a_i a_i^T \qquad \Sigma_Y = \sum_{i=1}^{2n} b_i b_i^T \qquad V = \sum_{i=1}^{2n} a_i b_i^T$$

El problema se reduce a encontrar vectores $(a_i, b_i)_i$ para maximizar

$$\mathrm{Tr}(\sum_{i=1}^{2n} a_i b_i^T + b_i a_i^T) \quad \text{sujeto a} \quad \Sigma_X = \sum_{i=1}^{2n} a_i a_i^T \qquad \Sigma_Y = \sum_{i=1}^{2n} b_i b_i^T$$

Los puntos criticos de este problema son los del problema con multiplicadores de Lagrange siguiente,

$$\mathrm{Tr}(\sum_{i=1}^{2n} a_i b_i^T + b_i a_i^T) + \mathrm{Tr}((\sum_{i=1}^{2n} a_i a_i^T) * \Lambda) + \mathrm{Tr}((\sum_{i=1}^{2n} b_i b_i^T) * \Gamma)$$

para Λ, Γ dos matrices libre simétricas. Las condiciones de criticalidad dan

$$b_i^T = a_i^T \Lambda$$
$$a_i^T = b_i^T \Gamma$$

Lo que da en particular $V = \Sigma_X \Lambda$ y $\Lambda \Sigma_X \Lambda = \Sigma_Y$.
Si Σ_X es invertible, podemos tomar $\Lambda = \Sigma_X^{-1/2} R \Sigma_X^{-1/2}$ con R simétrica de tal manera que $\Sigma_Y = \Sigma_X^{-1/2} R^2 \Sigma_X^{-1/2}$ y $R^2 = \Sigma_X^{1/2} \Sigma_Y \Sigma_X^{1/2}$. En una base de diagonalización de R^2 devalores propios $\lambda_i \geq 0$, podemos diagonalizar también R asignando los valores propios $\pm\sqrt{\lambda_i}$. Pero

$$\mathrm{Tr}(V) = \mathrm{Tr}(\Sigma_X \Lambda) = \mathrm{Tr}(\Sigma_X^{1/2} R \Sigma_X^{-1/2}) = \mathrm{Tr}(R)$$

Entonces la solución al problema está dado por $\pm R = \pm(\Sigma_X^{1/2} \Sigma_Y \Sigma_X^{1/2})^{1/2}$ o finalmente

$$\mathrm{Tr}(V) = \mathrm{Tr}(\Sigma_X^{1/2} \Sigma_Y \Sigma_X^{1/2})^{1/2}$$

□

1.3 Ejemplos de cálculos de divergencia de Kullback-Leibler

Example 3. *Si* $p(0) = p(1) = 1/2$ *y* $q(0) = \frac{1-\epsilon}{2}$, $q(1) = \frac{1+\epsilon}{2}$, *entonces*

$$KL(q||p) = -\frac{1}{2}\log(1-\epsilon^2) \qquad y \qquad KL(p||q) = \frac{1}{2}\log(1-\epsilon^2) + \frac{\epsilon}{2}\log\left(\frac{1-\epsilon}{1+\epsilon}\right)$$

Para dos variables gaussianas,

Example 4. *Si* $P \sim \mathcal{N}(\mu_0, \Sigma_0)$ *y* $Q \sim \mathcal{N}(\mu_1, \Sigma_1)$,

$$2KL(P||Q) = \mathrm{Tr}(\Sigma_1^{-1}\Sigma_0) + (\mu_1 - \mu_0)^T \Sigma_1^{-1}(\mu_1 - \mu_0) + \log\left(\frac{|\Sigma_1|}{|\Sigma_0|}\right) - d$$

La divergencia KL forma parte de familias más grandes de divergencias usadas para varias razones estadísticas.

f-divergencias De la misma manera que la divergencia KL, las f-divergencias no son distancias.

Definition 2. *Sea $f : (0, +\infty) \to \mathbb{R}$ convexa y tal que $f(1) = 0$. Para $\mu, \nu \in \mathcal{P}(\mathbb{R}^n)$ y tal que $\mu << \nu$, definimos*

$$D_f(\mu||\nu) = \mathbb{E}_\nu \left[f\left(\frac{d\mu}{d\nu}\right)\right]$$

En el caso discreto, podemos escribir

$$D_f(\mu||\nu) = \sum_{x\in\mathcal{X}} \nu(x) f\left(\frac{\mu(x)}{\nu(x)}\right)$$

Exercice 2.

