統計学: 尤度

数学和尚 ダイナマイト関根 *

2019 年 9 月 29 日

1 統計学: 尤度

1.1 はじめに

- 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.

元記事は `https://phasetr.com` に置いてあります.

1.2 尤度の基礎

「尤度」=「尤もらしさ」=「likelihood」という用語も相当にミスリーディングでかつ正常な理解を阻害していると思う. 尤度の定義は, モデルの確率分布 $p(x)$ がサンプル $X_1, \dots, X_n$ を生成する確率密度 $\prod_{k=1}^{n} p(X_k)$ です. これを「モデルのもっともらしさ」と呼んでしまうとまるで「モデルの正しさの度合い」のように聞こえてしまうのですが, 実際にはモデルが現実世界から得られたサンプルと同じサンプルを生成する確率密度が尤度の定義なの

* phasetr@gmail.com

で, 「モデルのサンプルへの確率的な適合度」に過ぎません.

我々が欲しいのはサンプルに適合するモデルではなく, サンプルを生成した未知の分布 (未知の法則) に適合するモデルです. その区別を明瞭にするためには, 「モデルの尤度」と呼ばずに「モデルのサンプルへの適合度」と呼ぶ方が分かりやすいし, 無用な誤解も防げる.

最尤法については「有限個のパラメーターを持つ確率モデルのサンプルへの適合度を最大にするパラメーターを求める推定法」と説明することができる. このように言い直すと「尤度」=「もっともらしさ」という用語を使った説明の分かりにくさが明らかになると思う. そして過学習については「サンプルへの適合度は上昇したがサンプルを生成した未知の分布への適合度が下がること」と説明できる. パラメーターを増やしたりハイパーパラメーターを調節してサンプル (学習データ) へのフィッティングを強化すると, 過学習が比較的容易に発生する.

数学的に定義された概念を理解するときには, 歴史的な経路によって固定された「尤度」=「もっともらしさ」=「likelihood」のような呼び名に惑わされることなく, 定義に戻ってすなおに解釈する方が無駄に混乱せずに納得できる場合が多い. 呼び名は多くの場合にミスリーディング.

1.3 コメントの返信

- URL

  「モデルの尤もらしさ」ではなく「パラメータの尤もらしさ」なら問題ないんじゃないのかしら. 人為的にモデルを選んだ上での尤もらしさである, というのを忘れるとダメなのは当然だけど, 「尤もらしさ」の用語が問題ではなく「モデルの」と修飾するのが問題なように思った.

「モデルのもっともらしさ」を最尤法の「モデルのパラメーターのもっともらしさ」と言い換えても大して変わらない. 理解するべき最重要ポイント

は, 尤度は我々が真に知りたいサンプルを生成した未知の分布ではなく, サンプルへの適合度に過ぎないことである.

サンプル=データに最も良くフィットするパラメーターを探しただけなのに, 「このパラメーターがもっともらしい」などと言ってしまうと誤解の原因になる. どんなに過学習を起こしていても「このパラメーターがもっともらしい」と言えてしまうのは良くない.

複数のモデルの情報量規準を計算して「この中ではこのモデルがもっともらしい」と言うのであれば, 「もっともらしい」という言葉を普通の意味で正しく使っていることになる. 尤度と情報量規準は違う. 尤度は決してもっともらしさではない. サンプル=データへの適合度に過ぎない.

1.4 ここまでのまとめ

確率分布 $p(x)$ のサンプル $X_1, \ldots, X_n$ に関する尤度の定義は, 確率分布 $p(x)$ が $X_1, \ldots, X_n$ を生成する確率密度 $\prod_{k=1}^{n} p(X_k)$ である.

尤度はサンプルへの適合度を表している. サンプルへの適合度が大きくなっても, サンプルを生成した未知の分布への適合度が小さくなることがある (過学習).

