情報理論の基礎

村田昇

SGC ライブラリ 37 (20050

2. 情報理論の基礎事項

情報源とはある確率法則に従って文字列を生成するメカニズムであり、数学的にはすべての有限シンボル列に確率が割り当てられた確率空間である: (Σ^*, P)。ここで Σ はシンボルの集合。

記憶のない情報源 (定常) では Σ^* の元は iid 変数列である:

$$P(X^N) = p(X_N) \cdots p(X_1). \tag{1}$$

もちろん Markov 的情報源などが考えられる。

通信路: あるシンボル列を入れたとき別のシンボル列 (シンボルも違っていていい) をある確率で出力するデバイス。$P(Y^N|X^N)$ などを考え得る。無記憶通信路とは一字一字拾い読みして送る通信路である。記憶のある通信路や隠れマルコフ的な通信路もある (ノイズのある状態とない状態がマルコフ的に変わる)。

符号化: 符号語間の間隔をうまくとればノイズに強くなりうる。

Kraft の不等式。

$$\sum 2^{-l_i} \leq 1. \tag{2}$$

これを証明するのは枝の統合と剪定を使えばいい。平均符号長は $\langle i \rangle = \sum p_i l_i$。$l_i = -\log_2 Q_i$ であるとすると

$$\langle l \rangle = \sum p_i \log_2(1/Q_i). \tag{3}$$

Kullback-Leibler 情報量が正であることから

$$\sum p_i \log_2(1/Q_i) - \sum p_i \log_2(1/p_i) \geq 0 \tag{4}$$

したがって、最も能率のよい符号は $Q_i = p_i$、つまり

$$l_i = -\log_2 p_i \tag{5}$$

これを文字 i の情報と呼ぶ。公理的には、その文字の出る確率を p とするとき情報量 $f(p)$ は 1. 非負性: $f(p) \geq 0$、2. 単調減少性: $f(p)$ は p が大きいと小さい、3. 加法性: 独立事象の担う情報量は和になる ($f(pq) = f(p) + f(q)$)。

3. Mode of thinking of information geometry

$\mathcal{S}$ is the space of the totality of probability measures. A parameter family of measures defines a subspace called a *model manifold* $\mathcal{M}$. Our problem is to choose p on $\mathcal{M}$ closest to an empirical measure.

To this end we need a metric. Let

$$D(p, q) = \sum p \log(p/q). \tag{6}$$

This is called the *Kullback-Leibler divergence*. Notice that

$$D(p, q) - [D(p, r) + D(r, q)] = \sum (p_i - r_i)(\log_2 r_i - \log_2 q_i) \tag{7}$$

Therefore, if $\boldsymbol{p}-\boldsymbol{r}$ and $\log \boldsymbol{r}-\log \boldsymbol{q}$ are orthogonal, then we have

$$D(p,q)=D(p,r)+D(r,q). \tag{8}$$

35

m-geodesic (this is an interpolation point of distributions, so no extra normalization is needed)

$$R=\{(1-t)p+tq, t\in[0,1]\}, \tag{9}$$

e-geodesic (this is an interpolation of log of distribution so it is not a distribution. Consequently, we need an extra normalization)

$$R=\{(1-t)\log p+t\log q-\phi, t\in[0,1]\}, \tag{10}$$

where ϕ is a normalization factor. These geodesics can introduce flat surfaces, so we can foliate $\mathcal{S}$. If a model manifold is e-flat, then minimizing $D(p,r)$ wrt r gives the best model; if a model manifold is m-flat, then minimizing $D(r,p)$ wrt r gives the best model. Here, 'best' means the 'orthogonal' projection

Remark Minimizing $D(p,r)$ wrt r to choose the optimal model is natural from the large-deviation point of view: Suppose r is the true distribution. Then, p maximizing D is the most probably observable distribution empirically. Therefore, if p is actually observed to choose r is rational. This is the max likelihood estimate. If the model manifold is e-flat, ML estimate is unique. Otherwise, there is no guarantee of uniqueness.

