Machine Learning

Lecture 10

Lecturer:Haim Permuter *Scribe: Omer Luxembourg*

I. Introduction

In this lecture we introduce the *f-Divergence* definition which generalizes the *Kullback-Leibler Divergence*, and the *data processing inequality* theorem. Parts of this lecture are guided by the work of T. Cover's book [1], Y. Polyanskiy's lecture notes [3] and Z. Goldfeld's lecture 6 about *f-Divergences* [2]. This lecture assumes the student is familiar with basic probability theory. The notations here are similar to those of the previous lectures.

II. *f-Divergence*

Definition 1 (*Kullback-Leibler Divergence*) Recall the *Kullback-Leibler Divergence* (a.k.a. KL-Divergence) definition:

$$D_{KL}(P_X||Q_X) \triangleq \mathbb{E}_P\left[\log\left(\frac{P(x)}{Q(x)}\right)\right]. \tag{1}$$

For discrete probabilities eq. (1) becomes:

$$D_{KL}(P_X||Q_X) \triangleq \sum_{x\in\mathcal{X}} P(x)\log\left(\frac{P(x)}{Q(x)}\right), \tag{2}$$

and for continuous probabilities:

$$D_{KL}(P_X||Q_X) \triangleq \int_{x\in\mathcal{X}} P(x)\log\left(\frac{P(x)}{Q(x)}\right)dx, \tag{3}$$

for P, Q such that if $Q(x) = 0$ then $P(x) = 0$ for the same x.

There are two main properties for *Divergence*, which were proved in previous lectures.

a. $D_{KL}(P_X||Q_X) \geq 0$, and equality hold if and only if $P = Q$.
b. $D_{KL}(P_X||Q_X)$ is *convex* in (P_X, Q_X).

Definition 2 **(*f-Divergence*)** For two distributions P and Q, the *f-Divergence* is defined as:

$$D_f(P_X||Q_X) \triangleq \mathbb{E}_Q\left[f\left(\frac{P(x)}{Q(x)}\right)\right], \tag{4}$$

for P, Q, such that if $Q(x) = 0$ then $P(x) = 0$ for the same x, and for f that satisfies the following:

- f is *convex* for $\mathbb{R}^+$.
- $f(1) = 0$.

The following are special cases of *f-Divergences*:

a. ***Kullback-Leibler Divergence:*** a.k.a. relative entropy, $f(x) = x\log x$,

$$\begin{aligned} D_f(P_X||Q_X) &\triangleq \mathbb{E}_Q\left[f\left(\frac{P(x)}{Q(x)}\right)\right] \\ &\overset{(a)}{=} \sum_{x\in\mathcal{X}} Q(x)\cdot\frac{P(x)}{Q(x)}\log\left(\frac{P(x)}{Q(x)}\right) \\ &= \sum_{x\in\mathcal{X}} P(x)\log\left(\frac{P(x)}{Q(x)}\right) \\ &\triangleq D_{KL}(P_X||Q_X), \end{aligned} \tag{5}$$

where (a) follows from the definition of *f*. Note that $f(1) = 0$ and f is *convex* for all $t \geq 0$. ($f''(t) = \frac{1}{t}$).

b. ***Negative Log:*** $f(x) = -\log(x)$,

$$\begin{aligned} D_f(P_X||Q_X) &\triangleq \mathbb{E}_Q\left[f\left(\frac{P(x)}{Q(x)}\right)\right] \\ &\overset{(a)}{=} \sum_{x\in\mathcal{X}} -Q(x)\log\left(\frac{P(x)}{Q(x)}\right) \\ &\triangleq D(Q_X||P_X), \end{aligned} \tag{6}$$

where (a) is the definition of divergence, which is non-negative, and 0 if $P = Q$. Note that $f(1) = 0$ and f is *convex* for all $t \geq 0$. It is worth noting that, in general, $D(P||Q) \neq D(Q||P)$.

c. ***Total Variation:*** $f(x) = \frac{1}{2}|x-1|$,

$$D_{TV}(P,Q) \triangleq D_{f_{TV}}(P_X||Q_X) \tag{7}$$

$$
\begin{aligned}
&= \mathbb{E}_Q \left[f_{TV} \left(\frac{P(x)}{Q(x)} \right) \right] \\
&= \sum_{x \in \mathcal{X}} Q(x) \cdot \frac{1}{2} \left| \frac{P(x)}{Q(x)} - 1 \right| \\
&= \frac{1}{2} \sum_{x} |P(x) - Q(x)| \, .
\end{aligned}
$$

Note that $f(1) = 0$ and f is *convex* for all $t \geq 0$. In addition $D_{TV}(P, Q) = D_{TV}(Q, P)$ means that the *total variation* is a *metric* on the space of probability distributions. That is because it is a divergence function and a symmetric function of P and Q .