1.5 学習誤差, 汎化誤差

統計学における尤度の対数の $-1/n$ 倍は, 機械学習の用語では「学習誤差」=「訓練誤差」=「training error」と呼ばれることがあります. 学習誤差はモデルの分布によるサンプルの予測の誤差の指標になる. モデルの分布による未知の母集団分布の予測の誤差の指標は汎化誤差と呼ばれる.

ただし汎化誤差の計算には未知の母集団分布が使われるので, 現実の統計分析では計算できない. そこで代わりに使われるのが汎化誤差の推定量. 汎化誤差の推定量は統計学において情報量規準と呼ばれている.

1.6 区別するべき 3 つの量

(1) モデルの分布 $p(x)$ のサンプル $X_1, \ldots, X_n$ への適合度
(2) モデルの分布 $p(x)$ の未知の母集団分布への適合度 (現実の統計分析では計算不可能)
(3) モデルの分布 $p(x)$ の未知の母集団分布への適合度の推定量

この 3 つを区別したい. 尤度と呼ばれる量は (1) である.

我々が真に欲しいモデルの分布は未知の母集団分布に近い分布なので, 上記 (3) の「モデルの分布 $p(x)$ の未知の母集団分布への適合度の推定量」(情報量規準の-1 倍) を「$p(x)$ の真のもっともらしさ」だと思った方が「もっともらしさ」という用語の統計学的な使い方として正しいように思われる.

1.7 最尤法

最尤法の開発者はフィッシャーさんです. フィッシャーさんが最尤法を考えたときには, 過学習の問題にも配慮している情報量規準が存在しなかったので, 「もっともらしさ」を「サンプルへのフィットの度合い」として定義してしまったのは仕方がなかったかもしれません. ・ 統計学が過学習にも配慮した真の「もっともらしさ」=情報量規準の概念を明確にするには, 赤池弘次さんの登場を待つ必要があった. 赤池さんの 1980 年の論説は非常におすすめ! 相対エントロピー, KL 情報量, 大偏差原理などの概念が統計学に持ち込まれた.

- 参考連続ツイート

  1980 年の赤池弘次さんの 2 つの論説 赤池弘次, 1980, 統計的推論のパラダイムの変遷について, 赤池弘次, 1980, エントロピーとモデルの尤度. 赤池さんによれば Fisher さんは尤度 (ゆうど) について十分に

理解していなかった.

1.8 コメントへの応答

- URL

  モデルを 1 つ固定した上で「このモデルを選ぶ限りこのパラメータが最も尤もらしい」と言うのは, どんなに過学習が起きていても正しいから, パラメータ数の異なるモデル同士を比較するには拡張する余地があるとはいえ「尤度は尤もらしさの指標ではない」と言われると私はむしろ混乱する.

  【モデルを 1 つ固定した上で「このモデルを選ぶ限りこのパラメータが最も尤もらしい」と言うのは, どんなに過学習が起きていても正しい】 正しいわけがない. 過学習が起こっている疑いがあるケースで尤度を「もっともらしさ」の指標に使うのは自明に誤り. 難しく考える必要はない. 定義に忠実に解釈して余計なことを考えないだけでよい. 歴史的に「尤度」と呼ばれているものの定義は「モデルの確率分布が得られたサンプルと同じサンプルを生成する確率密度」でしかない. 一言で言えば「モデルの分布のサンプルへの適合度」. モデルの確率分布のサンプルへの適合度はモデルの確率分布の (未知の) 母集団分布への適合度とは異なる. モデルの確率分布のサンプルへの適合度が高くても, モデルの (未知の) 確率分布の母集団分布への適合度が低い例は簡単に作れる.

実際にはノイズに過ぎないサンプルの詳細な形状に「真実」を「発見」した人が「真実」にもフィットするようにパラメーターを調節できるモデルを設定して最尤法を実行すれば, 「真実」にフィットした高い尤度が得られるだろう. しかしそれと同時に誤差の大きな予測分布が出来上がる. 一方, その

「真実」を疑わしいと思っていた別の人は, 同じモデルの別のパラメーターの方がもっともらしいと思っていた. そして, その人は尤度の高さは「もっともらしさそのもの」ではないこともよく理解していた.

