Information Theory and Statistics, Part I

Information Theory 2013
Lecture 6

George Mathai

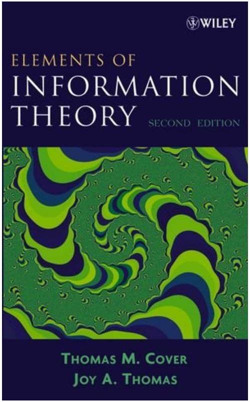
May 16, 2013

Outline

This lecture will cover

- Method of Types.
- Law of Large Numbers.
- Universal Source Coding.
- Large Deviation Theory.
- Examples of Sanov's Theorem.

All illustrations are borrowed from the book.



Method of Types

Definition : The type P_x of a sequence $x_1, x_2 \ldots x_n$ is the relative proportion of the occurrences of each symbol of $\mathcal{X}$

$$P_x^n(a) = \frac{N(a|x^n)}{n}$$

Ex: $\mathcal{X}$ ={1,2,3}. Let x= 11321.
then $P_x(1) = 3/5$, $P_x(2) = 1/5$, $P_x(3) = 1/5$.

Hence $P_x = \{3/5, 1/5, 1/5\}$

Let $\mathcal{P}_n$ denotes the set of types with denominator n .

Ex: $\mathcal{P}_5$={(0/5, 0/5, 5/5), (0/5, 1/5, 4/5)..(0/5, 5/5, 0/5)..(5/5, 0/5, 0/5)}

If $P \in \mathcal{P}_n$, then the set of sequences of length n and type P is called the type class of P, denoted by $T(P)$

$$T(P) = \{x \in \mathcal{X}^n : P_x = P\}$$

Ex: $T(P_x) = \{11123, 11132, 11213, \ldots 32111\}$

Method of Types

Theorem:

$$|\mathcal{P}_n| \leq (n+1)^{|\mathcal{X}|}$$

This bound is very high. One can achieve much better than this

Method of Types

Theorem: If $X_1, X_2, \ldots X_n$ are drawn IID according to $Q(x)$, the probability of x depends only on its type and is given by

$$Q^n(x^n) = 2^{-n(H(P_x)+D(P_x||Q))}$$

Proof:

$$\begin{aligned}
Q^n(x^n) &= \prod_{i=1}^{n} Q(x_i) \\
&= \prod_{a\in\mathcal{X}} Q(a)^{N(a|x^n)} \\
&= \prod_{a\in\mathcal{X}} Q(a)^{nP_{x^n}(a)} \\
&= \prod_{a\in\mathcal{X}} 2^{nP_{x^n}(a)\log Q(a)} \\
&= 2^{-n(H(P_x)+D(P_x||Q))}
\end{aligned}$$

Corollary : If x^n is in the type class of Q, then

$$Q^n(x) = 2^{-nH(Q)}$$

Method of Types

Theorem: Size of a type class $T(P)$

For any type $P \in |\mathcal{P}|$

$$\frac{1}{(n+1)^{|\mathcal{X}|}} 2^{nH(P)} \leq |T(P)| \leq 2^{nH(P)}$$

where the exact value of $|T(P)|$ is given by

$$|T(P)| = \binom{n}{nP(a1), nP(a2), \dots nP(a_{|\mathcal{X}|})}$$

So we look for the bounds. Upper Bound

$$\begin{aligned} 1 &\geq P^n(T(P)) \\ &= \sum_{x^n \in T(P)} P^n(x^n) \\ &= \sum_{x^n \in T(P)} 2^{-nH(P)} \qquad\qquad = |T(P)| 2^{-nH(P)} \end{aligned}$$

Method of Types

Theorem:(Probability of type class) for any $P \in \mathcal{P}_n$ and any distribution Q , the probability of the type class $T(P)$ under Q^n is $2^{-nD(P||Q)}$ to first order in the exponent. More precisely,

$$\frac{1}{(n+1)^{|\mathcal{X}|}} 2^{-nD(P||Q)} \leq Q^n(T(P)) \leq 2^{-nD(P||Q)}$$

