四川大學
SICHUAN UNIVERSITY

Information Theory

王治国

四川大学数学学院

wangzhiguo@scu.edu.cn

1. Probability Distributions

2. Properties of Common Functions

3. Entropy and Kullback-Leibler divergence

Common Probability Distributions

- Uniform distribution

$$P(x = x_i) = \frac{1}{k}, \qquad i = 1, \dots, k$$

- The Bernoulli distribution is a discrete distribution having two possible outcomes

$$\left.\begin{array}{r} P(\mathrm{x} = 1) = \phi \\ P(\mathrm{x} = 0) = 1 - \phi \end{array}\right\} \boxed{P(\mathrm{x} = x) = \phi^{x}(1-\phi)^{1-x}}$$

$$E_{\mathrm{x}}[\mathrm{x}] = \phi, \ \ \mathrm{Var}_{\mathrm{x}}(\mathrm{x}) = \phi(1-\phi)$$

- Multinoulli distribution (categorical distribution) is a generalization of the Bernoulli distribution. If you perform an experiment that can have K outcomes, its joint probability mass function is

$$p_X(x_1, \dots, x_K) = \begin{cases} \prod_{j=1}^{K} \phi_j^{x_j} & \text{if } (x_1, \dots, x_K) \in R_X \\ 0 & \text{otherwise} \end{cases}$$

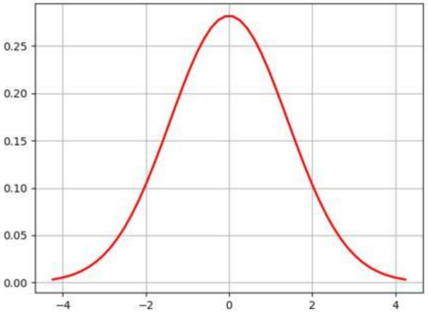
one-hot vector

where $R_X = \left\{ x \in \{0,1\}^K : \sum_{j=1}^{K} x_j = 1 \right\}$, $p_1, \dots, p_K$ be K strictly positive numbers such that $\sum_{j=1}^{K} \phi_j = 1$

Common Probability Distributions

- The Gaussian distribution (normal distribution)

$$\mathcal{N}\left(x;\mu,\sigma^2\right)=\sqrt{\frac{1}{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$



- Multivariate normal distribution toR^n

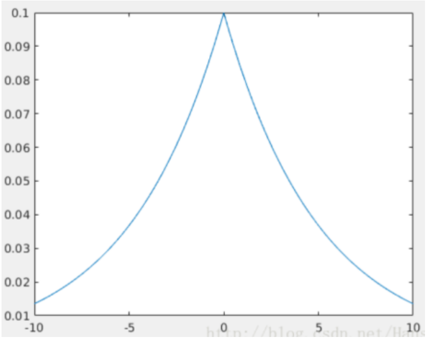
$$\mathcal{N}(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma})=\sqrt{\frac{1}{(2\pi)^n\det(\boldsymbol{\Sigma})}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

- Exponential distribution

$$p(x;\lambda)=\lambda\mathbf{1}_{x\geq 0}\exp(-\lambda x)$$

- Laplace distribution ($\gamma > 0$)

$$\text{Laplace}(x;\mu,\gamma)=\frac{1}{2\gamma}\exp\left(-\frac{|x-\mu|}{\gamma}\right)$$



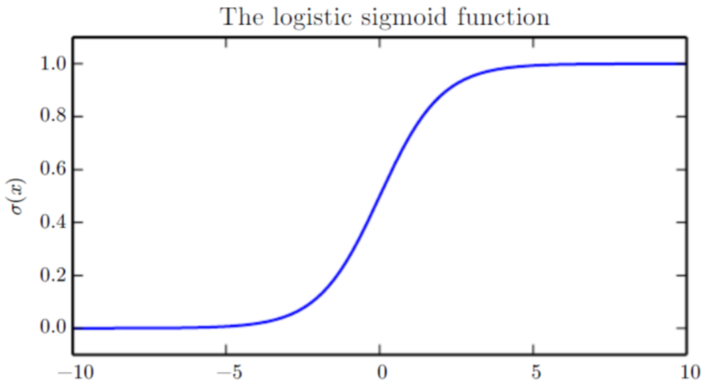
2 Properties of Common Functions

Useful Properties of Common Functions

- Logistic sigmoid function

$$\sigma(x) = \frac{1}{1+\exp(-x)}$$

- Logistic sigmoid function is used to produce the ϕ parameter of a Bernoulli distribution because its range is (0,1)