1.9 データへの適合=予測精度の向上ではない

- URL

  時系列予測の問題において, 機械学習のモデルより既存の統計モデル (ARMA モデルなど) の方が予測精度において優良な結果が出るという研究. データへの適合=予測精度の向上ではないことも実験で示している. 機械学習の研究では統計モデルとの比較も入れるべきという提言をしている. URL.

1.10 過学習の例

- URL nbviewer

動画の右半分の赤の線が下がると尤度は高くなる. 青の線は予測誤差 + 定数. 尤度を最大にするパラメーターを探す過程のある所から先では, 新たに「真実」が発見されて赤の線がぴょこっと下がる (尤度が上がる) たびに予測誤差もぴょこっと大きくなる.

- training error = 対数尤度の $-1/n$ 倍
- generalization error = 予測誤差 + 定数

予測誤差をさげたい. しかし, 尤度を最大にするパラメーターを探す過程で尤度が上がると, 予測誤差が悪化することがよくある.

- 動画へのリンク

以上の動画の様子は「サンプルへの適合度 (尤度) を最大にするパラメーターを探す過程」では典型的です. 途中で新たに「詳細な構造」がぴょこっと「発見」されるたびに予測誤差が悪化するというようなことが起こる. これが過学習の典型的なパターンです. 動画の「ぴょこっ」に注目!

1.11 汎化誤差との逆相関

汎化誤差との逆相関は AIC や WAIC や LOOCV でも起こります. ただしその意味での逆相関は上の動画のそれとは違って, サンプルが確率変数であることに由来する揺らぎの逆相関です. そのことが原因でモデル選択に失敗する場合がある. 汎化誤差の推定は極めて重要かつ相当に困難な問題.

1.12 過学習を起こしていない場合の動画

- URL

1.13 汎化誤差

私は, 統計学を学ぶときには, 歴史的な理由で尤度と呼ばれることになったものよりも先に汎化誤差の概念を学ぶべきだと思う. 我々が欲しいのはサンプルに適合したモデルではなく, 母集団分布に適合するモデルであり, 母集団分布への適合度は汎化誤差の小ささで測られる. 汎化誤差の概念を理解することは, ほぼ Kullback-Leibler 情報量の Sanov の定理を理解することと同じになります. だから私は統計学で使う確率論の三種の神器は次の 3 つだと主張している.

- 大数の法則
- 中心極限定理
- Sanov の定理

Sanov の定理は普通教えてくれない.

最近作ったリポジトリ に置いてある易しく書いたつもりの 2 つの解説ノート.

- ベイズ統計の枠組みと解釈について
- ベイズ統計の手書きのノート

これらも KL 情報量の Sanov の定理の話から始めている.

1.14 用語法

用語法について実は少し悩んでいて, 汎化誤差の式を一般的な場合に使用するときに「汎化誤差」 (generalization error) という呼び名が適切かどうか? 私は $G(q \parallel p) = \int q(x)(-\log p(x))dx$ に「○○誤差」という名前をつけたい. 現時点では「汎化誤差」と呼んでいる.

確率分布 $p(x)$ が過去に得られた情報から作った未知の分布 $q(x)$ の予測モデルの分布のとき, $p(x)$ による分布 $q(x)$ の予測誤差の大小は $G(q \parallel p)$ の大小で判別できます. 過去に得られたデータを超えた $p(x)$ の予測精度を測るので, これは「汎化誤差」の名に恥じない.

しかし $G(q \parallel p)$ を $q(x)$ がサンプル $X_1, \ldots, X_n$ の分布 $q(x) = \frac{1}{n}\sum_{k=1}^{n}\delta(x - X_k)$ の場合にも使いたい. そのとき $G(q \parallel p) = \frac{1}{n}\sum_{k=1}^{n}(-\log p(X_k))$ は対数尤度の $-1/n$ 倍の「学習誤差」です. さすがにこれは「汎化誤差」とは呼べない.