- *Si $f(x) = x \log x$, $D_f(\mu||\nu) = KL(\mu||\nu)$.*
- *Si $f(x) = \frac{1}{2}|x-1|$, $D_f(\mu||\nu) = d_{TV}(\mu||\nu)$.*
- *Si $f(x) = (1-\sqrt{x})^2$, $D_f(\mu||\nu) = \sum_{x\in\mathcal{X}}(\sqrt{\mu(x)} - \sqrt{\nu(x)})^2$ (distancia de Hellinger cuandrada)*

2 Distancias/Divergencias y acoplamientos

Aquí presentamos las interpretaciones en términos de acoplamientos de las distancias de las secciones anteriores. Nos referiremos a acoplamiento de dos medidas μ, ν como la distribución π de una variable (X, Y) donde $X \sim \mu$ y $Y \sim \nu$.

2.1 Variación total

Tenemos la interpretación siguiente para la distancia en variación total.

Proposition 3. *Para dos medidas $\mu, \nu \in \mathcal{P}(\mathbb{R}^n)$, $d_{TV}(\mu, \nu) = \min \mathbb{P}(X \neq Y)$ donde el mín se toma sobre los acoplamientos (X, Y) tal que $X \sim \mu$ y $Y \sim \nu$.*

Proof. De nuevo sin perder en generalidad probamos el teorema en el caso discreto. Sea $p = 1 - d_{TV}(\mu, \nu)$.
Si $p = 0$, $d_{TV(\mu,\nu)} = 1$ y los soportes de μ y ν son disjuntos lo que implica directamente $\mathbb{P}(X \neq Y) = 1$.
Si $p = 1$, $d_{TV(\mu,\nu)} = 0$ y $\mu = \nu$. Tomando $X = Y \sim \mu$, tenemos $\mathbb{P}(X \neq Y) = 0$.
Si $0 < p < 1$, tomamos (U, V, W) tal que

$$U \sim \frac{1}{p}(\mu \wedge \nu) \qquad V \sim \frac{1}{1-p}(\mu - \mu \wedge \nu) \qquad W \sim \frac{1}{1-p}(\nu - \mu \wedge \nu),$$

y finalmente sea $B \sim \text{Ber}(p)$ e independiente de las otras variables. Luego definimos

$$(X, Y) = \begin{cases} (U, U) & \text{si } B = 1 \\ (V, W) & \text{si } B = 0 \end{cases}$$

de tal manera que $X \sim \mu$, $Y \sim \nu$. Adicionalmente, $\text{supp}(V) \cap \text{supp}(W) = \emptyset$ entonces,

$$\mathbb{P}(X \neq Y) = \mathbb{P}(B = 0) = 1 - p = d_{TV}(\mu, \nu)$$

lo que demuestra la desigualdad $\geq$. Para la otra, uno nota que

$$1 - d_{TV}(\mu, \nu) = \sum_{x \in \mathcal{X}} \mu(x) \wedge \nu(x) \geq \sum_{x \in \mathcal{X}} \mathbb{P}(X = x, Y = x) = \mathbb{P}(X = Y).$$

□

2.2 Kullback-Leibler

Presentamos un resultado que involucra un acoplamiento en su demostración. Tiene que ver con una propiedad de sub-gaussianidad de una variable aleatoria.
Si $\forall \lambda > 0$,

$$\log \mathbb{E}\left[e^{\lambda(Z - \mathbb{E}[Z])}\right] \leq \frac{v\lambda^2}{2}$$

decimos que Z es sub-gaussiana de constante v. En particular eso implica, $\forall t > 0$,

$$\mathbb{P}(Z - \mathbb{E}[Z] \geq t) \leq e^{-\frac{t^2}{2v}}.$$

Theorem 1 (Ver Lemma 4.18 en [1])**.** *Sea P una medida de probabilidad sobre $\mathbb{R}$. Los siguientes son equivalentes,*

1. $\forall \lambda > 0$, $\log \mathbb{E}_P\left[e^{\lambda(Z - \mathbb{E}_P[Z])}\right] \leq \frac{v\lambda^2}{2}$
2. $\forall Q << P$, *y tal que* $KL(Q||P) < +\infty$,

$$\mathbb{E}_Q[Z] - \mathbb{E}_P[Z] \leq \sqrt{2vKL(Q||P)}.$$

La demostración de este teorema ofrece un acoplamiento entre las medidas P y Q del segundo punto.

2.3 Distancia de Wasserstein

De manera natural, la distancia de Wasserstein define un acoplamiento óptimo (para la distancia considerada) entre dos variables $X \sim \mu$ y $Y \sim \nu$.