d. Jensen-Shannon divergence (symmetrized KL): $f(x) = x \log \frac{2x}{x+1} + \log \frac{2}{x+1}$,

$$
\begin{aligned}
D_{JS}(P||Q) &\triangleq D_{f_{JS}}(P_X||Q_X) \\
&= \mathbb{E}_Q \left[f \left(\frac{P(x)}{Q(x)} \right) \right] \\
&= \sum_{x \in \mathcal{X}} Q(x) \left(\frac{P(x)}{Q(x)} \log \frac{2\frac{P(x)}{Q(x)}}{\frac{P(x)}{Q(x)} + 1} + \log \frac{2}{\frac{P(x)}{Q(x)} + 1} \right) \\
&= \sum_{x \in \mathcal{X}} P(x) \log \left(\frac{P(x)}{\frac{P(x)+Q(x)}{2}} \right) + Q(x) \log \left(\frac{P(x)}{\frac{P(x)+Q(x)}{2}} \right) \\
&\overset{(a)}{=} D\left(P|| \frac{P+Q}{2} \right) + D\left(Q|| \frac{P+Q}{2} \right),
\end{aligned} \tag{8}
$$

where (a) is the definition of divergence.

$f(1) = 0$ and f is a *convex* function. ($f''(x) = \frac{1}{x^2+x} \geq 0$ for all $x > 0$).

Theorem 1 (Properties of *f-Divergence*).

- ***Non-negativity:*** *For a f function that is strictly convex around 1,* $D_f(P||Q) \geq 0$*. The equality holds if and only if* $P = Q$*.*

  Proof:

$$
\begin{aligned}
D_f(P||Q) &= \mathbb{E}_Q \left[f \left(\frac{P}{Q} \right) \right] \\
&\overset{(a)}{\geq} f\left(\mathbb{E}_Q \left[\frac{P(x)}{Q(x)} \right] \right) \\
&\overset{(b)}{=} f(1)
\end{aligned} \tag{9}
$$

$$\stackrel{(c)}{=} \quad 0,$$

where (a) is from *Jensen's inequality* for a *convex* function *f*, (b) is due to the fact that $\frac{P(x)}{Q(x)}$ is fixed $\forall x$ because $P = Q$, (c) is from the definition of *f*. Note that if *f* is not *strictly convex* around 1, the equality can hold from *Jensen's inequality* and not from $P = Q$.

- ***Joint convexity:*** $(P, Q) \longmapsto D_f(P||Q)$ *is a jointly convex function. Consequently,* $P \longmapsto D_f(P||Q)$ *for fixed* Q *and* $Q \longmapsto D_f(P||Q)$ *are also convex functions.*

  Proof: From the *Perspective Transform Preserve Convexity* lemma we learned that if $f(x)$ is convex $\Rightarrow t \cdot f\left(\frac{x}{t}\right)$ is convex in (x, t).

  $$D_f(P||Q) = \sum_x Q(x) f\left(\frac{P(x)}{Q(x)}\right), \tag{10}$$

  f is a convex function; thus, from the *Perspective Transform Preserve Convexity Lemma*, $Q(x) \cdot f\left(\frac{P(x)}{Q(x)}\right)$ is convex in (x, t). Therefore $D_f(P||Q)$ is the sum of convex functions in (P, Q) by eq. (10); thus it is a convex function in (P, Q).

Theorem 2 ***Conditioning Increases f-Divergence:*** Define the conditional *f-Divergence*

$$D_f(P_{Y|X}||Q_{Y|X}|P_X) \triangleq \mathbb{E}_{P_{X,Y}}\left[D_f\left(P_{Y|X}||Q_{Y|X}\right)\right]. \tag{11}$$

Let P_Y be the output of the system $P_{Y|X}$ for input P_X, and Q_Y be the output of the system $Q_{Y|X}$ for input P_X, see figure 1.

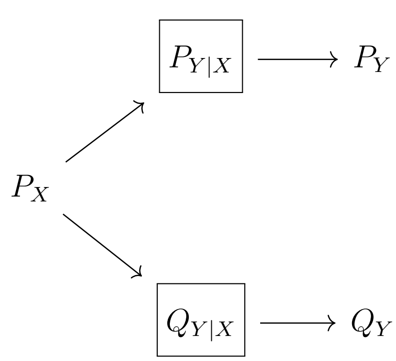

Fig. 1. Channel transition matrices

Then

$$D_f(P_Y||Q_Y) \leq D_f\left(P_{Y|X}||Q_{Y|X}|P_X\right). \tag{12}$$

One can view P_Y and Q_Y as the output distributions after passing P_X through the channel transition matrices $P_{Y|X}$ and $Q_{Y|X}$, respectively. The above relation tells us that the average *f-Divergence* between the corresponding channel transition rows is at least the *f-Divergence* between the output distributions.