- Logistic sigmoid function saturates when its argument is very positive or very negative, meaning that the function becomes very flat and insensitive to small changes in its input.

$$1-\sigma(x) = \sigma(-x)$$

$$\frac{d}{dx}\sigma(x) = \sigma(x)(1-\sigma(x))$$

The gradient becomes 0 when x is very positive or very negative

Useful Properties of Common Functions

- Softplus function

$$\zeta(x) = \log(1 + \exp(x))$$

- The name of the softplus function comes from the fact that it is a smoothed or "softened" version of (rectified linear unit)

$$x^{+} = \max(0, x)$$

$$\log \sigma(x) = -\zeta(-x)$$

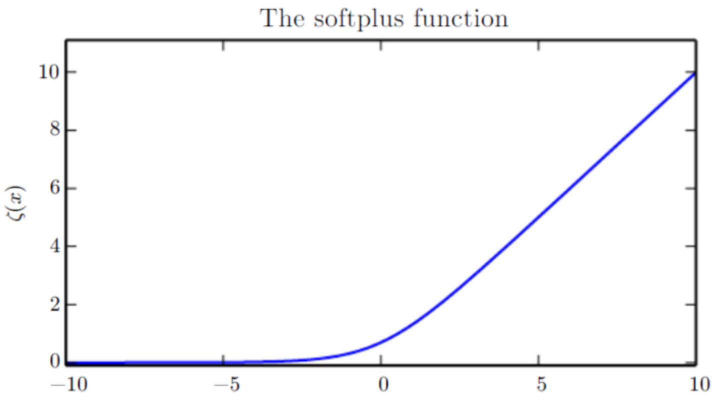
$$\zeta(x) - \zeta(-x) = x$$

$$\frac{d}{dx}\zeta(x) = \sigma(x)$$

$$\zeta(x) = \int_{-\infty}^{x} \sigma(y) dy$$

$$\forall x \in (0,1), \sigma^{-1}(x) = \log\left(\frac{x}{1-x}\right)$$

$$\forall x > 0, \zeta^{-1}(x) = \log(\exp(x) - 1)$$



Softmax function

- The softmax function is often used to predict the probabilities associated with a multinoulli distribution. The softmax function is defined to be

$$\operatorname{softmax}(\boldsymbol{x})_i = \frac{\exp(x_i)}{\sum_{j=1}^{n} \exp(x_j)}$$

$$x = \begin{bmatrix} 3 \\ 0 \end{bmatrix} \qquad \begin{aligned} \operatorname{softmax}(\boldsymbol{x})_1 &= \frac{\exp(3)}{\exp(3)+\exp(0)} = 0.953 \\ \operatorname{softmax}(\boldsymbol{x})_2 &= \frac{\exp(0)}{\exp(3)+\exp(0)} = 0.047 \end{aligned}$$

- The softmax function is rather a smooth approximation to the argmax function

$$\operatorname{argmax}(\boldsymbol{x})_1 = 1 \qquad \operatorname{argmax}(\boldsymbol{x})_2 = 0$$