3 Relaciones e interpretaciones con estadística

3.1 Dominación en distancias

Las 3 distancias que consideramos tienen algunas relaciones entre ellas. En particular tenemos

Proposition 4 (Pinsker)**.** *Sean $P, Q \in \mathcal{P}(\mathcal{X})$ con $Q << P$.*

$$d_{TV}(P,Q)^2 \leq \frac{1}{2} KL(P||Q)$$

Proof. Si $\frac{dQ}{dP} = f$ y $A^* = \{x : f(x) \geq 1\}$, definimos $Z = \mathbb{1}_{A^*}$ de tal manera que

$$d_{TV}(P,Q) = Q(A^*) - P(A^*) = \mathbb{E}_Q[Z] - \mathbb{E}_P[Z]$$

Pero como Z es acotada, el lema de Hoeffding dice

$$\log \mathbb{E}\left[e^{\lambda(Z - \mathbb{E}[Z])}\right] \leq \frac{\lambda^2}{8}$$

lo que implica el resultado por el teorema anterior. □

Si consideramos la distancia Lipschitz acotada

$$d_{BL}(\mu,\nu) = \sup_{\|f\|_{BL} \leq 1} \left| \int f d\mu - \int f d\nu \right|$$

con

$$\|f\|_{BL} = \max\left\{ \|f\|_\infty, \sup_{x \neq y} \frac{|f(x) - f(y)|}{|x - y|} \right\}$$

entonces tenemos

Proposition 5 (Ver [5])**.**

$$d_{BL}(\mu,\nu) \leq \min\{d_{TV}(\mu,\nu), W_1(\mu,\nu)\}$$

Exercice 3. *Usar la desigualdad de Hölder para mostrar que para $p \leq q$,*

$$W_p(\mu,\nu) \leq W_q(\mu,\nu)$$

En general, las diferentes distancias son todas equivalentes baja condiciones de log-concavidad de las medidas μ, ν. Más información está presente en [4].

3.2 Interpretación en estadística

Variación total y optimal testing Queremos crear un test $T : \mathcal{X} \to \{0, 1\}$ para distinguir entre las hipótesis

$$\begin{aligned} H_0 : &\quad X \sim \mu_0 \\ H_1 : &\quad X \sim \mu_1 \end{aligned}$$

minimizando el riesgo total

$$E_T = \mathbb{P}_{X\sim\mu_0}\left(T(X) = 1\right) + \mathbb{P}_{X\sim\mu_1}\left(T(X) = 0\right)$$

entonces tenemos

Proposition 6. *Sean $\mu, \nu \in \mathcal{P}(\mathcal{X})$. Para todo test T,*

$$E_T \geq 1 - d_{TV}(\mu, \nu)$$

con igualdad para $T^* = \mathbb{1}_{\overline{A^*}}$

Proof. (Caso discreto)
Sea $R = \{T(x) = 1\}$ y $R^* = \overline{A^*}$ entonces,

$$\begin{aligned} E_T &= \mathbb{P}_{X\sim\mu_0}\left(T(X) = 1\right) + \mathbb{P}_{X\sim\mu_1}\left(T(X) = 0\right) \\ &= 1 + \mu_0(R) - \mu_1(R) \\ &= 1 + \sum_{R\cap R^*} (\mu_0 - \mu_1)(x) + \sum_{R\cap\overline{R^*}} (\mu_0 - \mu_1)(x) \\ &= 1 - \sum_{R\cap R^*} |\mu_0 - \mu_1|(x) + \sum_{R\cap\overline{R^*}} |\mu_0 - \mu_1|(x) \\ &= 1 - \sum_{x\in\mathcal{X}} |\mu_0 - \mu_1|(x) \left(\mathbb{1}_{R\cap R^*} - \mathbb{1}_{R\cap\overline{R^*}}\right) \end{aligned}$$

que se minimiza por $R = R^*$ y entonces $E_T = 1 - \sum_{x\in\overline{A^*}} |\mu_0 - \mu_1|(x) = 1 - d_{TV}(\mu, \nu)$. □

Divergencia KL y testing/estimación De nuevo si nos interesamos en

$$\begin{aligned} H_0 : &\quad X \sim f_0 \\ H_1 : &\quad X \sim f_1 \end{aligned}$$

donde f_0, f_1 son densidades, entonces

$$KL(f_0||f_1) = \mathbb{E}_0\left[\log\left(\frac{f_0(X)}{f_1(X)}\right)\right]$$

Si $X_1, \ldots, X_n$ son variables i.i.d. con $X_i \sim f(x|\theta_0)$ y

$$\mathcal{F} = \{x \mapsto f(x|\theta) : \theta \in \Theta\}$$

entonces bajo "buenas condiciones" sobre Θ o $\mathcal{F}$ tenemos,

$$\hat{\theta}_n^{MLE} \longrightarrow \text{argmin}_{\theta\in\Theta} KL(f(\cdot|\theta_0)||f(\cdot|\theta))$$