統計学: カルバック-ライブラ情報量とサノフの定理: 黒木さんのツイートまとめ

数学和尚 ダイナマイト関根 *

2019 年 9 月 29 日

1 統計学: カルバック-ライブラ情報量とサノフの定理: 黒木さんのツイートまとめ

1.1 ツイート

- URL

ここからの連続ツイートをまとめます. (自分にとって) 読みやすくするため, 適当に編集します.

1.2 はじめに: 大偏差原理について

黒木さんの PDF にもコメントがなかったと思うので, はじめに大偏差原理について書いておきましょう.

まずカルバック-ライブラ情報量とサノフの定理を大偏差原理の観点から説明します. サノフの定理は大偏差原理の 1 種 (簡単な 1 バージョン) です.

* phasetr@gmail.com

大偏差原理といってもいろいろなバージョンがあり, 各大偏差原理にはそれ特有のレート関数と呼ばれる関数があります. このレート関数として表れるのは典型的にはエントロピーです.

大偏差原理は大数の法則・中心極限定理と並ぶ確率論の基本定理です. 中心極限定理で正規分布が出てくるからこそ正規分布が大事だ, というのと同じように, その基本定理で出てくる関数だからエントロピーが大事, そういう構造にもなっています.

さて, 大偏差原理が何かという話をごく簡単にします. 詳しくは次の PDF や文献を見てください: ここではそれらから適当に記述をつまんできます.

- 福島竜輝, 大偏差原理について
- 福島竜輝, 大偏差原理の基本の “き”
- 若木宏文, ラプラス近似とその応用
- [1, Dembo, Zeitouni, Large Deviations Techniques and Applications]

Explanation 1.1

まずは Dembo, Zeitouni の Introduction から記述を抜き出してきます. 独立な標準正規分布の列 $X_1, \ldots, X_n$ を考え, これの経験平均 $\hat{S}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ を考えましょう. この $\hat{S}_n$ は平均 0, 分散 $1/n$ の正規分布で, 全ての $\delta > 0$ に対して次の大数の法則が成り立ちます.

$$\lim_{n\to\infty} P\left(\left|\hat{S}_n\right| \geq \delta\right) = 0. \tag{1.1}$$

さらに全ての区間 A に対して次の中心極限定理が成り立ちます.

$$\lim_{n\to\infty} P\left(\sqrt{n}\hat{S}_n \in A\right) = \frac{1}{\sqrt{2\pi}} \int_A e^{-x^2/2}\, dx. \tag{1.2}$$

ここで次の大雑把な評価に注意しましょう.

$$P\left(|\hat{S}_n| \geq \delta\right) = 1 - \int_{-\delta\sqrt{n}}^{\delta\sqrt{n}} e^{-x^2/2}\, dx. \tag{1.3}$$

ここから次の評価が出ます.

$$\lim_{n\to\infty}\frac{1}{n}\log P\left(|\hat{S}_n|\geq\delta\right)=-\frac{\delta^2}{2}. \tag{1.4}$$

これが大偏差原理の典型的な言明です: つまり $\hat{S}_n$ の典型的な値は中心極限定理によって $\frac{1}{\sqrt{n}}$ のオーダーである一方, $e^{-n\delta^2/2}$ というオーダーの小さな確率で $|\hat{S}_n|$ は比較的大きな値を取るのです.

1.3　ラプラス原理と大偏差原理

もう 1 つの視点として, ラプラス原理や鞍点法から見た視点も説明します.

Example 1.2

ラプラス原理は応用上よく出てくる次の形の積分評価です: 正値の連続関数 f,g に対して N が十分大きいときに次の漸近評価が成り立ちます.

$$\int_a^b e^{Nf(x)}g(x)\,dx\approx\exp\left[N\sup_{x\in[a,b]}f(x)\right]\quad(n\to\infty). \tag{1.5}$$

ここで $\approx$ は両辺の対数の比が 1 に収束するという意味です.

福島竜輝, 大偏差原理の基本の"き"にしたがって, バラダン (Varadhan) による大偏差原理の定式化を書いておきます. 確率測度は確率分布 (確率密度関数) と読み替えてかまいません. 位相空間の部分集合 A に対して $\operatorname{Int}A$ をその開核, $\operatorname{clos}A$ を閉包とします.