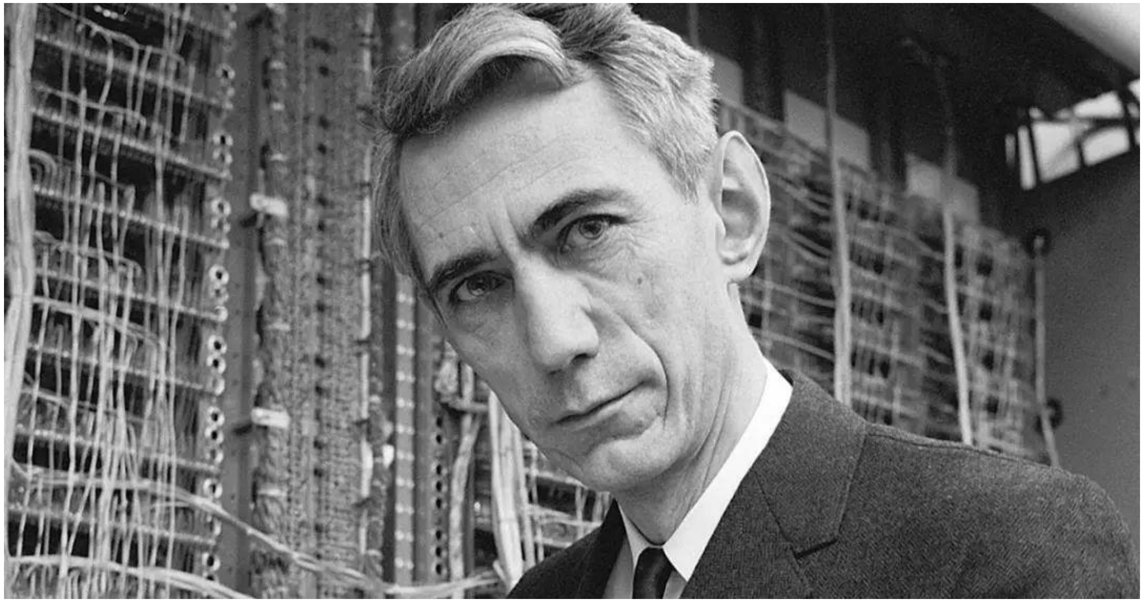
- Softmax function has a small problem: overflow or underflow

By instead evaluating softmax $(\boldsymbol{z})$ where $\boldsymbol{z} = \boldsymbol{x} - \max_i x_i$.

- The gradient of softmax function?

If $\boldsymbol{y} = \operatorname{softmax}(\boldsymbol{x}),\ \boldsymbol{x}, \boldsymbol{y} \in R^n$. Solve $\frac{\nabla \boldsymbol{y}}{\nabla \boldsymbol{x}}$

3 Entropy and Kullback-Leibler divergence



信息论之父克劳德 · 香农 (1916-2001)

$$C = B\log_2(1+\frac{S}{N})$$

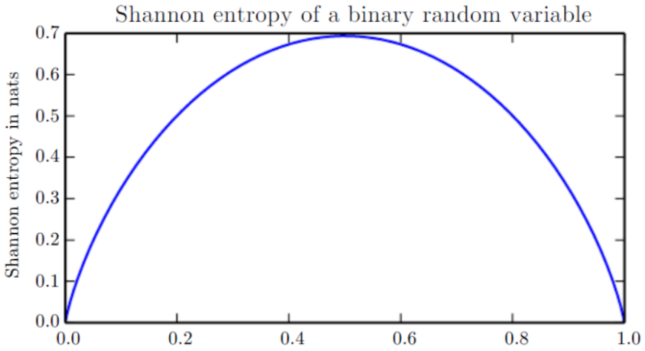
https://www.bilibili.com/video/BV1YV411z7qo?from=search&seid=12422131076821505031

Shannon entropy

- Quantify information
  - 非负
  - 连续依赖概率吧
  - 独立事件的信息可加
  - 信息量大小可能跟结果数量有关系

- Shannon entropy (differential entropy)

$$H(x) = -E_{x\sim P}[\log P(x)]$$



Example (Bernoulli random variables): Let $f(p) = -p\log p - (1-p)\log(1-p)$ denote the binary entropy, which is the entropy of a Bernoulli (p) random variable.

