Statistic

for machine learning

Tran Trong Khiem

AI lab tranning

2024/05/29

1. Entropy
2. Relative entropy (KL divergence)
3. Mutual information

Entropy

Entropy

- The **entropy** of a **probability distribution** measure of **uncertainty**, or **lack of predictability**.
- E.g : $X_n \sim p$ generated from distribution p.
  - If p has **high entropy, hard to predict** the value of each X_n
  - entropy is zezo, all X_n are the same

Entropy for discrete random variables

- The **entropy** of a **discrete random variable** X with distribution p over K state :

$$H(X) = -\sum_{k=1}^{K} p(X=k)\log_2 p(X=k) = -\mathbb{E}_X[\log_2 p(X)]$$

- Discrete distribution with **maximum entropy** is the **uniform distribution**.
- Distribution with **minimum entropy** is any delta-function.

Entropy

Cross entropy

- The **cross entropy** between distribution p and q is defined by :

$$H_{ce}(p,q) \hat{=} -\sum_{k=1}^{K} p_k log(q_k)$$

- be minimized by setting q = p.

Joint entropy

- The **joint entropy** of two random variables X and Y is defined as :

$$H(x,y) = -\sum_{x,y} p(x,y) \log_2(x,y)$$

- If X and Y are **independent** : $H(x,y) = H(x) + H(y)$
- $H(X,Y) \geq \max\{H(X), H(Y)\} \geq 0$

Conditional entropy

Conditional entropy of Y given X:

- is the **uncertainty we have in Y** after **seeing X**, **averaged** over possible values for X :

$$H(Y \mid X) = \mathbb{E}_{p(X)}[H(p(Y \mid X))] = H(X,Y) - H(X)$$

- If Y is a **deterministic function** of X, then $H(Y|X) = 0$
- $H(Y|X) \leq H(Y)$, with equality **iff X and Y are independent**.
- The **chain rule for entropy** is given by:

$$H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i \mid X_1, X_2, \ldots, X_{i-1})$$

Perplexity

- defined as : $\text{perplexity}(p) = 2^{H(p)}$
- often used to **evaluate the quality** of **statistical language** models.

Differential entropy for continuous random variables

Differential entropy

- X is a **continuous** random variable with **pdf p(x)**, **the differential entropy** is given as :

$$h(X) \hat{=} -\int_{\mathcal{X}} p(x)\log(p(x))dx$$

- **Differential entropy** can be **negative**.
- describe X to n bits of accuracy only requires $n-3$ bits, $h(X)=-3$

Connection with variance

- The entropy of a Gaussian **increases monotonically** as the variance increases.

Discretization

- **Problem** : computing the differential entropy can be **difficult**.
- Using the heuristic: $B = N^{1/3}\frac{\max(D)-\min(D)}{3.5\sigma(D)}$

1. Entropy
2. Relative entropy (KL divergence)
3. Mutual information

Relative entropy (KL divergence)

Definition

- **Discrete distributions**, the KL divergence is defined as follows:

$$\mathcal{D}_{KL} = \sum_{k=1}^{K} p_k \log(\frac{p_k}{q_k})$$

- **Continuous distributions** as well : $\mathcal{D}_{KL} = \int p(x) \log(\frac{p(x)}{q(x)})dx$
- **Interpretation** :

$$D_{\mathrm{KL}}(p\|q) = \sum_{k=1}^{K} p_k \log \frac{p_k}{q_k} = \sum_{k=1}^{K} p_k \log p_k - \left(\sum_{k=1}^{K} p_k \log q_k\right)$$

$$= -H(p) + H_{ce}(p, q)$$

Non-negativity of KL

Theorem 6.2.1: $D_{\mathrm{KL}}(p\|q) \geq 0$ with equality if and only if $p = q$.

- since $-\log(x)$ is convex, we have
  $:\sum p(x)\log(\frac{q(x)}{p(x)}) \leq \log \sum p(x)\frac{q(x)}{p(x)}$
-

$$-D_{KL}(p \parallel q) = -\sum_{x \in A} p(x) \log \frac{p(x)}{q(x)} = \sum_{x \in A} p(x) \log(\frac{q(x)}{p(x)})$$

$$\leq \log \sum_{x \in A} p(x) \frac{q(x)}{p(x)} = \log \sum_{x \in A} q(x) = log(1) = 0$$

Corollary 6.2.1: (Uniform distribution maximizes the entropy) $H(X) \leq \log |X|$, where $|X|$ is the number of states for X, with equality if and only if $p(x)$ is uniform.

