计算机应用数学-2025 秋

Lecture 1.6: Distance & Divergence

2025.11.14

Lecturer: 宋骐 *Scribe:* 王亦涵, 赵奇, 王馨语, 郑岭

$$\left.\begin{array}{l} Euclidean\ Distance \\ Manhattan\ Distance \\ Minkowski\ Distance \\ Cosine\ Distance \end{array}\right\} Point\ Distance$$

$$\left.\begin{array}{l} Hamming\ Distance \\ Edit\ Distance \end{array}\right\} String\ Distance$$

$$Jaccard\ Distance\ -\ Set\ Similarity$$

$$\left.\begin{array}{l} Person\ Correlation \\ Entropy\ and\ Cross\ Entropy \\ Relative\ Entropy\ and\ K-L\ Divergence \\ Wasserstein\ Distance \end{array}\right\} Distribution\ Similarity$$

1 Point Distance

Distance between objects.

Suppose the data points are from $M \subseteq \mathbb{R}^d$ or $M \subseteq \{0,1\}^d$

Metric:

$D.\ M \times M \to \mathbb{R}$ if for all $x, y, z \in M$

$\Delta\ D(x,y) = 0 \Leftrightarrow x = y$

$\Delta\ D(x,y) = D(y,x)$

$\Delta\ D(x,z) \leq D(x,y) + D(y,x) - triangle\ inequality$

Note that $D(x,y) \geq 0$

Proof: $D(x,y) + D(y,x) \geq D(x,x)$

which gives $2D(x,y) \geq D(x,x) \geq 0$ and thus $D(x,y) \geq 0$

We call D the Distance function.

D can be used in clustering, etc.

I. Euclidean Distance 欧几里德距离

$$D_{l_2}(x,y) = \|x-y\|_2 = \sqrt{\sum_{i=1}^{d}|x_i - y_i|^2}$$

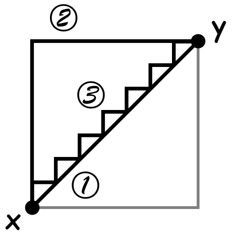
II. Manhattan Distance

$$D_{l_1}(x,y) = \sum_{i=1}^{d}|x_i - y_i|$$

III. Minkowski Distance

$$D_{l_p}(x,y) = \|x-y\|_p = (\sum_{i=1}^{d}|x_i - y_i|^p)^{\frac{1}{p}}$$

It's a generalized version of Euclidean & Manhattan Distance



① Euclidean 直线距离

② Manhattan 出租车距离

③ Equivalent Manhattan

IV. Standardized Euclidean Distance 标准化欧氏距离

Make each component have the same mean and variance.

将各个分量“标准化”到方差、均值相同的区间.

$$x^* = \frac{x-m}{s}$$

$$d(x,y) = \sqrt{\sum_{i=1}^{d}(\frac{x_i - y_i}{s_i})^2}$$

2 String Distance

I. Hamming Distance

The number of positions at which the corresponding symbols are different.

The minimum number of substitutions required to change one string into another.

$c = a\ \ XOR\ \ b.$ (a, b have the same length)

Calculate how many “1”s c have.

Widely used in Network.

II. Edit distance

Counting the minimum number of operations required to transform one string into the other.

Operators:

- Insertion of a single symbol.

- Deletion of a single symbol.

- Substitution of a single symbol x for a single symbol $y \neq x$, $u\ x\ v \rightarrow u\ y\ v$.

It is a generalized version of Hamming Distance.

	In sertion	Deletion	Substitution
Levenshtein Distance	✓	✓	✓
Lougest Common Subsequence (LCS)	✓	✓	
Hamming Distance			✓

莱文斯坦距离

$$lev(a,b) = \begin{cases} |a| & if\ |b| = 0 \\ |b| & if\ |a| = 0 \\ lev(tail(a), tail(b)) & if\ head(a) = head(b) \\ 1 + min(lev(tail(a), b), lev(a, tail(b)), lev(tail(a), tail(b))) & otherwise \end{cases}$$

3 Set Distance

Jaccard Distance and Similarity.