Proof:

$$\begin{aligned} D_f(P_{Y|X}||Q_{Y|X}|P_X) &\triangleq \sum_x P_X \sum_y Q(Y|X) f\left(\frac{P(Y|X)}{Q(Y|X)}\right) \\ &\overset{(a)}{=} \sum_x P_X D_f\left(P(Y|X=x)||Q(Y|X=x)\right) \\ &\overset{(b)}{\geq} D_f\left(\left(\sum_x P_X P(Y|X=x)\right)||\left(\sum_x P_X Q(Y|X=x)\right)\right) \\ &\overset{(c)}{=} D_f\left(\mathbb{E}_{P_X}\left[P(Y|X)\right]||\mathbb{E}_{P_X}\left[Q(Y|X)\right]\right) \\ &\overset{(d)}{=} D_f\left(P(Y)||Q(Y)\right), \end{aligned} \tag{13}$$

where (a) follows from the definition of *f-Divergence*, (b) follows from *Jensen's inequality*, because D_f is convex in P, Q, (c) is the definition of expectation, and (d) follows from the *Law of Total Expectation*.

Remark 1 (equality for $D_f(P_{Y|X}||Q_{Y|X}|P_X)$): We can notice the following equality holds:

$$\begin{aligned} D_f(P_{Y,X}||\tilde{Q}_{Y,X}) &\triangleq \mathbb{E}_{\tilde{Q}_{Y,X}}\left[f\frac{P_{Y,X}}{\tilde{Q}_{Y,X}}\right] \\ &= \sum_{y,x} \tilde{Q}(y,x) f\left(\frac{P(y,x)}{\tilde{Q}(y,x)}\right) \\ &= \sum_x P(x) \sum_y Q(y|x) f\left(\frac{P(y,x)}{Q(y,x)}\right) \\ &\overset{(a)}{=} \sum_x P(x) \sum_y Q(y|x) f\left(\frac{P(y|x)P(x)}{Q(y|x)P(x)}\right) \end{aligned} \tag{14}$$

$$\stackrel{(b)}{=} \sum_x P(x) \sum_y Q(y|x) f\left(\frac{P(y|x)}{Q(y|x)}\right)$$
$$= D_f(P_{Y|X}||Q_{Y|X}|P_X),$$

where (a) follows from the definition of conditional probability, and $\tilde{Q}_{(}y,x) \triangleq P(x)Q(y|x)$, and (b) is from the definition of divergence.

III. Data Processing Inequality

The data processing inequality for KL divergence extends to all f-Divergences.

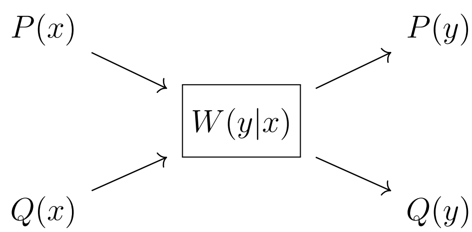

Fig. 2. One channel transition [3]

The intuition behind the following inequality is that processing the observation x by a channel $W_{Y|X}$ makes it more difficult to determine whether it came from P_X or Q_X. In neural networks, for instance, the divergence of the system output will decrease as we move to the next layer.

Theorem 3 (Data Processing Inequality): Consider a channel that produces Y given X based on the law $W_{Y|X}$. If P_Y and Q_Y are distributions of Y when X is generated by P_X and Q_X, respectively, then for *any f-Divergence*,

$$D_f(P_X||Q_X) \geq D_f(P_Y||Q_Y), \tag{15}$$

as for the KL divergence.

Proof:

$$D_f(P_X||Q_X) \triangleq D_f(P_X W_{Y|X}||Q_X W_{Y|X}) \tag{16}$$
$$= \sum_{y,x} Q(x,y) f\left(\frac{P(x,y)}{Q(x,y)}\right)$$

$$
\begin{aligned}
&\overset{(a)}{=} \sum_y Q(y) \sum_x Q(x|y) f\left(\frac{P(x,y)}{Q(x,y)}\right) \\
&\overset{(b)}{\geq} \sum_y Q(y) f\left(\sum_x Q(x|y) \frac{P(x,y)}{Q(x,y)}\right) \\
&= \sum_y Q(y) f\left(\sum_x Q(x|y) \frac{P(x,y)}{Q(y)Q(x|y)}\right) \\
&\overset{(c)}{=} \sum_y Q(y) f\left(\frac{P(y)}{Q(y)}\right) \\
&= D_f(P_Y||Q_Y),
\end{aligned}
$$

where (a) follows from conditioning, (b) is *Jensen's inequality* for convex f in P, Q, and (c) is from *Law of Total Probability*. Note that $P_{X,Y} = P_X W_{Y|X}$ and $Q_{X,Y} = Q_X W_{Y|X}$.

REFERENCES

[1] T. M. Cover and J. A. Thomas. *Elements of Information Theory, Chap. 1*. ISBN, 1991.

[2] Z. Goldfeld. Lecture 6: f-divergences.
Available at http://people.ece.cornell.edu/zivg/ECE_5630_Lectures6.pdf, 2020.

[3] Y. Polyanskiy. Lecture notes on information theory, chap. 6.
Available at http://people.lids.mit.edu/yp/homepage/data/itlectures_v5.pdf, 2017.