KL divergence and MLE

- **Goal**: find the distribution q that is as **close as possible to** p.

$$q^* = \arg\min_q D_{KL}(p \parallel q) = \arg\min_q \left(\int p(x) \log p(x)\, dx - \int p(x) \log q(x)\, dx \right)$$

- p is the empirical distribution: $p_D(x) = \frac{1}{N} \sum_{n=1}^{N} \delta(x - x_n)$
- Sifting property of delta functions we get:

$$\mathcal{D}_{KL}(p_D || q) = C - \int p_D(x) \log(q(x)) dx$$

$$= -\int [\frac{1}{N} \sum_{n=1}^{N} \delta(x - x_n)] \log(q(x)) dx + C = C - \frac{1}{N} \sum_{n} \log(q(x_n))$$

- Where $C = \int p(x) \log(p(x)) dx$ is a constant independent of q.

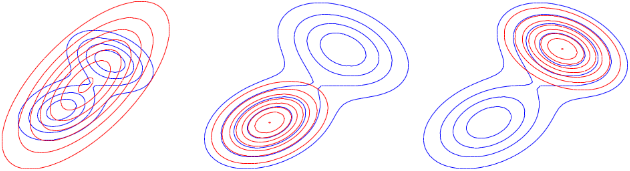
Forward vs reverse KL

forwards KL: defined by $\mathcal{D}_{KL}(p||q) = \int p(x)\log(\frac{p(x)}{q(x)})$

- Minimize by $q(x) > 0$ where $p(x) > 0$

reverse KL: defined by $\mathcal{D}_{KL}(q||p) = \int q(x)\log(\frac{q(x)}{p(x)})$

- Minimize by $q(x) = 0$ where $p(x) = 0$



(a) (b) (c)

1. Entropy
2. Relative entropy (KL divergence)
3. Mutual information

Mutual information

Definition:

- The mutual information between X and Y is defined as :

$$I(X;Y) = D_{KL}\left(p(x,y) \parallel p(x)p(y)\right) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

- MI is always **non-negative**, even for **continuous** rv, achieve 0 iff $p(x,y) = p(x)p(y)$.

Interpretation:

- $I(X;Y) = H(X) - H(X \mid Y) = H(Y) - H(Y \mid X)$
- $I(X;Y) = H(X,Y) - H(X|Y) - H(Y|X)$
- $I(X;Y) = H(X) + H(Y) - H(X,Y)$

Conditional mutual information

Define the **conditional mutual information** as :

$$I(X;Y|Z) = \mathbb{E}_{p(Z)}\left[I(X;Y)|Z\right] = \mathbb{E}_{p(x,y,z)}\left[\log \frac{p(x)p(y|z)}{p(x|z)p(y|z)}\right]$$

$$= H(X \mid Z) + H(Y \mid Z) - H(X,Y \mid Z) = H(X \mid Z) - H(X \mid Y,Z)$$

$$= H(Y \mid Z) - H(Y \mid X,Z) = H(X,Z) + H(Y,Z) - H(Z) - H(X,Y,Z)$$

$$= I(Y;X,Z) - I(Y;Z)$$

We can rewrite as : $I(Z,Y;X) = I(Z;X) + I(Y;X|Z)$

- we get the **chain rule for mutual information** :

$$I(Z_1,\ldots,Z_N;X) = \sum_{n=1}^{N} I(Z_n;X \mid Z_1,\ldots,Z_{n-1})$$

MI as a "generalized correlation coefficient"

Suppose that (x, y) are jointly Gaussian:

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{pmatrix}\right)$$

We find that the **entropy** is :

$$h(X, Y) = \frac{1}{2}\log\left((2\pi e)^2 \det \Sigma\right) = \frac{1}{2}\log\left((2\pi e)^2\sigma^4(1-\rho^2)\right)$$

Since X and Y are **individually normal with variance** σ^2 :

$$h(X) = h(Y) = \frac{1}{2}\log(2\pi e\sigma^2)$$

Hence,

$$I(X, Y) = h(X) + h(Y) - h(X, Y) = -\frac{1}{2}\log[1-\rho^2]$$

Normalized mutual information

Note that:

- $I(X;Y) = H(X) - H(X|Y) \leq H(X)$
- $I(X;Y) = H(Y) - H(Y|X) \leq H(Y)$
- $0 \leq I(X;Y) \leq \min(H(X), H(Y))$

Define the **normalized mutual information** as follows:

- $\text{NMI}(X,Y) = \frac{I(X;Y)}{\min(H(X),H(Y))} \leq 1$
- If $NMI(X,Y) = 0$, then X and Y are independent.
- If $NMI(X,Y) = 1$, then :
  - If $H(X) < H(Y)$ then $H(X|Y) = 0$, X is a deterministic function of Y
  - If $H(Y) < H(X)$ then $H(Y|X) = 0$, Y is a deterministic function of X.

Maximal information coefficient

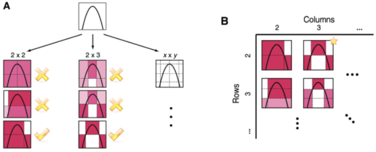
The **maximal information coefficient** (MIC) is defined as :

- $\mathrm{MIC}(X, Y) = \max_G \frac{I((X,Y)|_G)}{\log ||G||}$
- G is the **set of 2d grids**.
- $(X, Y)\ |_G$ represents a **discretization of the variables** onto **grid.**
- $||G||$ is $\min(G_x, G_y)$
  - G_x is the **number of grid cells** in the x direction.
  - G_y is the **number of grid cells** in the y direction.
- $0 \leq MIC(X, Y) \leq 1$

The **intuition** behind:

- If there is a **relationship** between X and Y,
  - Some **discrete gridding of the 2d input space** that **captures** this.
- MIC **searches** over **different grid resolutions** (e.g. 2x2, 2x3, etc)

Maximal information coefficient



Data processing inequality

- **Contex**:
  - Have an unknown variable X.
  - observe a noisy function of it, Y.
  - process Y in some way to create a new variable Z.
- **Theorem 6.3.1**: Suppose $X \to Y \to Z$ forms a Markov chain, so that $X \perp Z \mid Y$. Then $I(X;Y) \geq I(X;Z)$.

Sufficient Statistics

Suppose we have the chain $\theta \to D \to s(D)$.

- $I(\theta; s(D)) \leq I(\theta; D)$
- $s(\mathcal{D})$ is a **sufficient statistic** of the data D for the purposes of inferring θ.
  - $T(X)$ is sufficient for $\theta \quad \text{if} \quad P(X \mid T(X), \theta)$ is independent of θ.
- Since we can **reconstruct the data** from knowing $s(\mathcal{D})$.
  - $\theta \to s(\mathcal{D}) \to \mathcal{D}$

Minimal sufficient statistic for $\mathcal{D}$ is $s(\mathcal{D})$:

- For all sufficient statistics $s'(D)$, there exists some function f such that $s'(D) = f(s(D))$.
- $\theta \to s(D) \to s'(D) \to D$

Fano's inequality

Denote:

- An estimator $\hat{Y} = f(X)$ such that $Y \to X \to \hat{Y}$.
- E be the event $Y \neq \hat{Y}$. $P_e = P(Y \neq \hat{Y})$ be the probability of error.

We have:

- $H(E, Y|\hat{Y}) = H(Y|\hat{Y}) + H(E|Y, \hat{Y}) = H(E|\hat{Y}) + H(Y|E, \hat{Y})$
- so, $H(Y|X) \leq H(Y|\hat{Y}) \leq H(E|\hat{Y}) + H(Y|E, \hat{Y})$
- $H(E|\hat{Y}) \leq H(E)$
- $H(Y \mid E, \hat{Y}) = P(E = 0)H(Y|\hat{Y}, E = 0) + P(E = 1)H(Y|\hat{Y}, E = 1)$

$$H(Y \mid E, \hat{Y}) \leq (1 - P_e) \cdot 0 + P_e \log |Y|$$

- Thus, $H(Y|X) \leq H(E|\hat{Y}) + H(Y|E, \hat{Y}) \leq H(E) + P_e \log |Y|$
- Final, we get : $P_e \geq \frac{H(Y|X) - 1}{\log |Y|}$