Definition 1.3

可分な距離空間 X 上の加法族 $\mathcal{F}$ 上の確率測度の族 $(\mu_n)_{n\in\mathbb{N}}$ が次の条件をみたすとき, 大偏差原理をみたすという: ある下半連続関数 $I\colon X\to[0,\infty)$

が存在して, 全ての可測集合 Γ に対して次の評価が成り立つ.

$$
\begin{aligned}
&- \inf_{x \in \operatorname{Int} \Gamma} I(x) && (1.6)\\
&\leq \liminf_{n \to \infty} \frac{1}{n} \log \mu_n(\Gamma) && (1.7)\\
&\leq \limsup_{n \to \infty} \frac{1}{n} \log \mu_n(\Gamma) && (1.8)\\
&\leq - \inf_{x \in \operatorname{clos} \Gamma} I(x). && (1.9)
\end{aligned}
$$

この I をレート関数 (rate function) と呼ぶ. 任意の非負の数 $l \geq 0$ に対してレベル集合 $\{x \in X \,|\, I(x) \leq l\}$ がコンパクトならば, I はよいレート関数 (good rate function) と呼ぶ.

大雑把に言えば大偏差原理が成り立つとき $d\mu_n(x) \approx e^{-nf(x)}$ が成り立つと言えます.

次の形でラプラス原理の一般化としてバラダンの補題が成り立ちます.

Lemma 1.4

可分な距離空間 X 上の加法族 $\mathcal{F}$ 上の確率測度の族 $(\mu_n)_{n \in \mathbb{N}}$ が I をよいレート関数として大偏差原理をみたすとき, X 上の有界な連続関数 f に対して次の評価が成り立つ.

$$
\int e^{nf(x)} \, d\mu_n(x) \approx \exp \left[n \sup_{x \in X} \left(f(x) - I(x) \right) \right]. \tag{1.10}
$$

1.4　大偏差原理小まとめ

ごちゃごちゃと書きましたが, まとめると次の 5 点です.

- サノフの定理は大偏差原理の 1 バージョンである.
- 大偏差原理はめったに取らない値を取る確率を問題にする.
- 大偏差原理は大数の法則・中心極限定理に並ぶ確率論の基本定理である.

- 大偏差原理はレート関数と呼ばれる関数を持ち, この関数が挙動を制御している.
- サノフの定理のレート関数としてカルバック-ライブラ情報量が出てくる.

次からツイートをまとめます.

1.5 カルバック-ライブラ情報量の謎

よい解説が見当たらなかったので自分で解説を書いた。

- Kullback-Leibler 情報量と Sanov の定理
- ツイート引用

  私も去年の今頃, 機械学習に入門して, いたるところで「KL ダイバージェンス」なる量が導入されていて, 「距離でもないしなんなんこいつ? 」みたいな感じで釈然としなかったのだが, これは機械学習入門あるあるなのではないだろうか.

Kullback-Leibler 情報量のよい解説を次のように定義しよう.

- Sanov の定理についての簡単な解説を含む. Kullback-Leibler 情報量 $D(q \| p)$ が「p による q のシミュレーションの予測誤差」を表すことの必然性の理解に必要.
- その応用としてカノニカル分布の導出も書いてある.
- 他にも色々書いてある.

一方で Kullback-Leibler 情報量の悪い解説を次のように定義しよう.

- Kullback-Leibler divergence をもっと一般の divergence の特別な場合としてのみ解説していて, Sanov の定理のような大偏差原理との関

係に言及していない.

- Sanov の定理に触れずに, 平均符号長との関係のみに触れている.

最近書いたよりコンパクトな Kullback-Leibler 情報量の解説が以下のファイルの最初の方にある.

- ベイズ統計の枠組みと解釈について
- Kullback-Leibler 情報量と記述統計

1.6 補足・関連

KL 情報量の Sanov の定理には「モデルの分布 p による分布 q のシミュレーションの予測誤差」という使い方の他に, 「真の事後分布から最も出てき易いより簡単な形の近似的な事後分布を求めるため」 (変分推論) にも使われます.

- Sanov の定理の使い方 2

  Sanov の定理の使い方 2: 事後分布 ϕ を直接計算するのが大変なので, 特別な形の分布 ψ で事前分布から最も出て来易いものを求めることを考える. そのためには $D(\psi \parallel \phi)$ を最小にする特別な形の分布 ψ を求めればよい. これが所謂「変分推論」 (平均場近似の一般化) です.

真の事後分布から最も出てき易いより簡単な形の近似的な事後分布を KL 情報量の意味で作ると, 近似的な事後分布の台は真の事後分布より小さくなります. 複雑な分布から簡単な形の分布で最も出て来やすいものを求めるとどういうことが起こりそうかについては次の連続ツイートを参照.

- 正規分布と混合正規分布の KL 情報量のプロット

例を色々知っていることが大事.