Entropy of Gaussian

Example (Entropy of normal random variables): The entropy of a normal random variable is straightforward to compute. Indeed, for $X \sim \mathrm{N}\left(\mu, \sigma^2\right)$ we have $p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$, so that

$$H(X) = -\int p(x)\left[\frac{1}{2}\log\frac{1}{2\pi\sigma^2} - \frac{1}{2\sigma^2}(x-\mu)^2\right] = \frac{1}{2}\log\left(2\pi\sigma^2\right) + \frac{E\left[(X-\mu)^2\right]}{2\sigma^2} = \frac{1}{2}\log\left(2\pi e\sigma^2\right)$$

For a general multivariate Gaussian, where $X \sim N(\mu, \Sigma)$ for a vector $\mu \in R^n$ and $\Sigma \succ 0$ with density $p(x) = \frac{1}{(2\pi)^{n/2}\sqrt{\det(\Sigma)}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right)$, we similarly have

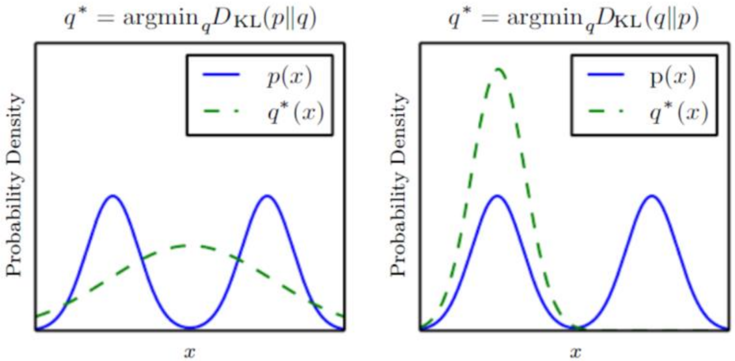
$$\begin{aligned} H(X) &= \frac{1}{2} E\left[n \log(2\pi) + \log\det(\Sigma) + (X-\mu)^\top \Sigma^{-1}(X-\mu)\right] \\ &= \frac{n}{2}\log(2\pi) + \frac{1}{2}\log\det(\Sigma) + \frac{1}{2}\operatorname{tr}\left(\Sigma\Sigma^{-1}\right) = \frac{n}{2}\log(2\pi e) + \frac{1}{2}\log\det(\Sigma) \end{aligned}$$

Kullback-Leibler (KL) divergence

- Kullback-Leibler (KL) divergence can measure how different these two distributions

$$D_{KL}(P\|Q) = E_{x\sim P}\left[\log\frac{P(x)}{Q(x)}\right] = E_{x\sim P}[\log P(x) - \log Q(x)]$$

- it is not symmetric: $D_{\mathrm{KL}}(P\|Q) \neq D_{\mathrm{KL}}(Q\|P)$ for some P and Q



Example. Let $\mathcal{X} = \{0, 1\}$ and consider two distributions p and q on $\mathcal{X}$. Let $p(0) = 1 - r, p(1) = r$, and let $q(0) = 1 - s, q(1) = s$. Then

$$D(p\|q) = (1-r)\log\frac{1-r}{1-s} + r\log\frac{r}{s}$$

$$D(q\|p) = (1-s)\log\frac{1-s}{1-r} + s\log\frac{s}{r}$$

If $r = s$, then $D(p\|q) = D(q\|p) = 0$.

If $r = \frac{1}{2}, s = \frac{1}{4}$, we can calculate

$$D(p\|q) = 1 - \frac{1}{2}\log 3 = 0.2075,$$

$$D(q\|p) = \frac{3}{4}\log 3 - 1 = 0.1887$$

Kullback-Leibler (KL) divergence

Recall the definition of a convex function $f : R^k \to R$ as any function satisfying

$$\boxed{f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y)}$$

for all $\lambda \in [0, 1]$, all x, y. The function f is strictly convex if the convexity inequality is strict for $\lambda \in (0, 1)$ and $x \neq y$.

Proposition (Jensen's inequality). Let f be convex. Then for any random variable X,

$$f(E[X]) \leq E[f(X)]$$

$$D_{\mathrm{KL}}(P\|Q) = -E\left[\log \frac{Q(x)}{P(x)}\right] \geq -\log E\left[\frac{Q(x)}{P(x)}\right]$$