4. Coding and various information quantity

42 Since KL information is not symmetric, we can consider two projections.

$$B_\epsilon^m(p) = \{q|D(p,q)\le\epsilon\}, \tag{11}$$
$$B_\epsilon^e(p) = \{q|D(q,p)\le\epsilon\}. \tag{12}$$

When we say 'coding,' we are creating an information source (the 'true distribution'). Therefore, $B_\epsilon^m(p)$ is to create an optimum coding scheme q, On the other hand, $B_\epsilon^e(p)$ is the set of sources that is approximated by the coding scheme p. Therefore, if the best coding is needed for the source p, we find q that minimize $D(p,q)$. If the best source q is needed for the coding scheme p, we minimize $D(q,p)$.

43 Shannon's first theorem: The optimal average code length L_m satisfies

$$H(X)\le L_m\le H(X)+1, \tag{13}$$

Shannon-Fano coding is explained.

Asymptotic equipartition property: This is the SMB theorem.

51 Mutual information $I(X,Y)$: intuitively, this is the information about X that Y has (and vice versa). $H(Y|x)$ measures how x is blurred when it is sent. This implies that on the average $2^{nH(Y|X)}$ sequences are obtained by sending a single sequence. Therefore, the totality of n Y symbol sequences $2^{nH(Y)}$ can carry

$$2^{nH(Y)-nH(Y|X)}=2^{nI(Y;X)} \tag{14}$$

distinguishable sequences. That is, without being corrupted, we can send this much of n symbol sequences.

$$C = \max_{p(X)} I(Y;X) \tag{15}$$

is called the *channel capacity*. Here $p(X)$ is determined by coding. Shannon's second theorem guarantees the existence of p that realizes C. The rate R is defined by

$$R = \frac{\#\text{ of actually transmitted letters}}{\#\text{ of letters put into the channel}} \tag{16}$$

If $I = 0$, X and Y are independent.

If I is large, then from Y we can guess X. To this end, Blahut-Arimoto algorithm to determine p that realizes C.

5. Selecting models

To choose the best parameteric model is the model selection problem. If the true model with parameter θ exists in the search set, then the accuracy of the estimated parameter is bounded by the Cramér-Rao inequality. Let $\hat{\theta}$ be the unbiased estimator of θ. Then,

$$V(\hat{\theta}) \geq 1/nJ, \tag{17}$$

where n is the number of data, and J is the Fisher information

$$J = E_\theta \left[\left(\frac{\partial}{\partial \theta} \log p(X;\theta) \right)^2 \right] = -E_\theta \left(\frac{\partial^2}{\partial \theta^2} \log p \right) \tag{18}$$

Here, p is the model distribution with parameter θ.
[Demo] First let us check the equality (18). Since $\sum p_i(\theta) = 1$,

$$E_\theta \left(\frac{\partial}{\partial \theta} p \right) = 0. \tag{19}$$

72 Here E_θ is the expectation wrt $p_i(\theta)$. Differentiating this again

$$E_\theta \left(\frac{\partial^2}{\partial \theta^2} p \right) + E_\theta \left(\left[\frac{\partial}{\partial \theta} p \right]^2 \right) = 0. \tag{20}$$

The estimation of the parameter is given by ($\hat{\theta}$ is unbiased)

$$\sum \hat{\theta}(X_1, X_2, \cdots X_n) p(X_1, \theta) p(X_2, \theta) \cdots p(X_n, \theta) = \theta. \tag{21}$$

Therefore,

$$\sum \hat{\theta} \frac{\partial}{\partial \theta} \prod p_k(\theta) = 1. \tag{22}$$

On the other hand $\sum \prod_k p_k = 1$ implies

$$\sum \frac{\partial}{\partial \theta} \prod p_k(\theta) = 0. \tag{23}$$

Therefore,

$$\sum(\hat{\theta}-\theta)\frac{\partial}{\partial\theta}\prod_k p_k(\theta)=1. \tag{24}$$

We know

$$\frac{\partial}{\partial\theta}\prod_k p_k(\theta)=\left(\sum\frac{\partial}{\partial\theta}\log p_k\right)\prod p_k(\theta). \tag{25}$$

