Metrics between probability distributions

Fundamental question in machine learning, statistics and indeed in science: How far is the learned probability from the real one? In order to answer this quantitatively one needs a notion of distance between probability distributions.

What do we want from a notion of distance? X: set, later we will specialize to the case where X is a set of probability distributions. $d: X \times X \to \mathbb{R}^{\geq 0}$

(i) distances should be non-negative

(ii) $\forall x \in X \quad d(x,x) = 0$

~~Rem.~~ (i) & (ii) are minimum requirements

(iii) $d(x,y) = 0 \Rightarrow x = y$

This is not always adopted. Sometimes it makes sense, sometimes not.

(iv) $d(x,y) = d(y,x) \to$ symmetry.

Again this makes sense when one is talking about the geometric distance in, say, $\mathbb{R}^n$ but just ask someone riding a bicycle up a hill if this is reasonable.

(v) $d(x,y) \leq d(x,z) + d(z,y)$ TRIANGLE

Guided by geometric intuition: shortest distance between 2 points is the direct path. But not all distances come from geometry.

A map $d : X \times X \to \mathbb{R}^{\geq 0}$ satisfying all of the above is called a <u>metric</u>. A map satisfying (i), (ii), (iv) & (v) but <u>not</u> (iii) is called a <u>pseudometric</u>. A map satisfying (i), (ii), (iii) & (v) is called a <u>quasi metric</u>.

A very popular ~~metric~~ distance used in information theory is the <u>relative entropy</u> or KL - divergence it satisfies (i), (ii) & (iii) but <u>not</u> (iv) or (v). This distance function violates geometric intuition but it has a good information theoretic intuition based on entropy & coding. Adam described Bregman divergences in his lecture on scoring rules.

I will focus on metrics & pseudo-metrics specifically between probability distributions. The main metric in the subject comes from transportation theory and is called the Kantorovich metric. It is widely <u>misnamed</u> the Wasserstein metric. There is <u>nobody</u> named Wasserstein and this name is a complete mistake. It was invented by many people at different times but Kantorovich was the main person who developed the theory.

Recall the concept of random variable

$(\Omega, \mathbb{P})$ ~~$(X, \mathbb{P})$~~ is a probability space, a random variable on Ω is a real-valued function $X : \Omega \to \mathbb{R}$ such that ....

(I am suppressing measure theory details)

$$\mathbb{E}[X] = \int_{\mathbb{R}} X \, dx$$

Given X we get an induced prob. measure on $\mathbb{R}$ $\mathbb{P}(X^{-1}(B))$ where $B \subseteq \mathbb{R}$. We write $\mathbb{P}_X$ for this measure on $\mathbb{R}$. We often write $\mathbb{P}\{X \in B\}$ for $\mathbb{P}(X^{-1}(B))$. The distribution function of a measure Q on $\mathbb{R}$ is a function $F : \mathbb{R} \to [0,1]$ s.t. $\forall x \in \mathbb{R}$ $F(x) = Q((-\infty, x])$. So for a random variable X we write F_X for the distribution function of $\mathbb{P}_X$.

Some basic metrics between RV's:

(1) $EN(X,Y) = |\mathbb{E}[X] - \mathbb{E}[Y]|$ where $\mathbb{E}[X], \mathbb{E}[Y] < \infty$

(2) $\rho(X,Y) = \sup_{x \in \mathbb{R}} \{ |F_X(x) - F_Y(x)| \}$; KOLMOGOROV

(3) $L(X,Y) = \inf_{\varepsilon > 0} F_X(x-\varepsilon) - \varepsilon \le F_Y(x) \le F_X(x+\varepsilon) + \varepsilon$ LEVY metric

Note (1) is not a metric but (2) & (3) are.

(4) $$K(X,Y) = \int_{\mathbb{R}} |F_X(x) - F_Y(x)| \, dx$$

KANTOROVICH metric

Metrics defined directly on measures:

TOTAL VARIATION METRIC:

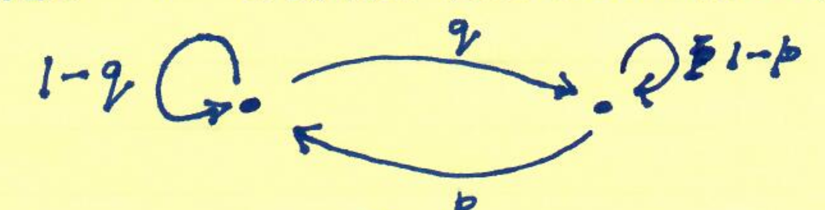
$$\Delta \text{ or } TV(P,Q) = \sup_{A \subseteq X} | P(A) - Q(A) |.$$

PINSKER'S inequality

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

Very useful reasoning about Markov chains and mixing times.

