Class 8

Shikhar Saxena

February 13, 2023

Contents

Recap

Recall $x_1, x_2, \dots x_n \sim p$

Can construct a density estimate $\hat{p}_\theta(x) \approx p$. We can then sample from $y \sim \hat{p}_\theta(x)$ to generate.

This is one approach of generation.

But GAN uses two distributions and we'll see how that follows.

So, at the EOD we want to measure $\hat{p}_\theta(x)$ wrt p. **So how to measure two distributions?**

Some papers that highlight some measures:

- Adhikari and Joshi 1956 "measures of distance"
- Rao 1952, "measures of separations"
- Chernoff and Kullback, "measures of discriminatory intent"
- Kolmogorov 1963, "measures of variation distance"

Example 1. *Two-norm* $\|x-y\|_2^2$ *preserves differentiability. While one-norm* $\|x-y\|_1$ *preserves sparsity.*

Thus, we see different measures measure (and quantify) different sort of information.

Another way to measure is to divide the PMF at each point.

$$\phi(x) = \frac{p_1(x)}{p_2(x)}$$

Then we can take log *of this which will be zero when both distributions are equal. We can go further and take expectation of this* $E_{x \sim p_1} \log \frac{p_1(x)}{p_2(x)}$.

Remark. *A common property for each measure is that they* ***increase*** *as two distributions* ***more apart****.*

Kullback-Liebler (KL) measure

$$D(p \| p_\theta) = \sum_x p(x) \log \frac{p(x)}{p_\theta(x)} = E_{x \sim p} \log \frac{p(x)}{p_\theta(x)}$$

KL measure is very equivalent to say a loss function. Whatever p_θ we get, we want to minimize this measure.

Minimize KL divergence

$$\begin{aligned}
\inf_\theta D(p \| p_\theta) &= \inf_\theta \sum_x p(x) \log \frac{p(x)}{p_\theta(x)} \\
&= \inf_\theta \sum_x p(x) \log \frac{1}{p_\theta(x)} - \sum_x p(x) \log \frac{1}{p(x)} \\
&\textit{Ignore the 2nd term because it's not dependent on } \theta \\
&= \inf_\theta \left(\sum_x p(x) \log \frac{1}{p_\theta(x)} \right) \\
&= \sup_\theta \left(\sum_x p(x) \log p_\theta(x) \right) \\
&= \sup_\theta E_{x \sim p} \log p_\theta(x) \ \textit{ which is the log-MLE estimate}
\end{aligned}$$

Some properties of KL-divergence

1. KL-Divergence is convex in the pair (p, q)

Claim:

$$KL(\lambda p_1 + (1-\lambda) p_2 \| \lambda q_1 + (1-\lambda) q_2) \leq \lambda KL(p_1 \| q_1) + (1-\lambda) KL(p_2 \| q_2)$$

Proof. We use the log-sum inequality (exercise: try to prove),

$$\sum_{i=1}^n a_i \log \frac{a_i}{b_i} \geq \left(\sum_{i=1}^n a_i \right) \log \frac{\sum_{i=1}^n a_i}{\sum_{i=1}^n b_i}$$

$$
\begin{aligned}
&KL(\lambda p_1 + (1-\lambda)p_2 \| \lambda q_1 + (1-\lambda)q_2) \\
&= \sum_x \left[(\lambda p_1(x) + (1-\lambda)p_2(x)) \log \frac{\lambda p_1(x) + (1-\lambda)p_2(x)}{\lambda q_1(x) + (1-\lambda)q_2(x)} \right] \\
&\textit{Take } a_1 = \lambda p_1, a_2 = (1-\lambda)p_2 \\
&\textit{and } b_1 = \lambda q_1, b_2 = (1-\lambda)q_2 \\
&\textit{On applying log sum ineqality we have} \\
&\leq \sum \left[\lambda p_1(x) \log \frac{p_1(x)}{q_1(x)} + (1-\lambda)p_2(x) \log \frac{p_2(x)}{q_2(x)} \right] \\
&= \lambda KL(p_1 \| q_1) + (1-\lambda) KL(p_2 \| q_2)
\end{aligned}
$$

□

Example

x	0	1	2
$p(x)$	9/25	12/25	4/25
$q(x)$	1/3	1/3	1/3

Now compute $D_{KL}(p\|q)$.

Properties (contd.)

KL-divergence is not a **metric**.