Jaccard Similarity $J(A,B) = \frac{|A \cap B|}{|A \cup B|}$

Jaccard Distance

$$\begin{aligned} J_S(A,B) &= 1 - J(A,B) \\ &= \frac{|A \cup B| - |A \cap B|}{|A \cup B|} \end{aligned}$$

4 Distance between variables and distributions

I. Entropy

Def 1.6.1 The (Shannon) entropy of a distance random variable X is

$$\begin{aligned} H[x] &= -\sum_x P(X=x)\log P(X=x) \quad (Discrete\ case) \\ &= -E[\log P(X)] \end{aligned}$$

Conditional Entropy of X given Y,

$$\begin{aligned} H[X|Y] &= \sum_y P(Y=y)\sum_x P(X=x|Y=y)\log P(X=x|Y=y) \\ &= -E[\log P(X|Y)] \\ &= H[X,Y]-H[Y] \end{aligned}$$

Properties of the Shannon entropy:

1. $H[x]\geq 0$
2. $H[x]=0$, if $\exists x_0 : X=x_0$
3. If x can take on $n<\infty$ different values (with positive probability), then $H[x]\leq\log n$

   $H[x]=\log n$, if x is uniformly distributed.
4. $H[X]+H[Y]\geq H[X,Y]$ with equality iff X and Y are independent.
5. $H[X,Y]\geq H[X]$
6. $H[X|Y]\geq 0$, with equality iff X is constant given Y, for almost all Y.
7. $H[X|Y]\leq H[X]$, with equality iff X is independent of Y.
8. $H[f(X)]\leq H[X]$, for any measurable function f, with equality iff f is invertible.

Lemma 1.6.1 (Chain Rule for Shannon Entropy).

Let $X_1, X_2, \cdots, X_n$ be discrete-valued random variables on a common probability space, Then

$H[X_1, X_2, \cdots, X_n] = H[X_1]+\sum_{i=2}^{n} H[X_i|X_1, X_2, \cdots, X_{i-1}]$

Proof: 作业

Def 1.6.2 (Shannon Entropy General Case).

The Shannon entropy of a random variable X with distribution μ, with respect to a reference measure ρ, is

$$H_\rho[x] = -E_\mu[\log \frac{d\mu}{d\rho}]$$

II. Cross Entropy

The cross entropy between two probability distributions p and q, over the same underlying set of events, measures the average number of bits needed to identify an event drawn from the set when the coding scheme used for the set is optimized for an estimated probability distribution q, rather than the true distribution p.

给定真实分布下，使用非真实分布分布指定策略消除系统不确定性所需付出努力的大小。

Assume p is the true distribution and q is an estimated (not true 真实) distribution.

Using p to identify an event, the average number of bits.

$$H(p) = -\sum_{i=1}^{n} p_i \log p_i$$

while using q to represent the number

$$\begin{aligned} H(p,q) &= -\textstyle\sum_{i=1}^{n} p_i \log q_i \\ &= \textstyle\sum_{i=1}^{n} p_i \log \frac{1}{q_i} \qquad \leftarrow Cross\ Entropy\ Discrete\ Case \end{aligned}$$

For continuous case

$$\begin{aligned} H(p,q) &= E_p[\log q] \\ &= -\textstyle\int_x p(x) \log q(x) dx \end{aligned}$$

Application- Cross Entropy loss function and logistic regression.

The true probability p_i is the true label, and the given distribution q_i is the predicted value of the current model.

Let’s consider a binary regression model.

In logistic regression, the probability is modeled using the logistic function $g(z) = \frac{1}{1+e^{-z}}$ where z is a linear function of input x.

The probability of the output $y = 1$ is given by

$$q_{y=1} = \hat{y} = g(w \cdot x) = \frac{1}{1+e^{-wx}}$$
$$q_{y=0} = 1 - \hat{y}$$

From the definition we have

$$p \in \{y, 1-y\}, q \in \{\hat{y}, 1-\hat{y}\}$$

we use cross entropy to measure the difference between p and q,

$$H(p,q) = -\sum_i p_i log q_i = -y log \hat{y} - (1-y) log(1-\hat{y})$$

The logistic loss is sometimes called cross-entropy loss, or log loss.