EXAMPLE : 2 state random walk



TRANSITION MATRIX

$$\begin{array}{c|cc} & E & W \\ \hline E & 1-p & p \\ W & q & 1-q \end{array}$$

Stationary distribution $\pi = \left(\frac{q}{p+q}, \frac{p}{p+q} \right)$

Assume we start on E $\mu_0 = (1,0)$ & define

$$\mu_{t+1} = T \mu_t$$

Define $\Delta_t = \mu_t(E) - \pi(E)$

one can show $\Delta_t = (1-p-q)^t \Delta_0$

$TV(\mu_t, \pi) = |\Delta_t|$ so

TV distance goes to zero exponentially fast.

Why I don't like TV:

Take $X = [0,1]$ and define $\forall x \in [0,1]$

$$\delta_x(A) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A \end{cases}$$

The "point mass" or Dirac measure.

$TV(\delta_x, \delta_y) = 1$ if $x \neq y$

So even if x, y are very close the TV is insensitive to this. ~~changes~~. If x & y are moving closer the TV stays constant.

Why I like the Kantorovich metric:
Let X be a metric space with metric d. Let $P(X)$ be the space of probability distributions on X & K the Kantorovich metric. Then $K(\delta_x, \delta_y) = d(x,y)$ i.e. there is an embedding $e: X \to P(X)$ by

$e(x) = \delta_x$ which is an <u>isometry</u>.

Of course, K can only be defined when there is an <u>underlying</u> metric space.

— x —

Some background: Given 2 measures P, Q on a space X we define a <u>coupling</u> of P, Q to be a probability measure π on $X \times X$ such that the marginals

$A \subseteq X$ $\pi_x(A) := \pi(A \times X)$ & $B \subseteq X$ $\pi_y(B) = \pi(X \times B)$

are P, Q respectively.

Suppose P, Q are two measures on $\mathbb{R}$ a coupling is a joint measure π on $\mathbb{R}\times\mathbb{R}$ with P, Q as its marginals. We can also define it to be a pair of random variables X, Y on $(\Omega, \mathbb{P})$ s.t.

$$\mathbb{P}_X = P \ \& \ \mathbb{P}_Y = Q.$$

If we have such a pair of RV's we can define $\pi(A\times B) = \mathbb{P}(X^{-1}(A) \cap Y^{-1}(B))$.

Easy to see that the marginals of π are P and Q.

If we have a coupling in the first sense it is easy to define a pair of RV's.

<u>Ex</u> Let $\mathbb{X} = \{H, T\}$ be the sample space of a coin. Let P, Q both be the fair (uniform) dist. Let (X, Y) be a pair of RV's on $\mathbb{X}$ so that $\mathbb{P}\{X = x, Y = y\} = 1/4$ for all $x, y \in \{H, T\}$

Another coupling is (X, Y) with

$$\mathbb{P}\{X = Y = H\} = 1/2 \ \& \ \mathbb{P}\{X = Y = T\} = 1/2.$$

In this case $\mathbb{P}\{X \neq Y\} = 0$.

<u>Prop.</u> $TV(P, Q) = \inf_{\substack{\text{couplings}\\(X,Y)}} \{\mathbb{P}\{X \neq Y\}\}$

<u>Proof.</u> I will only prove a part namely

$$TV(P,Q) \leq \inf \mathbb{P}\{X \neq Y\}$$

For any coupling (X, Y) we have

$$P(A) - Q(A) = \mathbb{P}\{X \in A\} - \mathbb{P}\{Y \in A\}$$

$$= \cancel{\mathbb{P}\{X \in A, Y \in A\}} + \mathbb{P}\{X \in A, Y \notin A\} - \cancel{\mathbb{P}\{X \in A, Y \in A\}} - \mathbb{P}\{X \notin A, Y \in A\}$$

$$\leq \mathbb{P}\{X \in A, Y \notin A\} \leq \mathbb{P}\{X \neq Y\}$$

so $\sup |\mathbb{P}(A) - Q(A)| \leq \inf\{\mathbb{P}\{X \neq Y\}\}$

Levy-Prokhorov metric

(X, d) a metric space

$A \subset X$ we define $A^\varepsilon = \{p \in X \mid \exists q \in A, d(p,q) < \varepsilon\}$

This is an open set $A^\varepsilon = \bigcup_{p \in A} B_\varepsilon(p)$

where $B_\epsilon(p) = \{q \in X \mid d(p,q) < \varepsilon\}$.

Let P, Q be probability distributions

$$LP(P, Q) = \inf\{\epsilon > 0 \mid P(A) \leq Q(A^\varepsilon) + \varepsilon \text{ and } Q(A) \leq P(A^\varepsilon) + \varepsilon \ \forall A \in X\}$$

* If (X, d) is a complete separable metric space then so is $(\mathcal{P}(X), LP(\cdot,\cdot))$.

Convergence in LP is equivalent to weak convergence of measures.