Combining the above two, we get

$$\sum(\hat{\theta}-\theta)\left(\sum\frac{\partial}{\partial\theta}\log p_k\right)\prod p_k(\theta)=1. \tag{26}$$

Using Cauchy-Schwarz, we get

$$\left\{\sum(\hat{\theta}-\theta)^2\prod p_k(\theta)\right\}\left\{\sum\left(\sum\frac{\partial}{\partial\theta}\log p_k\right)^2\prod p_k(\theta)\right\}\geq 1. \tag{27}$$

Since the average of the log likelihood

$$\sum\left(\sum\frac{\partial}{\partial\theta}\log p_k\right)\prod p_k(\theta)=0, \tag{28}$$

the crossterms in (27) vanish, so

$$E_\theta\left[\left(\sum\frac{\partial}{\partial\theta}\log p_k\right)^2\right]=E_\theta\left[\sum\left(\frac{\partial}{\partial\theta}\log p_k\right)^2\right]=nE_\theta\left[\left(\frac{\partial}{\partial\theta}\log p(\theta)\right)^2\right] \tag{29}$$

That is,

$$V(\hat{\theta})nJ\geq 1. \tag{30}$$

The inequality implies that the best estimator is the one that attains the minimum $1/nJ$. It can be shown as follows that if a model is smooth in θ, the model with KL information minimum is in this sense the best.

C Large deviation or using Sanov's theorem, minimizing the KL

If $\hat{\theta}$ is estimated by the log-likelihood maximization, then it obeys $N(\theta, 1/NJ)$.

$$\sum\frac{\partial}{\partial\theta}\log p(X_i,\theta)=0 \tag{31}$$

Taylor-expanding this around the true value we obtain

$$\sum\frac{\partial}{\partial\theta}\log p(X_i;\theta)+(\hat{\theta}-\theta)\sum\frac{\partial^2}{\partial\theta^2}\log p(X_i;\theta)=0. \tag{32}$$

The central limit theorem tell us that

$$\frac{1}{\sqrt{N}}\sum\frac{\partial}{\partial\theta}\log p(X_i,\theta) \tag{33}$$

obeys asymptotically the normal distribution with average zero and the variation J according to (18). Combining this and (32), the desired result is obtained.

86 Resampling (bootsrap) method

If we know the true distribution P, then the best model should be the one minimizing $D(P, P(\hat{\theta}))$. If we can have many empirical distributions, then the average of $D(P_i, P(\hat{\theta}))$
75 should be minimized.

AIC

The best model is the one that minimizes $D(P, P(\hat{\theta}))$ for the true P. However, we do not know P; we only no the empirical result $\hat{P}$, so the distribution estimated by the data is the one that minimizes $D(\hat{P}, P(\hat{\theta}))$. Akaike's idea is to estimate the average of $D(P, P(\hat{\theta}))$ with the aid of the asymptotic normality of $\hat{\theta}$.

Remark Suppose an empirical distribution $\hat{P}$ is given. Then, $\hat{\theta}$ minimizing $D(\hat{P}, P(\hat{\theta}))$ should be the best empirical result. However, if we assume that the true distribution P is known, then $\hat{\theta}$ minimizing $D(P, P(\hat{\theta}))$ should be the best model. If we replace this P with the best estimate from $\hat{P}$, this should be the estimate we should use to optimize the parameters.

A bootstrap version is to minimize the bootstrap average of $D(P', P(\theta))$, where the primes denote bootstrap results.

