黒木玄 **Gen Kuroki**
@genkuroki

mathtod.online/@ta_to_co/21612…

多項分布の漸近挙動でKL情報量を出す話は、多項分布を決める確率分布 q_i とその確率分布でシミュレートしたい先の確率分布 p_i を比較する話になっています。

確率分布 q_i でサンプルを生成したときに、そのサンプルは多項分布に従います。

確率分布 q_i が生成したサンプルから計算された経験分布がほぼ p_i になる確率の対数の $-1/n$ 倍がほぼKL情報量になるというのがSanovの定理。

確率分布 q_i がそれでシミュレートしたい先の確率分布 p_i にどれだけ近いかを、確率分布 q_i によるサンプル生成で確率分布 p_i が偶然再現される確率の高さで測るのは１つの自然な考え方。

そういう考え方をすればSanovの定理よりKL情報量を考えなければいけなくなります。

2017年06月03日 09:14 · Web · 0 · 3 · Webで開く

黒木玄 **Gen Kuroki** @genkuroki Saturday at 9:44am
大数の法則によって、確率分布 q_i でサイズ n のサンプルを生成すると、そのサンプル内での分布は $n \to \infty$ で q_i に近付きます。

そして、サンプル内での分布(経験分布)が q_i からある一定以上離れる確率は指数函数的に減少する。その指数函数的減少の部分を具体的に評価しているのが、Sanovの定理です。

すなわち、指数函数的減少の速さがKL情報量で測られる。KL情報量が小さいと指数函数的減少の速さは小さくなります。その速さで分布の違いを評価することは一つの自然な考え方であり、数学的性質もよいので、KL情報量が多用されています。

黒木玄 **Gen Kuroki** @genkuroki Saturday at 1:26pm
Sanovの定理のより正確な形は

$$\begin{aligned} &P(P_n \in A) \\ &= \exp\left[-n \inf_{p \in A} D(p\|q) + o(n)\right] \end{aligned}$$

の形になります(A は確率分布の「良い」集合、P_n は確率分布 q が生成するサイズ n のサンプルにおける経験分布)。

大雑把に上の式は「A に制限した条件付き確率分布を考えるとき、n が大きなときに効いてくるのはKL情報量 $D(p\|q)$ が最小になる分布 p だけであること」を意味しています。

そして上の式は「漸近挙動を得るための最も基本的なテクニックであるLaplaceの方法」の弱いバージョンが成立しているとも読めます。この点については後回しにして、上の式の具体例を示します。

これは二項分布でのSanovの定理の話で、
mathtod.online/@genkuroki/2796...
mathtod.online/@genkuroki/2797...
の再掲になります。

続く

黒木玄 **Gen Kuroki** @genkuroki Saturday at 1:33pm
続き

以下の計算はSanovの定理の応用です。当然そうなるだろうなという結果がSanovの定理から導かれるという話です。

問題：$0 < q < 1$ であるとし、$0 \leqq a < b \leqq q$であると仮定する。次の極限を求めよ：

$$\lim_{n\to\infty} \frac{-1}{n} \log \sum_{a \leqq k/n \leqq b} \binom{n}{k} q^k (1-q)^{n-k}.$$

二項分布における経験分布 k/n が a 以上 b 以下になる確率の対数の $-1/n$ 倍の極限を求めよという問題です。

解答に続く

黒木玄 **Gen Kuroki** @genkuroki Saturday at 1:41pm
続き

当たりが確率 q で出るルーレットを n 回まわしたとき、当たりが出た回数 k の割合 k/n は、大数の法則によって n を大きくするとき q に近付きます。

$0 \leqq a < b \leqq q$ と仮定しているので、当たりの出た割合 k/n が a 以上 b 以下になる場合に制限すると、n を大きくしたとき、当たりの出た割合は q に最も近い b の近くに貼り付くようになるはずです。私はこれを「大数の法則による力」と呼んでいます。

単なる確率現象に過ぎないのに、大数の法則によって、条件付き確率分布において、まるで力が働いて見えるようなことが起こる。

だから、件の問題における極限は対数の中の和を $k/n \approx b$ に制限して計算すれば良さそうである。

Sanovの定理は実際にそうなることを主張しています。

続く

黒木玄 **Gen Kuroki** @genkuroki Saturday at 1:52pm
続き

だから次の問題を解けばよい。

問題：整数 k_n が $k_n/n \approx b$ を満たすという仮定のもとで次の極限を求めよ：

$$\lim_{n\to\infty} \frac{-1}{n}\log\binom{n}{k_n} q^{k_n}(1-q)^{n-k_n}.$$

これの計算はほぼ大学入試問題レベルなので略します。大学入試的方法ではなく、スターリングの公式をがちゃっと代入した方が簡単です。結果は

$$b\log\frac{b}{q} + (1-b)\log\frac{1-b}{1-q}$$

になります。これはKullback-Leibler情報量

$$p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}$$

の条件 $p \in [a,b]$ の下での最小値になっています。($0 \leqq a < b \leqq q$ と仮定してあった。p が q に近いほどKL情報量は小さくなる。)

黒木玄 **Gen Kuroki** @genkuroki Saturday at 2:05pm
mathtod.online/@h_okumura/2162…

Kullback-Leibler情報量でできることは二つの確率分布(もしくは乱数)の比較。

だから、比較したい二つの確率分布(もしくは乱数)が何であるかをきちんと述べることが大事。

現実には損失函数を明確にしてどちらの予報の損失が小さいかを問題にするべきかも。

KL情報量は二種類の乱数の発生のさせ方 A, B があるとき、「方法 A による乱数発生で方法 B による乱数発生をどれだけの精度でシミュレートできているか」を表わす指標が欲しいときに役に立ちます。それ以上のものでもそれ以下のものでもない。

だから、そのシミュレーションで損じる便益や損失はKL情報量では一切考慮されない。

現実の問題に立ち向かうときには、便益は何でどれくらいか、損失は何でどれくらいか、をきちんと扱う必要が出て来ます。

話題を大きく脱線させてしまった。

黒木玄 **Gen Kuroki** @genkuroki Saturday at 2:13pm
wolframalpha.com/input/?i=plot…
添付画像は上のリンク先より。

$q = 3/5$ の場合の

$$p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}$$

のグラフです。$p = q = 3/5$ で最小になります。

一般のKL情報量も次元が上がるだけで同じような形をしています。

mathtod.online/media/rYG2H10xP…

黒木玄 **Gen Kuroki** @genkuroki Saturday at 2:48pm

mathtod.online/@genkuroki/1882…
の一番下で大学入試問題を引用して、それと類似のKullback-Leibler情報量を理解するための大学入試問題レベルの問題を出しています。Sanovの定理入門にはその問題を解いておくといいと思う。解答は引用した大学入試問題の解答例と同様なので、リンクをたどってそれを読んで真似をすればよい。

あり @ta_to_co Saturday at 11:03am

@genkuroki

だからp_i側が経験分布を呼んでるんですね．相当すっきりしました．

私はまだ確率論的にものをみることになれていないようです．証明を追うと係数にしか見えていませんでした．

ありがとうございます．

mathtod.online powered by Mastodon