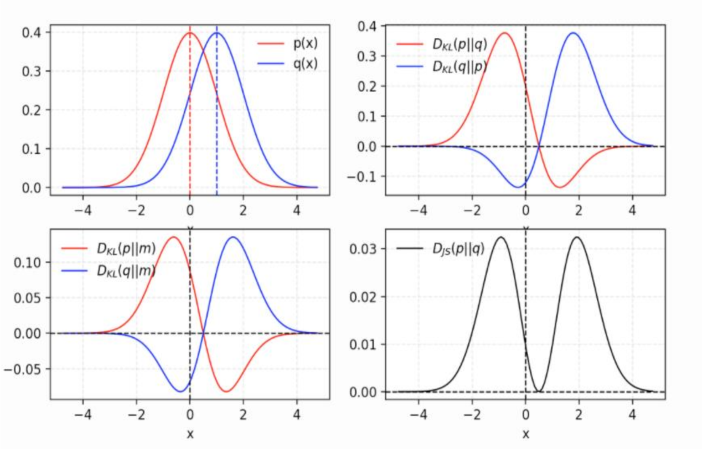
$$= -\log\left(\sum_x P(x)\frac{Q(x)}{P(x)}\right) = -\log(1) = 0$$

Example (Divergence between Gaussian distributions): Let P be the multivariate normal $\mathrm{N}(\mu_1, \Sigma)$, and Q be the multivariate normal distribution with mean μ_2 and identical covariance $\Sigma \succ 0$. Then we have that

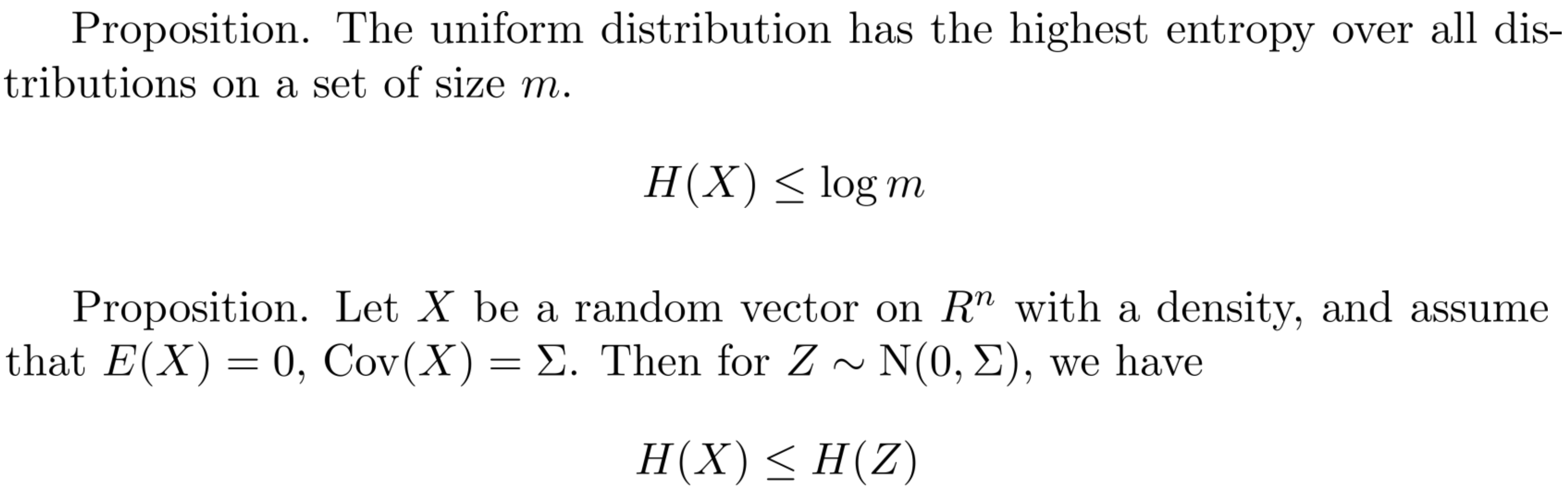
$$D_{\mathrm{KL}}(P \| Q) = \frac{1}{2} (\mu_1 - \mu_2)^\top \Sigma^{-1} (\mu_1 - \mu_2)$$

Jensen-Shannon Divergence is a smoother and symmetric version of measuring the similarity between two probability distributions and it is bounded by [0,1] .

