Keywords you should know. 第111回

知っておきたいキーワード

ノンパラメトリックベイズ

上田修功†

† NTTコミュニケーション 科学基礎研究所
"Nonparametric Bayes" by Naonori Ueda (NTT Communication Science Laboratories, Tokyo)
キーワード：ノンパラメトリックベイズ，機械学習，ベイズ統計

学生　先日，ノンパラメトリックベイズを用いた画像データのクラスタリングという論文を見つけましたが，ディリクレ過程といった専門用語や数式の嵐でまったく意味不明です．機械学習や統計の素人の私にもわかるように数式なしでノンパラメトリックベイズのエッセンスを教えて頂けないでしょうか？

先生　まずはベイズ推定から説明する必要がありますね．ベイズ推定とは，観測データの背後にある統計モデル（確率分布）のパラメータの事後分布を推定する方法です．この時，事前分布を導入してベイズの定理を用いて事後分布を求めるため，ベイズ推定と呼ばれます．

学生　すみません．具体例でもっとわかりやすい説明お願いします．

先生　失礼．それでは，例えば，ある目が他の目よりも高頻度で出るような不平等なサイコロを考えましょう．ベイズ推定では，サイコロの出る目を確率変数と見なし，それが従う分布のパラメータ（不平等サイコロの場合，各目の出る確率）を観測データから推定します．この時，もしあなたが“サイコロの目の出る確率は同じ”という事前知識をパラメータの事前の情報（事前確率分布）として利用すれば，観測データを得た後にこの事前確率がどう変わるかをベイズの定理を用いて事後確率分布として計算できます．

学生　ちょっと難しくなってきましたが，「事前分布を定義し，観測データからベイズの定理を用いて事後分布を計算する手順がベイズ推定」，という理解で良いですか？

先生　その通りです．観測データを得たことにより，事前知識である事前分布が事後分布という形でより現実に則した情報に修正されるわけです．

学生　観測データのみから推定してはいけないのですか？ なぜ，事前分布が必要なのですか？

先生　最尤（さいゆう）推定と呼ばれる事前分布を用いない推定法もあります．一般に，学習データ数が少数の場合は推定精度が劣化します．学習データが少数の場合でも，事前分布を付加情報として利用することで推定精度の劣化を抑止しようというのがベイズ学習の利点といえるでしょう．事前分布は先入観みたいなものですね．ついでながら，最尤推定では，パラメータの値そのものを推定するのに対し，ベイズ推定では，パラメータを確率変数と見なして，確率分布（事後分布）として推定できるため，信頼度なども同時に評価できるようになります．これが両者の大きな相違点です．

学生　わかりました．では，ノンパラメトリックベイズは，字面から推測すると，パラメータがないベイズ学習ですか？　パラメータがないのにパラメータの事後分布を計算するとはどういうことですか？

先生　良い質問です．ノンパラメトリックという言葉は，統計学では，分布を仮定しない統計分析手法の総称ですが，実は，ノンパラメトリックベイズにおける“ノンパラメトリック”の意味はそれとはまったく異なります．ノンパラメトリックベイズは，モデルの複雑さ，つまり，モデルのパラメータ数がデータに応じて自動調整され，原理的には無限個のパラメータを想定できる柔軟な統計モデルです．例えば，画像のクラスタリングでガウス混合モデルを用いる際，混合数をデータから自動学習できます．

学生　それは便利ですね！私が見た論文は，まさにそれを実現していたのですね．自動学習が可能になる仕組みをもう少し説明お願いします．

先生　具体的にクラスタリングの例で説明しましょう．☞

映像情報メディア学会誌 Vol. 70, No. 3, pp. 478～480（2016）

☞ ベイズ推定では事前分布という考え方が特徴的だと先に説明しましたが，ノンパラメトリックベイズは，モデルの複雑さの事前分布を考えるベイズ推定といえます．

学生　モデルの複雑さとはクラスタリングの場合，クラスタ数のことですか？

先生　そうです．

学生　クラスタが1個である確率，クラスタが2個である確率，という風なものだとすると，すべての可能性，つまり，クラスタ数が無限個までの確率を考えることになるのですか？ 事前には観測データが何個観測されるかもわからないので，クラスタ数が観測データ数よりも多いような場合の確率分布など事前に定義できるのでしょうか？

先生　質問が鋭くなってきましたね．実は，それを可能にするのがノンパラメトリックベイズの本質です．もちろん，データが何個観測されるかは事前にわからないので，クラスタ数の事前分布を確率過程としてモデル化します．つまり，観測データが逐次的に得られたとして，それらがどのクラスタに属するかが逐次的に得られ，結果としてクラスタリングされるわけです．この時，クラスタ数は予め固定されているのではなく，データ数に応じて適応的に増加するという事前分布です．

学生　抽象的で良くわからないので，先のサイコロの例でより具体的に説明お願いします．

先生　不平等サイコロの場合，ベイズ推定では1から6の出る目の確率の事前分布も考えられますが，ノンパラメトリックベイズでは，モデルの複雑さ（構造）の事前分布なので，出る目の確率の事前分布ではなく，不平等サイコロそのものの分布が構造の事前分布に相当します．