- 分布 $q(x)$ と同じ平均と分散を持つ正規分布は, $q(x)$ を最も小さな予測誤差でシミュレートする正規分布になる. ($q(x)$ が正規分布から程遠い場合にはベストな正規分布による予測誤差は大きくなる)
- 分布 $q(x)$ を最も小さな予測誤差でシミュレートする Laplace 分布 $p(x) = \frac{1}{2b}\exp\left[-\frac{|x-a|}{b})\right]$ の a は $q(x)$ の中央値になる.
- 混合正規分布から最も出て来やすい正規分布は混合正規分布のパラメーターについて不連続に変化する.
- などなど.

1.7 情報幾何をやっているカルバック-ライブラ情報量がわかる?

- ツイート

  KL ダイバージェンスが結局なんなのか, 情報幾何学やらないと永遠に分からないのでは? (適当)

情報幾何をやっても Kullback-Leibler 情報量のことは分からない. Fisher 情報行列は KL 情報量の Hessian matrix なので, KL 情報量を理解していないと, Fisher 情報行列を理解できない. だから, Kullback-Leibler 情報量について理解していないと情報幾何の出発点で躓く.

Fisher 情報行列が退化していると Fisher 情報行列の非退化性を仮定した理論全体が使えなくなってしまいますが, もとの Kullback-Leibler 情報量はそういう場合であっても非常に有効な数学的道具として働き, そのおかげで [2, 渡辺澄夫『ベイズ統計の理論と方法』] のような仕事が可能になったのです.

KL 情報量について理解していると, 統計学では通常天下り的に登場

する指数型分布族が自然に出て来る設定が分かる. $\int q(x)f_i(x)dx = c_i$ $(i = 1, \ldots, r)$ を満たす分布 $q(x)$ で, 分布 $p(x)$ から最も出て来やすいものは次のように書ける.

$$q(x) = \frac{1}{Z} \exp\left[-\sum \beta_i f_i(x)\right] p(x), \tag{1.11}$$

$$Z = \int \exp\left[-\sum \beta_i f_i(x)\right] p(x)\, dx. \tag{1.12}$$

1.8 指数型分布族とカルバック-ライブラ情報量

- ツイート

  KL ダイバージェンス, 実は指数型分布族じゃないとそんなに有効じゃない気もしている

Kullback-Leibler 情報量は指数型分布族ではないケースに非常に役に立つ. その典型例が [2, 渡辺澄夫『ベイズ統計の理論と方法』]. モデル $p(x|w)$ が指数型分布族だと KL 情報量の w 依存性が超絶易しくなる. 易しくない場合については [2, 渡辺澄夫『ベイズ統計の理論と方法』].

1.9 大偏差原理とエントロピー

エントロピー大と場合の数や確率大を同じようなものと思いたければ, Kullback-Leibler 情報量を相対エントロピーとは呼べなくなる. 相対エントロピーの定義は KL 情報量の -1 倍とするのが自然. 大偏差原理の話.

1.10 カルバック-ライブラ情報量の説明の仕方

- ツイート

> GAN の Adversarial Loss の構成とかだと, 割と明示的に KL ダイバージェンス (JS ダイバージェンス) を説明しなくちゃいけなくて, いかに細部を端折りながら雰囲気を伝えればいいのかは迷っていた (結局 2 つの確率分布における “妥当” な距離ですみたいな説明をしてしまっている)

このスレッド (注: この一連のツイートまとめ) に詳しく書きましたが (解説 PDF へのリンク付き), KL 情報量については Sanov の定理を知れば色々スッキリ理解できます. 対称化した JS div. が意味不明のことをしていることもわかる. KL 情報量 $D(q \,\|\, p)$ が p, q について非対称であることは極めて自然.

KL 情報量以外の “divergence” についても大偏差原理があれば分かりやすいのですが, その辺はどうなっているのでしょうか? 何らかの大偏差原理が無ければ divergence の統計学的な自然さが怪しくなると思う. それを最小化することの必然性や自然さがどこにあるかという問題.

分布 p の乱数列 $X_1, X_2, \ldots$ で分布 q のシミュレーションを行ったときの予測誤差 $D(q \,\|\, p)$ と, 分布 q の乱数列 $X_1, X_2, \ldots$ で分布 p のシミュレーションを行ったときの予測誤差 $D(p \,\|\, q)$ は一般に違うと考える方が自然. ゆえに, KL 情報量 $D(q \,\|\, p)$ が p, q について非対称であることは自然. Sanov の定理が重要.

参考文献

[1] Amir Dembo and Ofer Zeitouni. *Large Deviations Techniques and Applications*. Springer, 2009.

[2] 渡辺澄夫. 『ベイズ統計の理論と方法』. コロナ社, 3 2012.

索引

大偏差原理, 3

レート関数, 4