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



Proposition



Proposition. The uniform distribution has the highest entropy over all distributions on a set of size m.

$$H(X) \leq \log m$$

Proposition. Let X be a random vector on R^n with a density, and assume that $E(X) = 0$, $\text{Cov}(X) = \Sigma$. Then for $Z \sim \text{N}(0, \Sigma)$, we have

$$H(X) \leq H(Z)$$

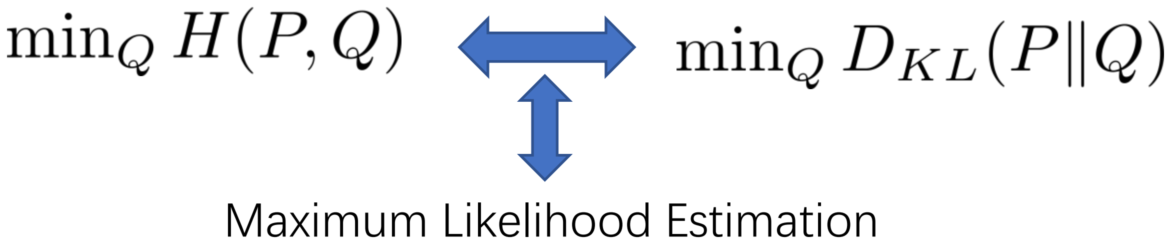
□ differential entropy for random variables with second moments is always maximized by the Gaussian distribution.

Cross-entropy

A quantity that is closely related to the KL divergence is the cross-entropy $H(P, Q) = H(P) + D_{\mathrm{KL}}(P\|Q)$, which is similar to the KL divergence but lacking the term on the left:

$$H(P, Q) = -E_{x\sim P} \log Q(x)$$

Minimizing the cross-entropy with respect to Q is equivalent to minimizing the KL divergence, because Q does not participate in the omitted term.



$$\min_Q H(P, Q) \Longleftrightarrow \min_Q D_{KL}(P\|Q)$$

$$\Updownarrow$$

Maximum Likelihood Estimation

https://adityashrm21.github.io/Information-Theory-In-Deep-Learning/

Maximum Likelihood Estimation

Consider a set of m examples $X = \{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}\}$ drawn independently from the true but unknown data generating distribution $p_{\text{data}}(\mathbf{x})$.

Let $p_{\text{model}}(\mathbf{x};\boldsymbol{\theta})$ be a parametric family of probability distributions over the same space indexed by $\boldsymbol{\theta}$. In other words, $p_{\text{model}}(\boldsymbol{x};\boldsymbol{\theta})$ maps any configuration $\boldsymbol{x}$ to a real number estimating the true probability $p_{\text{data}}(\boldsymbol{x})$.

The maximum likelihood estimator for $\boldsymbol{\theta}$ is then defined as

$$\boldsymbol{\theta}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\theta}} p_{\text{model}}(X;\boldsymbol{\theta})$$

$$= \arg\max_{\boldsymbol{\theta}} \prod_{i=1}^{m} p_{\text{model}}\left(\boldsymbol{x}^{(i)};\boldsymbol{\theta}\right)$$

$$= \arg\max_{\boldsymbol{\theta}} \textstyle\sum_{i=1}^{m} \log p_{\text{model}}\left(\boldsymbol{x}^{(i)};\boldsymbol{\theta}\right)$$

$$= \arg\max_{\boldsymbol{\theta}} E_{\mathbf{x}\sim\hat{p}_{\text{data}}} \log p_{\text{model}}(\boldsymbol{x};\boldsymbol{\theta})$$

$$\Updownarrow$$

$$\arg\min_{\boldsymbol{\theta}} D_{\mathrm{KL}}\left(\hat{p}_{\text{data}} \| p_{\text{model}\ (\boldsymbol{x};\theta)}\right) = E_{\mathbf{x}\sim\hat{p}_{\text{data}}}\left[\log \hat{p}_{\text{data}}(\boldsymbol{x}) - \log p_{\text{model}}(\boldsymbol{x};\theta)\right]$$

Conditional Entropy and mutual information

- □ The conditional entropy (also called conditional uncertainty) of a random variable X given a random variable Y is the average conditional entropy over Y

$$H(X|Y) = E_Y[H(X|y)] = -\sum_{y} p(y) \sum_{x} p(x|y)\log(p(x|y)) = -\sum_{x,y} p(x,y)\log(p(x|y))$$