The gradient of the cross-entropy loss for logistic regression is the same as the gradient of the squared-error loss for linear regression.

$$X = \begin{pmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{pmatrix} \in R^{n \times (p+1)}$$

$$\hat{y_i} = \hat{f}(x_{i1}, \dots, x_{ip}) = \frac{1}{1 + exp(-\beta_0 - \beta_1 x_{i1} - \dots - \beta_p x_{ip})}$$

$$L(\beta) = -\sum_{i=1}^{N} [y_i log \hat{y}_i + (1-y_i) log(1-\hat{y}_i)]$$

Then

$$\frac{\partial}{\partial \beta} L(\beta) = x^T(\hat{Y} - Y)$$

Proof:

$$\frac{\partial}{\partial \beta_0} \ln \frac{1}{1+e^{-\beta_0+k_0}} = \frac{e^{-\beta_0+k_0}}{1+e^{-\beta_0+k_0}}$$

$$\frac{\partial}{\partial \beta_0} \ln(1 - \frac{1}{1+e^{-\beta_0+k_0}}) = \frac{-1}{1+e^{-\beta_0+k_0}}$$

$$\frac{\partial L(\beta)}{\partial \beta} = -\sum_{i=1}^{N} [\frac{y_i \cdot e^{-\beta_0+k_0}}{1+e^{-\beta_0+k_0}} - (1-y_i)\frac{1}{1+e^{-\beta_0+k_0}}]$$

$$= -\sum_{i=1}^{N} [y_i - \hat{y_i}] = \sum_{i=1}^{N} (\hat{y_i} - y_i)$$

$$\frac{\partial}{\partial \beta_1} \ln \frac{1}{1+e^{-\beta_1 x_{i1}+k_1}} = \frac{x_{i1} e^{k_1}}{e^{\beta_1 x_{i1}} + e^{k_1}}$$

$$\frac{\partial}{\partial \beta_1} \ln[1 - \frac{1}{1+e^{-\beta_1 x_{i1}+k_1}}] = \frac{-x_{i1} e^{\beta_1 x_{i1}}}{e^{\beta_1 x_{i1}} + e^{k_1}}$$

$$\frac{\partial}{\partial \beta_1} L(\beta) = -\sum_{i=1}^{N} x_{i1}(y_i - \hat{y_i}) = \sum_{i=1}^{N} x_{i1}(\hat{y_i} - y_i)$$

III. Relative Entropy or K-L Divergence

Given two probabilistic distributions P and Q,

$$D_{KL}(P||Q) = \sum_{x \in \mathcal{X}} p(x) log(\frac{P(x)}{Q(x)}).$$

Note: its not symmetric!

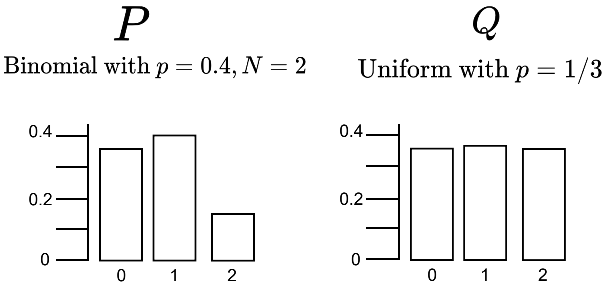
For continuous case

$$D_{KL}(P||Q) = \int_{-\infty}^{\infty} p(x) log(\frac{p(x)}{q(x)}) d(x).$$

More generally, if P and Q are probability measures on a measurable space $\mathcal{X}$, and P is absolutely continuous with respect to Q, then

$$D_{KL}(P||Q) = \int_{x \in \mathcal{X}} log(\frac{P(dx)}{Q(dx)}) P(dx).$$

Example:



P

Binomial with $p = 0.4, N = 2$

Q

Uniform with $p = 1/3$

$\mathcal{X}$	0	1	2
$P(x)$	$\frac{9}{25}$	$\frac{12}{25}$	$\frac{4}{25}$
$Q(x)$	$1/3$	$1/3$	$1/3$

$$
\begin{aligned}
D_{KL}(P||Q) &= \sum_{x\in\mathcal{X}} P(x)ln(\frac{P(x)}{Q(x)}) \\
&= \frac{9}{25}ln(\frac{9/25}{1/3}) + \frac{12}{25}ln(\frac{12/25}{1/3}) + \frac{4}{25}ln(\frac{4/25}{1/3}) \\
&= \frac{1}{25}(32ln2 + 55ln3 - 50ln5) \\
&= 0.0853 \\
D_{KL}(Q||P) &= \sum_{x\in\mathcal{X}} Q(x)ln(\frac{Q(x)}{P(x)}) \\
&= \frac{1}{3}ln(\frac{1/3}{9/25}) + \frac{1}{3}ln(\frac{1/3}{12/25}) + \frac{1}{3}ln(\frac{1/3}{4/25}) \\
&= 0.0975
\end{aligned}
$$

Application-Bayesian Updating

KL divergence can be used to measure the information gain in moving from a prior distribution to a posterior distribution $P(x) \to p(x|I)$.

