黒木玄 **Gen Kuroki** @genkuroki on Apr 27

これもツイッターの方に書いたはずのことですが、統計関係の数学的教養として

(1)大数の法則、(2)中心極限定理

の二つはすでに定番になっていますが、さらに

(3)Kullback-Leibler情報量(相対エントロピーの-1倍)に関する大偏差原理(Sanovの定理)

の類も定番になるべきだと思っています。KL情報量は統計的に得られたモデルの予測精度を測るための最もよく使われている指標になっているのですが、Sanovの定理を知らないとどうしてKL情報量でモデルの予測精度を測れるかがよくわからなくなってしまうと思います。

平均符号長の下限(KL情報量になる)からの乖離とモデルの予測誤差の類似で理解した気分になることもできるとは思いますが、Sanovの定理を使った理解と比較すると類似に頼っており、理解の仕方としては少し弱くなると思います。(しかし、どうしてKL情報量を「情報量」と呼ぶことが自然であるかについては、平均符号長の下限であるを知れば納得し易い。まあ、実際にはKL情報量のことをKLダイバージェンスと呼ぶことが多いのですが。)

黒木玄 **Gen Kuroki** @genkuroki on Apr 27

KL情報量に関するSanovの定理を現実に定番の教養だとみなせるためには「適切に数学的な厳密性を犠牲にすれば内容的に易しいこと」が必要です。理解がひどく困難な数学的結果を定番の教養にするのは現実的に無理。細かいことを言わなければSanovの定理(の易しいバージョン)は本質的に多項分布にスターリングの公式を適用するだけで簡単に出て来ます。

詳しくは
math.tohoku.ac.jp/~kuroki/LaTe…
を見て下さい。

実際の計算のときには、多項分布をポアソン分布の直積の条件付き確率分布とみなして、まずポアソン分布にスターリングの近似公式を適用した式を書いてみると色々納得し易くなります。計算の見通りも易しくなるし、ピアソンのカイ二乗統計量との関係も見易くなります。

そういう方針の計算が上のPDFの1.9節にあります。

黒木玄 **Gen Kuroki** @genkuroki on Apr 27

確率の総和は常に1になることを了解しておけば、確率が何に比例するかのみを指定すれば確率分布が決まります。

非負の整数に値を持つ確率変数 K_i がポアソン分布に従うことは

$$P(K_i = k_i) \propto \frac{\lambda_i^{k_i}}{k_i!}$$

と書けます。ここで $P(\ )$ は確率を表わし、λ_i は正の定数です。

これにスターリングの公式 $k_i! \sim \sqrt{2\pi k_i}\exp(k_i \log k_i - k_i)$ を適用すると、k_i が大きなとき漸近的に

$$P(K_i = k_i) \propto \frac{\exp(-k_i \log(k_i/\lambda) + k_i)}{\sqrt{2\pi k_i}}$$

右辺の指数函数の中にすでにKL情報量 $D(q||p) = \sum_i q_i \log(q_i/p_i)$ っぽい式が見えています。ポアソン分布の段階でここまで見える。

以下 K_i 達は独立と仮定します。続く

黒木玄 **Gen Kuroki** @genkuroki on Apr 27

続き

$\sum_{i=1}^r k_i = n$ と仮定し、比例定数は n には依存してよいということにすると

$$P(K_1 = k_1, \dots, K_r = k_r) \propto \frac{\exp(-\sum_{i=1}^r k_i \log(k_i/\lambda_i))}{\prod_{i=1}^r \sqrt{2\pi k_i}}.$$

さらに $p_i = \lambda_i/(\sum_{i=1}^r \lambda_i)$, $q_i = k_i/n$ とおくと、

$$P(K_i = k_i, \dots, K_r = k_r) \propto \frac{\exp(-n\sum_{i=1}^r q_i \log(q_i/p_i))}{\prod_{i=1}^r \sqrt{2\pi q_i}}.$$

右辺の指数函数の中身にある

$$D(q||p) = \sum_{i=1}^r q_i \log \frac{q_i}{p_i}$$

の部分がKL情報量です。このKL情報量は確率分布 p_i による確率分布 q_i の予測誤差を意味しています。

黒木玄 **Gen Kuroki** @genkuroki on Apr 27

続き

KL情報量 $D(q||p)$ は

$$P(K_i = k_i, \dots, K_r = k_r) \propto \frac{\exp(-nD(q||p))}{\prod_{i=1}^r \sqrt{2\pi q_i}}.$$