Proof:

$$\begin{aligned} Q^n(T(P)) &= \sum_{x^n \in T(P)} Q^n(x^n) \\ &= \sum_{x^n \in T(P)} 2^{-n(H(P_x)+D(P_x||Q))} \\ &= |T(P)| 2^{-n(H(P_x)+D(P_x||Q))} \end{aligned}$$

Using the bounds we found on $T(P)$ we get

$$\frac{1}{(n+1)^{|\mathcal{X}|}} 2^{-nD(P||Q)} \leq Q^n(T(P)) \leq 2^{-nD(P||Q)}$$

Law of Large Numbers

Given $\epsilon > 0$ we can define a typical set T_Q^ϵ of sequences for the distribution Q^n as

$$T_Q^\epsilon = \{x^n : D(P_{x^n}||Q) \leq \epsilon\}$$

Then the probability that x^n is not typical is

$$\begin{aligned}
1 - Q^n(T_Q^\epsilon) &= \sum_{P:D(P||Q)>\epsilon} Q^n(T(P)) \\
&\leq \sum_{P:D(P||Q)>\epsilon} 2^{-nD(P||Q)} \\
&\leq \sum_{P:D(P||Q)>\epsilon} 2^{-n\epsilon} \\
&\leq (n+1)^{|\mathcal{X}|} 2^{-n\epsilon} \\
&= 2^{-n(\epsilon - |\mathcal{X}| \frac{log(n+1)}{n})}
\end{aligned}$$

As $n \to \infty$, then $1 - Q^n(T_Q^\epsilon) \to 0 \implies Pr(T_Q^\epsilon) \to 1$

Law of Large Numbers

Theorem: Let$X_1, X_2 \dots X_n$ be $i.i.d. \sim P(x)$. Then

$$Pr\{D(P_{x^n}||P) > \epsilon\} \leq 2^{-n(\epsilon - |\mathcal{X}|\frac{log(n+1)}{n})}$$

$and\ consequently, D(P_{x^n}||P) \to 0 with\ probability\ 1$

Law of Large Numbers

Definition: Strongly typical set $A_\epsilon^*(n)$

$$A_\epsilon^{*(n)} = \left\{ x^n \in \mathcal{X}^n : \begin{array}{l} |\frac{1}{n}N(a|x) - P(a)| \leq \frac{\epsilon}{|\mathcal{X}|} if P(a) \geq 0 \\ N(a|x) = 0 \end{array} \right\}$$

Typical sets consists of sequences whose types does not differ from the true probabilities by more than $\frac{\epsilon}{|\mathcal{X}|}$

Universal Source Coding

Definition: A fixed - rate block code of rate R for source $X_1, X_2 \dots X_n$ which has an unknown distribution Qconsists of two mapping: the encoder,

$$f_n : \mathcal{X}^n \to \{1, 2, \dots 2^{nR}\}$$

and the decoder

$$\phi_n : \{1, 2, \dots 2^{nR}\} \to \mathcal{X}^n$$

where R is called the *rate* of the code. Probability of error for the code wrt distribution

$$P_e^{(n)} = Q^n(X^n : \phi_n(f_n(X^n)) \neq X^n)$$

Universal Source Coding

Definition: A rate R block code for a source will be called *universal* if the functions f_n and ϕ_n don't depend on the distribution Q and if $P_e^{(n)} \to 0$ as $n \to \infty$ if $R > H(Q)$

Theorem: There exists a sequence of $(2^{nR}, n)$ universal source codes such that $P_e^{(n)} \to 0$ for every source Q such that $H(Q) < R$

Universal Source Coding

Proof: Let

$$R_n = R - |\mathcal{X}| \frac{log(n+1)}{n}$$

Consider the sequence

$$A = \{x^n \in \mathcal{X}^n : H(P_x) \leq R_n\}$$

Then

$$\begin{aligned}
|A| &= \sum_{P \in \mathcal{P}_n : H(P) \leq R_n} |T(P)| \\
&\leq \sum_{P \in \mathcal{P}_n : H(P) \leq R_n} 2^{nH(P)} \\
&\leq \sum_{P \in \mathcal{P}_n : H(P) \leq R_n} 2^{nR_n} \\
&\leq (n+1)^{|\mathcal{X}|} 2^{nR_n} \qquad = 2^{n(R_n + |\mathcal{X}| \frac{log(n+1)}{n})} = 2^{nR}
\end{aligned}$$