$G(q \parallel p) = \int q(x)(-\log p(x))\,dx$ は $p(x)$ による $q(x)$ への適合度の指標 (小さいほどよく適合している) になっています. もしくは Sanov の定理によって $p(x)$ による $q(x)$ のシミュレーションの誤差の大きさと言ってもよい. $G(q \parallel p)$ は「汎化誤差」にも「学習誤差」にもなり得る. どう呼んだものか.

1.15 呼び方の試案

- $G(q \parallel p) =$ 分布 p の分布 q への不適合度
- 汎化誤差 $=$ 予測分布の母集団分布への不適合度
- 学習誤差 $=$ 対数尤度の $-1/n$ 倍 $=$ モデルの分布のサンプルへの不適合度

尤度は「モデルの分布のサンプルへの適合度」の一種. 不適合度は英語では badness of fit で良さそうかな.

1.16 コメントへの返信

- URL

  モデルを固定する, という条件付きの最適化問題では尤度と情報量基底のどちらを使っても最適点は一致するのだから, 「同じモデルで最尤点よりさらに尤もらしいパラメータが存在する」という状況はありえない. 条件を緩和したときに, より尤もらしさの高い点が存在する可能性があるというだけ.

【情報量基底】というような意味不明の用語を使わずに済むようになってから返答した方がよいと思いました. 【「同じモデルで最尤点よりさらに尤もらしいパラメータが存在する」という状況はありえない】というのは明瞭に誤り. すでにそういう例を上で示した.

おそらく, もっともらしさを測る方法を歴史的に尤度と呼ばれることになった量を使う以外に知らないのだと思います. だから「尤度を最大化するパラメーター=最ももっともらしいパラメーター」以外の考え方をできなくなっているのではないか? 推定用とテスト用のデータを分けておいて, 推定で得られた予測分布の精度をテストデータで測るというようなことはよくや

られています. モデルの尤度を最大化するパラメーターに対応する予測分布より, そうでないパラメーターに対応する予測分布の方が予測精度が高いことは普通にある.

あまりにもアレな話でよく知らないのですが, 「予測精度を上げるための学習=推定の過程の early stopping」という話が大真面目にされている世界がある. early stopping の話を初めて知ったときにはかなりジワった.

何度目の繰り返しになるのか分かりませんが, 我々が欲しいのは未知の母集団分布によくフィットしている可能性が高い分布であり, 推定用のデータの側によくフィットしている分布ではありません. 尤度最大化はモデルのパラメーターを調節して推定用のデータの側にフィットさせているだけ.

このスレッドを立てるときには「何を当たり前にことを言っているんだ? 藁人形を批判しても意味がないだろ!」と非難されることを恐れたのですが, 自ら実例になってくれる人が出て来たのでそう言われるリスクはだいぶ減ったかもしれませんね. データへの単なるフィッティングと予測精度に気を配った推測は全然違うという事実は非常に大事.

1.17 姉妹編へのリンク

- 連続ツイートへのリンク

1.18 Google で early stopping を検索

- URL

何度見てもジワる. 人生の大変さを感じる.

- 関連連続ツイート

1.19 質問への回答

- URL

  ウィキペディアによると, 尤度関数と確率密度関数は別物とありますが, 尤度が確率密度というのは妥当ですか?

「尤度関数と確率密度関数は別物」と私の発言は整合的です. だから問題なし. そうであることがわかるまで自分の頭で考えてみるとよいと思います.

パラメーター w を持つ x の確率密度函数 $p(x|w)$ はパラメーター w を決めると x について確率密度函数になる. サンプル X_1 に対して, パラメーター w の函数 $L(w) = p(X_1|w)$ はサイズ 1 のサンプルで決まる尤度函数になります. 尤度函数はパラメーター w をサンプル X_1 における確率密度 $p(X_1|w)$ に対応させる函数.

Problem 1.1
サイズ n のサンプル $X_1, \ldots, X_n$ に関する確率モデル $p(x|w)$ の尤度函数について説明せよ.

模範解答は書きません. 解答を書いても採点もしません.

参考文献