Lecture 5 : 19th Jan 2026 :

Distributional Divergence

objective : Given a pair of distribution fun^c defined on the same sample space, define / quantify "distance" b/w them.

Given an event $A \in \mathcal{F}$,

define $I(A) \triangleq -\log \mathbb{P}(A)$

likely info has less info

less likely info has more info

→ Info of Ω has zero info

→ null event has ∞ info

→ A & B are indpt, $\mathbb{P}(AB) = \mathbb{P}[A]\,\mathbb{P}[B]$

$$\log \mathbb{P}[AB] = \log \mathbb{P}[A] + \log \mathbb{P}[B]$$

→ $-\log \mathbb{P}(A)$ satisfies all the above Cond^n

→ to transfer the info, how many bits I need.

- Suppose X is DRV

then the "Surprisal" associated with it is

quantified as follows

$$\mathop{\mathbb{E}}_{\mathbb{P}_X}\left[-\log p_X(x)\right] \quad \text{by LOTUS.}$$

$$= \sum_{i=1}^{n} -p_X(x_i)\log p_X(x_i) \quad : \text{Avg info in } \mathbb{P}_X$$

Expected Value of $\mathbb{P}_X$

* This is the entropy of $\mathbb{P}_X$

$$H(\mathbb{P}_X) = \mathop{\mathbb{E}}_{\mathbb{P}_X}\left[-\log p_X(x)\right]$$

Suppose $\mathbb{P}_X$ & Q_X are 2 distributions on the same sample space

Define $\mathop{\mathbb{E}}_{\mathbb{P}_X} -\log Q_X(x) = H(\mathbb{P}_X, Q_X)$: Cross entropy

Samples from $\mathbb{P}_X$, estimate Q_X

Avg info Q_X Contain about $\mathbb{P}_X$

* info that One distribution Contains about another distributions.

* Cross entropy is not symmetric.

Consider $-H(P_x) + H(P_x, Q_x)$

$$= +\sum p_x(x_i) \log p_x(x_i) - \sum p_x(x_i) \log q_x(x_i)$$

Extra info of Q_x which is present

$$D_{KL}(P_x \| Q_x) = \sum p_x(x_i) \log \frac{p_x(x_i)}{q_x(x_i)}$$

Show:

1) $D_{KL}(P_x \| Q_x) \neq D_{KL}(Q_x \| P_x)$

2) $D_{KL} \geq 0$

3) $D_{KL} = 0$ iff $P_x = Q_x$

Does not obey Law of triangularity so its not a metric.

Now a similar defn when X is CRV

$$D_{KL}(P_x \| Q_x) = \int_X p_x(x) \log \frac{p_x(x)}{q_x(x)} dx.$$

Suppose we have "N" samples drawn i.i.d.

$$D = \{ v_1, v_2, \ldots, v_N \} \sim i.i.d \ \mathbb{P}_V$$

Goal is to estimate $\mathbb{P}_V$ given D.

Assume $p_\theta(v)$ as the model density & compute

$$D_{KL}(\mathbb{P}_V \,||\, \mathbb{P}_\theta) = \int p_V(v) \log \frac{p_V(v)}{p_\theta(v)} \, dv.$$

$$\theta^* = \arg\min_\theta D_{KL}(\mathbb{P}_V \,||\, \mathbb{P}_\theta)$$

$$= \arg\min_\theta \Big[\int p_V(v) \log p_V(v) \, dv - \int p_V(v) \log p_\theta(v) \, dv \Big]$$

Independent of θ

$$= \arg\min_\theta \Big[- \int p_V(v) \log p_\theta(v) \, dv \Big]$$

$$= \arg\min_\theta - \int p_V(v) \log p_\theta(v) \, dv$$

$$= \arg\min_\theta - \mathbb{E}_{\mathbb{P}_V} \log p_\theta(v)$$

wrt expectations can be approximated via samples

By the Law of Large numbers.

$$\mathop{\mathbb{E}}_{\mathbb{P}_V} \log p_\theta(v) \approx \frac{1}{N} \sum_{i=1}^{n} \log p_\theta(v_i)$$

where $v_i \sim i.i.d\ \mathbb{P}_V$

$$\frac{1}{N} \sum_{i=1}^{N} \log p_\theta(v_i) \longrightarrow \mathop{\mathbb{E}}_{\mathbb{P}_V} \log p_\theta(v) \quad \text{as } N \to \infty$$

Now our optimization problem is. (Negative Log Likelihood (NLL))

$$\therefore \quad \theta^* = \operatorname*{argmin}_\theta \left[\frac{-1}{N} \sum_{i=1}^{N} \log p_\theta(v_i) \right]$$

This is the Minimal D_{KL} estimator.