If some new fact $Y = y$ is discovered, it can be used to update the posterior distribution for $\mathcal{X}$ from p(xlI) to a new posterior distribution p(xly,I) using Bayes' theorem

$$p(x|y,I) = \frac{p(y|x,I)p(x|I)}{p(y|I)}$$

$$D_{KL}(p(x|y,I)||p(x|I)) = \sum_{x} p(x|y,I)log(\frac{p(x|y,I)}{p(x|I)})$$

IV. Jensen-Shannon (JS) divergence

It is a symmetrized and smoothed version of KL.

$$D_{JS} = \frac{1}{2}KL(p||\frac{p+q}{2}) + \frac{1}{2}KL(q||\frac{p+q}{2})$$

Application: Generative Adversarial Nets. To learn the generator's distribution p_g over data x, we define a prior on input noise variables $p_z(Z)$. represent a mapping to data space as $G(z;\theta_g)$, where G is a differentiable function represented by a multilayer porception with parameters θ_g.

We define a second multilayer perceptron $D(X;\theta d)$. Train D to maximize the probability of assigning the correct label to both training examples and samples from G. We simultaneously train G to minimize $log(1 - D(G(z))$. In other words, D and G play a two-player minimax game with value function $V(G,D)$:

$\min_G \max_D(D,G) = \underset{X \sim P_{data}(X)}{E}[logD(X)] + \underset{z \sim p_z(z)}{E}[log(1 - D(G(z)))]$

We first consider the optimal D for any given G.

Lemma 1.6.2 For G fixed, the optimal D is

$$D^*_G(x) = \frac{p_{data}(x)}{p_{data}(x) + p_g(x)}$$

Proof: The training criterion for D given any G, is to maximize the quantity

$$\begin{aligned} V(G,D) &= \int_x p_{data}(x)log(D(x))dx + \int_z p_z(z)log(1 - D(g(z)))dz \\ &= \int_x p_{data}(x)log(D(x)) + p_{g(x)}log(1 - D(x))dx \end{aligned}$$

For any $(a,b) \in \mathbb{R}^2 \setminus \{0,0\}$, $y \to alogy + blog(1-y)$ achieves the maximum in $[0,1]$ at $\frac{a}{a+b}$

The training objective for D can be interpreted as maximizing the log-likelihood for estimating the conditional probability $P(Y = y|x)$, where Y indicates whether x comes from p_{data} (with $y = 1$) or from p_g(with $y = 0$).

The minimax becomes:

$$\begin{aligned} C(G) &= \max_D V(G,D) \\ &= \underset{x \sim p_{data}}{E}[\log D^*_G(x)] + \underset{z \sim p_z}{E}[\log(1 - D^*_G(G_z))] \\ &= \underset{x \sim p_{data}}{E}[\log D^*_G(x)] + \underset{x \sim p_g}{E}[\log(1 - D^*_G(x))] \\ &= \underset{x \sim p_{data}}{E}[\log \frac{p_{data}(x)}{p_{data}(x) + p_g(x)}] + \underset{x \sim p_g}{E}[\log \frac{p_g(x)}{p_{data}(x) + p_g(x)}] \end{aligned}$$

Theorem 1.6.3

The global minimum of $C(G)$ is achieved iff $p_g = P_{data}$. At this point, $c(G)$ achieves value $-\log 4$.

Proof: For $p_g = P_{data}$, $D^*_G(x) = \frac{1}{2}$

Hence, we find $C(G) = \log \frac{1}{2} + \log \frac{1}{2} = -\log 4$.

To see that this is the best possible value of $C(G)$, reached only for $p_g = P_{data}$, observe that

$$\underset{x \sim p_{data}}{E}[-\log 2] + \underset{x \sim p_g}{E}[-\log 2] = -\log 4.$$

by substracting this expression from $C(G) = V(D^*_G, G)$, we obtain

$$
\begin{aligned}
C(G) &= -\log(4) + D_{KL}\left(p_{data}||\frac{p_{p_{data}} + p_g}{2}\right) + D_{KL}\left(p_g||\frac{p_{p_{data}} + p_g}{2}\right) \\
&= -\log(4) + 2 \cdot D_{JS}\left(p_{data}||p_g\right)
\end{aligned}
$$

Since JS divergence is always non-negative and zero only when they are equal, we show $c^* = -\log(4)$ and the only solution is $p_g = P_{data}$.