Universal Source Coding

Probability of decoding error $P_e^{(n)}$can be found by

$$\begin{aligned}
P_e^{(n)} &= 1 - Q^n(A) \\
&= \sum_{P \in \mathcal{P}_n : H(P) \leq R_n} Q^n(T(P)) \\
&\leq (n+1)^{|\mathcal{X}|} \max_{P : H(P > R_n)} Q^n(T(P)) \\
&\leq (n+1)^{|\mathcal{X}|} 2^{-n \min_{P : H(P > R_n)} D(P||Q)}
\end{aligned}$$

As $n \to \infty$, $P_e^{(n)} \to 0$

Large Deviation Theory

Theory of large deviations concerns the asymptotic behaviour of remote tails of sequences of probability distributions.
Let E be a subset of the set of probability mass functions.

$$Q^n(E) = Q^n(E \cap \mathcal{P}_n) = \sum_{x^n : P_{x^n} \in E \cap \mathcal{P}_n} Q^n(x^n)$$

$Q^n(E) \to 1$ If E contains relative entropy neighbourhood of Q
$Q^n(E) \to 0$ other wise

Large Deviation Theory

Sanov's Theorem: Let $X_1, X_2 \dots X_n$ be $i.i.d. \sim Q(x)$. Let $E \subseteq \mathcal{P}$ be a set of probability distributions. Then

$$Q^n(E) = Q^n(E \cap \mathcal{P}_n) \leq (n+1)^{|\mathcal{X}|} 2^{-nD(P^*||Q)}$$

where

$$P^* = \arg\min_{P \in E} D(P||Q)$$

is the distribution E that is closest to Q in relative Entropy. If in addition, the set E is the closure of its interior, then

$$\frac{1}{n} \log Q^n(E) \to -D(P^*||Q)$$

Note: E is not in the typical set

Large Deviation Theory

Proof:

$$
\begin{aligned}
Q^n(E) &= \sum_{P \in E \cap \mathcal{P}_n} Q^n(T(P)) \\
&\leq \sum_{P \in E \cap \mathcal{P}_n} 2^{-nD(P||Q)} \\
&\leq \sum_{P \in E \cap \mathcal{P}_n} \max_{P \in E \cap \mathcal{P}_n} 2^{-nD(P||Q)} \\
&= \sum_{P \in E \cap \mathcal{P}_n} 2^{-min_{P \in E \cap \mathcal{P}_n} nD(P||Q)} \\
&\leq \sum_{P \in E \cap \mathcal{P}_n} 2^{-min_{P \in E} nD(P||Q)} \\
&= \sum_{P \in E \cap \mathcal{P}_n} 2^{-nD(P^*||Q)} \\
&\leq (n+1)^{\mathcal{X}} 2^{-nD(P^*||Q)}
\end{aligned}
$$

Examples of Sanov's Theorem

Task:

$$Pr\{\frac{1}{n}\sum_{i=1}^{n} g_j(X_i) \geq \alpha_j, j = 1, 2, \ldots k\}$$

Set E is defined as

$$E = \{P : \sum_{a} P(a) g_j(a) \geq \alpha_j, j = 1, 2, \ldots, k\}$$

To find te closest distribution in $E$to Q. We minimize the $D(P||Q)$ subject tot he above constraint. The resulting functional

$$J(P) = \sum_{x} P(x) \log \frac{P(x)}{Q(x)} + \sum_{i} \lambda_i \sum_{x} P(x) g_i(x) + \nu \sum_{x} P(x)$$

Differentiating and we get

$$P^*(x) = \frac{Q(x) e^{\sum_i \lambda_i g_i(x)}}{\sum_{a \in \mathcal{X}} Q(a) e^{\sum_i \lambda_i g_i(x)}}$$