& we say it as Maximum Likelihood estimator

Another Algebra : Joint Likelihood of the Data

$$L_\theta(D) = \prod_{i=1} p_\theta(v_i)$$

taking a monotonic log func

$$L_\theta(D) = \sum_{i=1}^{N} \log p_\theta(v_i)$$

$$\theta^* = \underset{\theta}{\operatorname{argmax}} \frac{1}{N} \sum_{i=1}^{N} \log p_\theta(v_i)$$

How to solve the ML estimation problem in practice ?

a) how to choose p_θ ?

b) how to solve the optimization problem ?

Solution to the problems:

a) p_θ is choosen using a canonical func approx

eg : Assume $p_\theta(\,) \sim N(\,; h_\theta(v), I)$

$h_\theta(v)$: hypothesis func / Model

→ Now it boils to estimating parameters of this func

b) Optimization is mostly numerical

Consider the problem of supervised regression.

$x \in \mathbb{R}^d, \quad y \in \mathbb{R}$

Data $D = \{(x_i, y_i)\}_{i=1}^N \sim$ iid $\mathbb{P}_{xy}$

Estimate $\mathbb{P}_{y|x}$

$p_\theta(y|x) \sim N(y; h_\theta(x), I)$ Assumption. Model

$h_\theta(x) : x \to \mathbb{R}$

$$\theta^* = \arg\min_\theta -\frac{1}{N} \sum_{i=1}^N \log p_\theta(y_i|x_i)$$

$$\propto \arg\min_\theta \frac{-1}{N} \sum_{i=1}^N \log \exp\{-(y_i - h_\theta(x_i))^2\}$$

$$\theta^* \propto \arg\min_\theta \frac{1}{N} \sum_{i=1}^N (y_i - h_\theta(x_i))^2$$

This is the least square estimate

Lecture 6: 21st Jan 2026:

Risk Minimization Frome work

Given Data $D = \{(x_i, y_i)\}_{i=1}^{N} \sim$ i.i.d $\mathbb{P}_{XY}$

$x \in \mathbb{R}^d$, $y \in \mathbb{R}^k / \{1, \ldots, \kappa\}$

Goal is to estimate hypothesis func

$h : X \to Y$ $\quad h \in \mathcal{H}$ set of all func from $X \to Y$.

Quantify the "good ness" of an h

Define a notion of Loss-func L

$L : Y \times Y \to \mathbb{R}^+$

$L(h(x), y) \to \mathbb{R}^+$

Eg: a) $L(h(x), y) = \| h(x) - y \|_2^2$

squared error Loss.

b) $L(h(x), y) = \begin{cases} 0 & \text{if } h(x) = y \\ 1 & \text{if } h(x) \neq y \end{cases}$

Zero-One Loss.

Extending the Loss func^c for the population.

Define $R(h) \triangleq \mathbb{E}_{P_{xy}} L(h(x), Y)$ Risk function.

Problem : Estimate $h^*(x)$

$$h^*(x) = \underset{h \in \mathcal{H}}{\operatorname{argmin}} \; \mathbb{E}_{P_{xy}} \; L(h(x), Y)$$

This is known as the Risk minimization framework.

We can't solve this as we do not know P_{xy}

$\therefore$ we look @ its surrogate

True risk can't be minimized $\because$ P_{xy} is Unknown

$\therefore$ Consider the surrogate

$$\hat{R}(h) \triangleq \frac{1}{N} \sum_{i=1}^{N} L(h(x_i), y_i) \quad : \text{Empirical Risk.}$$

$$\hat{h}^*(x) = \underset{h \in \mathcal{H}}{\operatorname{argmin}} \; \hat{R}(h) \qquad x_i, y_i \sim iid \; P_{xy}$$

This is Empirical Risk minimization (ERMs)

We know $\hat{R}(h) \to R(h)$: Law of Large numbers.

However $\hat{h}^*(x)$ may not converge $h^*(x)$

Central question : when would $\hat{h}^*(x) \to h^*(x)$?