の形で登場します。$q_i = k_i/n$ は経験分布で q_i はパラメーターで与えられたモデルの確率

分布とみなせます。

この式は大雑把にモデルの確率分布 p_i のもとでの n 回の独立試行で経験分布 q_i が偶然得られる確率の対数がほぼ $-nD(q||p)$ になることを意味しています。

$D(q||p)$ が大きければ大きいほど確率の対数は急激に小さくなります。

KL情報量 $D(q||p)$ はモデル p で確率分布 q が偶然生じない程度の大きさを表わしています。

Sanovの定理の内容は大雑把にこういうものだと思って構いません。

多項分布にスターリングの近似公式をつっこむだけの話でしかありません。

黒木玄 **Gen Kuroki** @genkuroki on Apr 27

大雑把に言えば、確率分布 p のもとでの独立試行で経験分布 q が偶然生じる確率の対数の -1 倍(の $1/n$ 倍) がKullback-Leibler情報量 $D(q||p)$ です。 -1 倍しない単なる確率の対数の方は相対エントロピーと呼ばれています。

(確率の対数)=(相対エントロピー)=－(KL情報量)

「相対」や「KL」を外したバージョンは

(場合の数の対数)=(エントロピー)=－(情報量).

黒木玄 **Gen Kuroki** @genkuroki on Apr 27

「中心極限定理」よりも「Kullback-Leibler情報量に関するSanovの定理」の方が易しい話のように感じられるのは私だけ？

黒木玄 **Gen Kuroki** @genkuroki on Apr 28

連続な確率分布 $p(x)\,dx$ と $q(x)\,dx$ について、モデル p で真の分布 q を予測するときの「誤差」を測るKullback-Leibler情報量は

$$D(q||p) = \int q(x)\log\frac{q(x)}{p(x)}\,dx.$$

黒木玄 **Gen Kuroki** @genkuroki on Apr 28

mathtod.online/@genkuroki/2104…
の訂正

KL情報量になるのは「平均符号長の下限」ではなく、「平均符号長の下限からの乖離」の方です。括弧を挿入する場所を間違った。

確率分布 q に従うソースの平均符号長の下限は q の情報量

$$H(q) = -\sum_i q_i \log q_i.$$

確率分布 q に従うソースの確率モデル p での平均符号長は

$$G(q||p) = -\sum_i q_i \log p_i.$$

それらの差がKL情報量

$$\begin{aligned} D(q||p) &= G(q||p) - H(q) \\ &= \sum_i q_i \log(q_i/p_i). \end{aligned}$$

黒木玄 **Gen Kuroki**
@genkuroki

Kullback-Leibler divergenceを確率分布の違いの尺度として用いることの説明で、「平均符号長の下限からの乖離」の話をしてしまうパターンがよくあるのですが、率直に言ってあんまりよろしくないと思う。Sanovの定理の解説をするべきだと思う。

確率分布 q と異なる確率分布 p による q のシミュレーションでは試行回数 n が増えると大数の法則によって急速にボロが出ます。KLダイバージェンスはボロが出る速さを表わしています。

Sanovの定理は以上のように解釈できる結果なので、モデル p が真の分布 q をどれだけの精度で予測できるかの指標として、KLダイバージェンスがよく使われているわけです。

片方がもう一方をシミュレートする話なので当然 p と q について非対称になります。

大数の法則、中心極限定理の次に学ぶべき事柄として、Sanovの定理やCramerの定理のようなi.i.d.における大偏差原理が標準的な教養になるとよいと思います。

2017年05月13日 17:04 · Web · 2 · 5 · Webで開く

黒木玄 **Gen Kuroki** @genkuroki on May 13
数式入れないと500文字は相当な分量になる！

8zu @8zu@social.targaryen.house on May 14
@genkuroki ありがとうございました！自分が学んだ時もその「平均符号長の下限からの乖離」くらいしか教えられなかったけど、こっちのほうが本当にわかりやすい！

黒木玄 **Gen Kuroki** @genkuroki on May 14
@8zu
すでにお読みになったかもしれませんが、

mathtod.online/@genkuroki/2104…
のようにmathtod.online経由で読めば数式が読み易くなります。