Multidimensional case:

$$J_{ij} = E^X E_\theta \left(\frac{\partial}{\partial \theta_i} \log p(Y|X;\theta \frac{\partial}{\partial \theta_j} \log p(Y|X;\theta) \right), \tag{34}$$

where E^X is the average over X and E_θ is the average over $p(y|x,\theta)$. Let P be the true distribution, $P(\theta)$ be a model, and θ^* is the closest model to the truth:

$$\theta^* = \mathrm{argmin} D(P, P(\theta)). \tag{35}$$

$$G_{ij} = E_P \left(\frac{\partial}{\partial \theta_i} \log p(X;\theta^*) \frac{\partial}{\partial \theta_j} \log p(X;\theta^*) \right), \tag{36}$$

$$Q_{ij} = -E_P \left(\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log p(X;\theta^*) \right). \tag{37}$$

The asymptotic normality of the estimate reads

$$\hat{\theta} \in N(\theta, (1/N) Q^{-1} G Q^{-1}). \tag{38}$$

We know

$$\frac{\partial}{\partial \theta_i} D(P, P(\theta^*)) = 0. \tag{39}$$

Noting that

$$\frac{\partial^2}{\partial \theta_i \partial \theta_j} D(P, P(\theta^*)) = -E_P \left(\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log p(X;\theta^*) \right), \tag{40}$$

we have

$$D(P, P(\hat{\theta}) = D(P, P(\theta^*)) + \frac{1}{2} \sum Q_{ij} (\hat{\theta} - \theta^*)_i (\hat{\theta} - \theta^*)_j. \tag{41}$$

Therfore, using

$$E[(\hat{\theta} - \theta^*)_i (\hat{\theta} - \theta^*)_j] = \frac{1}{N} (Q^{-1} G Q^{-1})_{ij}, \tag{42}$$

we have

$$E[D(P, P(\hat{\theta})] = D(P, P(\theta^*)) + \frac{1}{2N} Tr(GQ^{-1}). \tag{43}$$

The first term cannot be computed, so $E(D(\hat{P}, P(\hat{\theta}))$ is estimated. $\hat{P}$ is closest to $\hat{\theta}$, and P to θ^*, so we can expand in two ways as

$$\begin{aligned} D(\hat{P}, P(\theta^*)) &= D(P, P(\theta^*)) + \text{scatter around true } P, & (44) \\ &= D(\hat{P}, P(\hat{\theta}) + \frac{1}{2} \sum \hat{Q}_{ij} (\hat{\theta} - \theta^*)_i (\hat{\theta} - \theta^*)_j & (45) \end{aligned}$$

The average of the first line should be

$$E(D(\hat{P}, P(\theta^*))] = D(P, P(\theta^*)). \tag{46}$$

The second line

$$D(\hat{P}, P(\theta^*)) = D(\hat{P}, P(\hat{\theta})) + \frac{1}{2} \sum \hat{Q}_{ij} (\hat{\theta} - \theta^*)_i (\hat{\theta} - \theta^*)_j + XXX \tag{47}$$

where XXX is the scattering perpendicular to the model manifold. Therefore,

$$E(D(\hat{P}, P(\theta^*)) = E(D(\hat{P}, P(\hat{\theta})) + \frac{1}{2N} Tr(GQ^{-1}). \tag{48}$$

Therefore,

$$E(D(\hat{P}, P(\hat{\theta})) = E(D(\hat{P}, P(\theta^*)) - \frac{1}{2N} Tr(GQ^{-1}). \tag{49}$$

Combining all

$$E(D(\hat{P}, P(\hat{\theta})) = E(D(P, P(\theta^*)) - \frac{1}{2N} Tr(GQ^{-1}). \tag{50}$$

That is

$$E(D(P, P(\hat{\theta})) = E(D(\hat{P}, P(\hat{\theta})) + \frac{1}{N} Tr(GQ^{-1}). \tag{51}$$

Now, G and Q are estimated with the aid of the emprical distribution. Usually $G = Q = J$, so the last term becomes m/N, where m is the number of parameters.

(51) reads

$$E(D(P, P(\hat{\theta})) = E(D(\hat{P}, P(\hat{\theta})) + \frac{m}{N}, \tag{52}$$

where $E(D(\hat{P}, P(\hat{\theta})) = \frac{1}{N} \sum_i \log[1/Np(X_i, \hat{\theta})]$, where N is the number of data, so ignoring the constant term the standard AIC

$$AIC = -2 \sum \log p(X_i; \hat{\theta}) + 2m. \tag{53}$$

becomes a measure of goodness.