Example of ERM for a Regression problem:

Given $D = \{(x_i, y_i)\}_{i=1}^{N}$ ~ iid $\mathbb{P}_{xy}$

$x_i \in \mathbb{R}^d$, $y \in \mathbb{R}$

Suppose $h_\theta(x)$ comes from a family of parametric func parameterized by θ

$h_\theta(x) = \theta^T x$ @ $h_\theta(x) = e^{\theta^T x}$

Define Loss func mean sq error

$$\hat{R}(h_\theta) = \frac{1}{N} \sum_{i=1}^{N} (h_\theta(x_i) - y_i)^2$$

$$\theta^* = \underset{\theta}{\operatorname{argmin}}\ \hat{R}(h_\theta)$$

The above is equivalent to estimating $p_\theta(y|x)$ using ML estimation with

$$p_\theta(y|x) \sim N(y; h_\theta(x), I)$$

ERM ⊂ MLE is solving same problem

Transformers : Solve ERM

- GMM : Solve MLE
- Diffusion model

We showed the equavalence in optimization Way now Lets look at its equavalence in Solution Space.

Equivalence between ERM & Divergence Minimization (MLE)

Question: what fun^c minimizes the true risk?

Consider $D = \{(x_i, y_i)\}_{i=1}^N \sim$ iid P_{xy}

$x_i \in \mathbb{R}^d$, $y \in \{0,1\}$

$$L(h(x), y) = \begin{cases} 0 & \text{if } h(x) = y \\ 1 & \text{otherwise} \end{cases}$$

$$R(h) = \mathop{\mathbb{E}}_{\mathbb{P}_{xy}} L(h(x), y)$$

Question: what h, minimizes the true risk for the above scenario?

Ans: Consider a hypothesis func^c as follows.

$$h_B(x=x) = \begin{cases} 0 & \text{if } p(y=0|x) > p(y=1|x) \\ 1 & \text{if } p(y=1|x) \geq p(y=0|x) \end{cases}$$

↑ Bayes' Classifier

Claim $R(h_B) \leq R(h) \quad \forall\ h \in \mathcal{H}$

Proof: Let $S_i(h) = \{ x \in \mathbb{R}^d : h(x) = i \}$ $i = 0, 1$

$$S_0(h) \cap S_1(h) = \phi$$

$$S_1(h) \cup S_0(h) = \mathbb{R}^d \qquad \forall\ h \in \mathcal{H}$$

Note: $R(h) = \mathbb{P}(I_{h(x) \neq y})$

$$R(h) = \mathop{\mathbb{E}}_{\mathbb{P}_{xy}} [L(h(x), y)]$$

$$= \mathbb{P}[I_{h(x) \neq y}]$$

$$= \mathbb{P}[h(x)=1,\ y=0] + \mathbb{P}[h(x)=0,\ y=1]$$

$$= \mathbb{P}[x \in S_1(h),\ y=0] + \mathbb{P}[x \in S_0(h),\ y=1]$$

$$= \mathbb{P}[y=0]\ \mathbb{P}_{x|y}[x \in S_1(h) \mid y=0] + \mathbb{P}[y=1]\ \mathbb{P}_{x|y}[x \in S_0(h) \mid y=1]$$

$$= \mathbb{P}[y=0] \int_{S_1(h)} p_{x|y=0}(x \mid y=0)\, dx + \mathbb{P}[y=1] \int_{S_0(h)} p_{x|y=1}(x \mid y=1)\, dx$$

→ for each example only one integral will be evaluated. since $S_1(h) \cap S_0(h) = \phi$.

→ $\forall\ x \in \mathbb{R}^d$ only one of two will be evaluated.

Till now its true for every hypothesis.

Now for the specific case of $h_B(x)$

we have

$$R(h) = \int_{S_1} p_{y=0}\ p_{x|y=0}\, dx + \int_{S_0} p_{y=1}\ p_{x|y=1}\, dx$$

Recall

$$h_B = \begin{cases} 1 & \text{if } p_{Y=1|X} > p_{Y=0|X} \\ & \Rightarrow p_{X|Y=1}\, p_{Y=1} > p_{X|Y=0}\, p_{Y=0} \\ 0 & \text{otherwise} \end{cases}$$

for the Bayes classifier the one with the minimal value will be evaluated.

$$R(h_B) = \int_{\mathbb{R}^d} \min\left(p_{Y=0}\, p_{X|Y=0},\ p_{Y=1}\, p_{X|Y=1}\right) dx$$

$$\therefore \quad \boxed{R(h_B) \leq R(h) \quad \forall h \in \mathcal{H}}$$

You can't design a classifier which is better than Bayes classifier.

(k-class classification same result)

Regression

Loss: Sq error loss

Optimal hypothesis is $\mathbb{E}_{P_{Y|X}}$