Metrics & Divergences

Lecture 1

GOALS// Define distance-like operations between probability measures:

- To state theoretical results.
- To define objective functions for ML.

For simplicity we work with probability measures on $\mathbb{R}^d$, $\mathcal{P}(\mathbb{R}^d)$ of two kinds:

- with density $\rho(\cdot)$ wrt Lebesgue measure
- sum of Diracs $\delta(\cdot - v)$ or $\delta_v(\cdot)$

DIVERGENCE //

Definition / The Kullback-Leibler divergence between ρ_1 & ρ_2 pdfs on $\mathbb{R}^d$ is

$$D_{KL}(\rho_1 \| \rho_2) = \int_{\mathbb{R}^d} \log\left(\frac{\rho_1(u)}{\rho_2(u)}\right)\rho_1(u)\,du.$$

Example / $\rho_1 = N(0, \Sigma_1)$, $\rho_2 = N(0, \Sigma_2)$ then

$$D_{KL}(\rho_1 \| \rho_2) = \frac{1}{2}\left(\mathrm{Tr}(\Sigma_2^{-1}\Sigma_1) - d + \log\left(\det\Sigma_2 / \det\Sigma_1\right)\right)$$

②

Definition / $D: P(\mathbb{R}^d) \times P(\mathbb{R}^d) \to \mathbb{R}$ is a **divergence** if $\forall\, \rho_1, \rho_2 \in P(\mathbb{R}^d)$:

- $D(\rho_1, \rho_2) \geq 0$
- $D(\rho_1, \rho_2) = 0$ iff $\rho_1 = \rho_2$.

METRIC/ Divergence $D: P(\mathbb{R}^d) \times P(\mathbb{R}^d) \to \mathbb{R}^+$ is a **metric** if, in addition, $\forall\, \rho_1, \rho_2, \rho_3 \in P(\mathbb{R}^d)$,

- $D(\rho_1, \rho_2) = D(\rho_2, \rho_1)$;
- $D(\rho_1, \rho_3) \leq D(\rho_1, \rho_2) + D(\rho_2, \rho_3)$.

Examples/

- **Total variation distance**:

$$D_{TV}(\rho_1, \rho_2) = \frac{1}{2} \| \rho_1 - \rho_2 \|_{L^1(\mathbb{R}^d; \mathbb{R})}$$

- **Hellinger distance**:

$$D_H(\rho_1, \rho_2) = \frac{1}{\sqrt{2}} \| \sqrt{\rho_1} - \sqrt{\rho_2} \|_{L^2}$$

Theorem/ $\frac{1}{\sqrt{2}} D_{TV}(\rho_1, \rho_2) \leq D_H(\rho_1, \rho_2) \leq \sqrt{D_{TV}(\rho_1, \rho_2)}$

Theorem/ $D_H(\rho_1, \rho_2)^2 \leq \frac{1}{2} D_{KL}(\rho_1 \| \rho_2)$

$$D_{TV}(\rho_1,\rho_2)^2 \leq D_{KL}(\rho_1 \| \rho_2)$$

Remark/ The metrics do not work well with Dirac measures:

- fail to be defined (D_H, D_{KL})
- $D_{TV} = 1$ between $\rho_1 = \delta_{u_1}$, $\rho_2 = \delta_{u_2}$, $u_1 \neq u_2$.

TRANSPORT METRICS/

Definition/ A coupling of $\rho_1, \rho_2 \in P(\mathbb{R}^d)$ is $\pi \in P(\mathbb{R}^d \times \mathbb{R}^d)$:

$$\int_{\mathbb{R}^d} \pi(z,u)\,du = \rho_1(z), \quad \int_{\mathbb{R}^d} \pi(z,u)\,dz = \rho_2(u).$$

The set of all couplings is Π_{ρ_1,ρ_2}

Definition/ Wasserstien p-distance between ρ_1, ρ_2 is

$$W_p(\rho_1,\rho_2)^p = \inf_{\pi \in \Pi_{\rho_1,\rho_2}} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|z-u\|^p_{\mathbb{R}^d}\, \pi(z,u)\,dz\,du$$

Example/ $\rho_i = N(0, \Sigma_i)$ then

$$W_2(\rho_1,\rho_2)^2 = Tr\left[\Sigma_1 + \Sigma_2 - 2\left(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2}\right)^{1/2}\right]$$

In particular if $\rho_1 = N(0, \Sigma)$, $\rho_2 = \delta_0$ then

$W_2(\rho_1, \rho_2)^2 = Tr(\Sigma)$. ρ_1, ρ_2 close if $\|\Sigma\| \ll 1$.

But $D_{TV}(\rho_1, \rho_2) = 1 \quad \forall \Sigma \neq 0$.

INTEGRAL PROBABILITY METRICS //

Definition/ Let $\mathcal{F}$ be a set of functions $f: \mathbb{R}^d \to \mathbb{R}$. Then, $\forall \rho_1, \rho_2 \in P(\mathbb{R}^d)$.

$$D_{\mathcal{F}}(\rho_1, \rho_2) := \sup_{f \in \mathcal{F}} |\mathbb{E}^{\rho_1}(f) - \mathbb{E}^{\rho_2}(f)|$$

Example/ $D_{TV} = D_{\mathcal{F}}$ if $\mathcal{F} = \{ f \in L^\infty(\mathbb{R}^d) : \|f\|_{L^\infty} \leq \frac{1}{2} \}$

Definition/ The energy distance between $\rho_1, \rho_2 \in P(\mathbb{R}^d)$ is D_E given by

$$D_E(\rho_1, \rho_2)^2 = 2\mathbb{E}^{(u,v) \sim \rho_1 \otimes \rho_2} |u - v|$$

$$- \mathbb{E}^{(u,u') \sim \rho_1 \otimes \rho_1} |u - u'| - \mathbb{E}^{(v,v') \sim \rho_2 \otimes \rho_2} |v - v'|$$

Example/ D_E computable for $\rho_i = \frac{1}{N} \sum_{n=1}^{N} \delta_{u_i^{(n)}}$.