V. Wasserstein Distance. （推土机距离）

If p and q are very dissimilar. i.e. far away from each other, and have no overlap, then their KL-divergence is meaningless, and $J-S$-divergence is a constant, thus the gradient becomes 0 .

$$
w(p, q) = \inf_{\gamma \in \Gamma(u,v)} \left(E_{(x,y)\gamma)} d(x, y)^p\right)^{1/p}
$$

where $\Gamma(u, v)$ is the set of all couplings of u and v, $W_\infty(u, v)$ is defined to be $\lim_{p \to +\infty} W_p(u, v)$

W-Distance can also be used to compare discrete and continuous distributions.

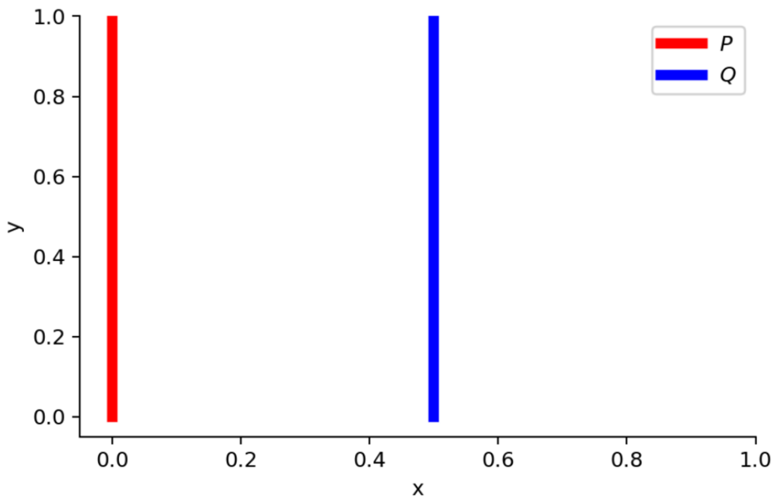
Application: Wasserstein GAN.

Why Wasserstein Distance is better than JS or KL?

Suppose we have two probability distributions P, Q.

$\forall(x, y) \in P, x = 0, y \sim U(0, 1)$

$\forall(x, y) \in Q, x = \theta, 0 \leq \theta \leq 1$ and $y \sim U(0, 1)$



P, Q have no overlap

When $\theta \neq 0$:

$$
\begin{aligned}
D_{\text{KL}}(P||Q) &= \sum_{x=0;y\sim U(0,1)} 1 \cdot \log\frac{1}{0} = +\infty \\
D_{\text{KL}}(Q||P) &= \sum_{x=\theta;y\sim U(0,1)} 1 \cdot \log\frac{1}{0} = +\infty \\
D_{\text{JS}}(P,Q) &= \frac{1}{2}\left(\sum_{x=0;y\sim U(0,1)} 1 \cdot \log\frac{1}{\frac{1}{2}} + \sum_{x=0;y\sim U(0,1)} 1 \cdot \log\frac{1}{\frac{1}{2}}\right) = \log 2 \\
W(P,Q) &= |\theta|
\end{aligned}
$$

When $\theta = 0$, P, Q are fully overlapped:

$$
\begin{aligned}
D_{\text{KL}}(P||Q) = D_{\text{KL}}(Q||P) = D_{\text{JS}}(P,Q) &= 0 \\
W(P,Q) = 0 &= |\theta|
\end{aligned}
$$

Only W provides a smooth measure.

Use W-distance as GAN loss function.

It is intractable to exhaust all the possible joint distributions in $\Pi(p_r, p_g)$ to compute $\inf_{\gamma\sim\Pi(p_r,p_g)}$.

Based on the Kantorovich-Rubinstein duality

$$
W(p_r, p_g) = \frac{1}{K} \sup_{||f||_L \leq K} \left(\underset{x\sim p_r}{E}[f(x)] - \underset{x\sim p_g}{E}[f(x)] \right)
$$

where $\sup$ is the opposite of $\inf$.

Now we want to measure the least upper bound (maximum value).