- Since not commutative
- Triangle Ineqality not satisfied (exercise: check)

Generalize KL-divergence

Now we will generalize KL to some variational approach.

$$D(p\|q) = \sum_x p(x) \log \frac{p(x)}{q(x)} = \sum p(x) \left\{ -\log \frac{q(x)}{p(x)} \right\}$$

So, we can replace $-\log$ with some other convex function f such that $f(1) = 0$.

f-divergence

Introduced by “A general class of coefficients of divergence of one distribution from another” by S. M. Ali, S. Silvey (1965).

Definition 1. *Let $f : R \to R$, convex lower-semicontinuous such that $f(1) = 0$. We define f-divergence between two densities p and q on X as*

$$D_f(p\|q) = \int_X q(x) f\left(\frac{p(x)}{q(x)}\right) dx$$

KL-divergence is a special case of F-divergence. Take $f(x) = x \log x$ to show that.

Definition 2. *f is lower-semicontinuous (lsc) if and only if*

$$\lim_{x \to x_0} f(x) \geq f(x_0)$$

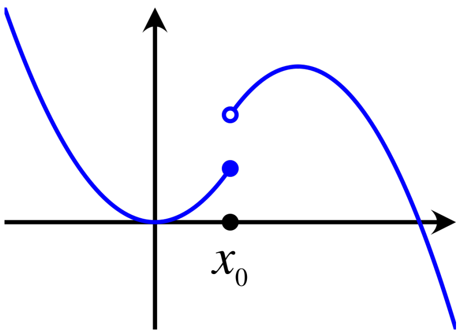

Figure 1: Lower Semicontinuity

Equivalent definitions for lsc

- $\{x \in X, f(x) \leq y\}$ Sub level sets are closed in X
- Epigraph is closed in X

(exercise) Write two functions that are both lsc and convex? Can we have a function that is cocave but lsc?

Tractable f-divergence

The integral is usually difficult to solve. So another researcher proposed this measure. Essentially we construct a lower bound that is tractable

Variational representation of f-divergence

Conjugate function revisited

(exercise)

- Find conjugate for $f(x) = ax + b$. Comes out to be $f^*(t) = -b$.
- Find conjugate for $f(x) = x \log x$. Comes out to be $f^*(t) = e^{t-1}$.
- Fenchel's inequality

$$f^*(y) + f(x) \geq y^T x \ \forall x, y$$

Conjugate of conjugate

Theorem 1 (Fenchel's - Moreau Theorem)**.** *If f is closed and a convex function then $f^{**} = f$.*

Proof can be taken as a project and found in the Rockafeller Convex analysis 1970 book.

Variational Approach

Proposed by Nguyen, Jordan 2010.

Claim: $D_f(p\|q) = \sup_{T:X\to R} E_{x\sim p}T(x) - E_{x\sim q}f^*(T(x))$

Using variational representation, $f(x) = sup_t \{tx - f^*(t)\}$.

It is ***variational approach*** since for varieties of t we are approximating f.

Proof. $f(x) = sup_t \{tx - f^*(t)\}$ (by fenchel's moreau theorem).

So,

$$
\begin{aligned}
D_f(p\|q) &= \int_X q(x) \sup_t \left[t\frac{p(x)}{q(x)} - f^*(t) \right] dx \\
D_f(p\|q) &= \int_X \sup_t \left[tp(x) - f^*(t)q(x) \right] dx
\end{aligned}
$$

Here, we cannot send the supremeum out and the integral in (because each t is calculated pointwise for all x to generate the function which is then integrated over). So, we try to construct operators T over X such that this can be mimicked and the operators can be interchanged.

$$
\begin{aligned}
D_f(p\|q) &= \sup_{T:X\to R} \int_X T(x)p(x) - f^*(T(x))q(x)dx \\
D_f(p\|q) &= \sup_{T:X\to R} E_{X\sim p}T(x) - E_{x\sim q}f^*(T(x))
\end{aligned}
$$

Using parameterized family of functions T_ϕ we try to approximate the optimal T (lower bound), can minimize an f-divergence between p and the pushforward p_θ by solving:

$$
arg \min_\theta \sup_\phi \left[E_{x\sim p}T_\phi(x) - E_{x\sim p_\theta} f^*(T_\phi(x)) \right]
$$

Thus, this is a template min-max problem (that will be specified in more detail in GANs). We can change T_ϕ family and f to get a variety of different approaches for generative modeling.

□