- □ The mutual information of X relative to Y is given by

$$I(X;Y) = \sum_{x,y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}$$

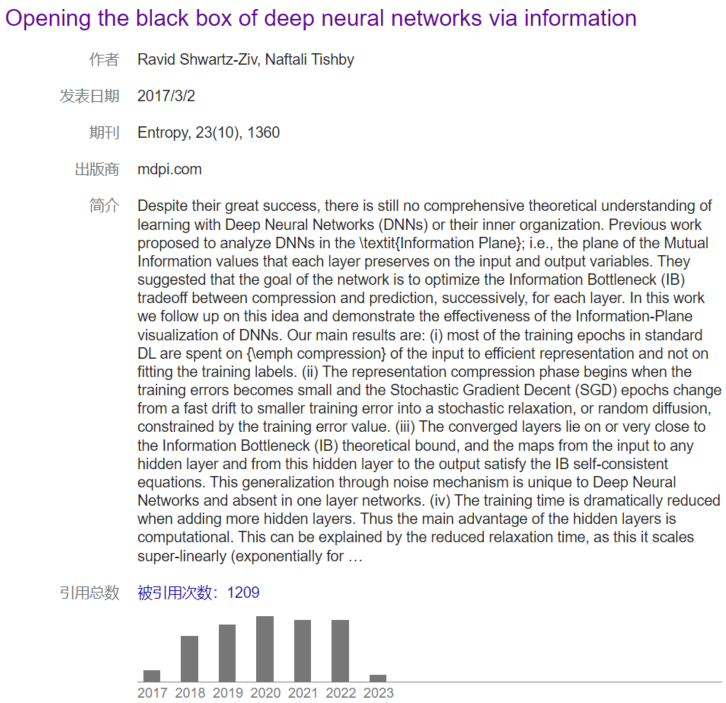
- □ A basic property of the mutual information is that

$$I(X;Y) = H(X) - H(X|Y)$$

- □ Assume we have the Markov chain $X \rightarrow Y \rightarrow Z$, Then we obtain the classical data processing inequality

$$I(X;Z) \leq I(X;Y)$$

课外阅读



Opening the black box of deep neural networks via information

作者 Ravid Shwartz-Ziv, Naftali Tishby

发表日期 2017/3/2

期刊 Entropy, 23(10), 1360

出版商 mdpi.com

简介 Despite their great success, there is still no comprehensive theoretical understanding of learning with Deep Neural Networks (DNNs) or their inner organization. Previous work proposed to analyze DNNs in the \textit{Information Plane}; i.e., the plane of the Mutual Information values that each layer preserves on the input and output variables. They suggested that the goal of the network is to optimize the Information Bottleneck (IB) tradeoff between compression and prediction, successively, for each layer. In this work we follow up on this idea and demonstrate the effectiveness of the Information-Plane visualization of DNNs. Our main results are: (i) most of the training epochs in standard DL are spent on {\emph compression} of the input to efficient representation and not on fitting the training labels. (ii) The representation compression phase begins when the training errors becomes small and the Stochastic Gradient Decent (SGD) epochs change from a fast drift to smaller training error into a stochastic relaxation, or random diffusion, constrained by the training error value. (iii) The converged layers lie on or very close to the Information Bottleneck (IB) theoretical bound, and the maps from the input to any hidden layer and from this hidden layer to the output satisfy the IB self-consistent equations. This generalization through noise mechanism is unique to Deep Neural Networks and absent in one layer networks. (iv) The training time is dramatically reduced when adding more hidden layers. Thus the main advantage of the hidden layers is computational. This can be explained by the reduced relaxation time, as this it scales super-linearly (exponentially for …

引用总数 被引用次数：1209

Geoffrey Hinton 曾对此研究评论道：信息瓶颈极其有趣，估计要再听 10000 遍才能真正理解它，当今能听到如此原创的想法非常难得，或许它就是解开谜题的那把钥匙。

- Elements of information theory, Thomas M. Cover, Joy A. Thomas