学生　サイコロの分布とは，出る目が1から6のサイコロだけではなく，7，8，9…の目もあるサイコロを想定し，それらがどのような確率で起こり得るかの確率分布ということでしょうか？

先生　そうです．そして，その不平等サイコロの生成過程を実現する確率過程が，君が最初にいっていたディリクレ過程です．

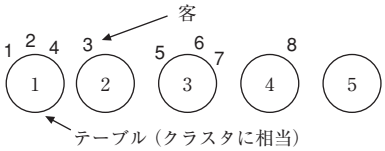
学生　まだ良く理解できていませんが，そのまえに，なぜ，ディリクレ過程と呼ばれるのですか？

先生　一般に，非負でかつ総和が1となる複数の確率変数の同時分布はディリクレ分布で表現できます．例えば，1から6の目の不平等サイコロの場合，各目が出る確率は当然非負で，かつ，1から6までの目の出る確率の総和は1なので，各目の出る確率の同時分布（この場合，6つの確率が同時にその値になる分布）がディリクレ分布から生成できるということです．言い換えると，ディリクレ分布とは，複数個（ただし固定）のシンボル（サイコロの場合目の数に相当）の生起確率の同時分布です．ですので，ディリクレ分布のパラメータをいろいろ変えることにより，各目の生起確率の組が異なる多様な不平等サイコロ（1が出やすいサイコロ，偶数が出にくいサイコロなど）を確率的に生成できると考えてください．ただし，ディリクレ分布では，目の数の種類は固定されていて，固定された目の中で，出る目の確率が多様なサイコロしか生成できません．これに対し，ディリクレ過程では，出る目の種類も多様な不平等サイコロも生成できます．つまり，7，8，9…の目も出るサイコロが生成できます．そして，ディリクレ過程にも基底分布と集中度パラメータという2つのパラメータがあり，これらのパラメータを変えることで不平等サイコロの多様性をコントロールすることができます．

つまり，ディリクレ過程とは，固定数の離散シンボルの生起確率の同時分布であるディリクレ分布を，理論的には，無限個の離散シンボルの生起確率の同時分布に拡張した確率過程ということです．

学生　ディリクレ分布の無限次元拡張ということですね．では，実際どのような確率過程なのですか？

先生　佳境に入ってきましたね．最も良く用いられるディリクレ過程の実現方法として，中華料理レストラン過程（Chinese Restaurant Process: CRP）があります．



第9番目の客は，各々，3，1，3，1，γに比例する確率で各々テーブル1，2，3，4，5（新規テーブル）を選択する

図1　CRP
これまでに多くの客に占有されているテーブルほど，選択確率が高くなるが，確率γで新規テーブルも選択される．

学生　中国人がなぜ関係するのでしょうか？？

先生　単なる比喩です．巨大な中華レストランがあり，無限個のテーブルがあると想定します．そこに客が1人ずつ訪れ，逐次，テーブルに着席していきます．今，第i-1番目までの客がすでに着席しているとします．この時，第i番目の客は，**図1**に示すように，すでに着席している人数の比に比例する確率であるテーブルに着席します．ただし，テーブル選択はすでに着席しているテーブルだけでなく，先ほどの集中度パラメータ（図1のγ）に比例する確率で誰も着席していないテーブルにも着席するとします．1つのテーブルに着席できる人数の制限はないものとすると，n人の客が逐次このルールにしたがって着席し，全員が着席し終えた時，K個のテーブルが客に占有された場合，n人はK個のクラスタ（グループ）にクラスタリングされたことになります．

学生　中国人はお話し好きなので，多くの人が着席しているテーブルを好むという比喩ですね．また，γの値を大きくすればするほどテーブル数（クラスタ数）が増大するわけですね．

先生　そして，先のルールでは，i-1人が着席した後，第i客が第kテーブルに着席する確率がルールとして与えられているので，$i = 1, 2, \ldots, n$で逐次計算することにより，n人の客の着席パターン（例えば，$n = 5$では，☞

☞ 5人の客が，順にテーブル番号1, 1, 2, 2, 3に着席した，あるいは，1, 2, 3, 3, 4に着席したなど）の同時確率分布（n人の客のすべての着席パターンの同時分布）が計算できます．ここで重要かつ興味深い結果として，その同時分布は，n人の客の着席順に依らず，各テーブルに何人着席したかのみに依存するという点です．この性質はディリクレ過程における交換可能性と呼ばれる極めて重要な性質です．

学生　それがなぜ重要なのですか？

先生　クラスタリングの事前分布とは，あくまでクラスタ構造の分布です．ですので，客個人は区別せず，もし，5人の客が順にテーブル番号1，1，2，2，3と着席した場合と，2，1，3，1，2と着席した場合，テーブルを1つのクラスタと見なすと両者はクラスタ1に2人，クラスタ2に2人，クラスタ3に1人ということになり，クラスタリングという点では同一です．にもかかわらず，両者の確率が異なるとクラスタリングの事前分布として適切ではないですね．

学生　なるほど納得しました．でも，そもそもクラスタリングはCRPのようなルールでクラスタリングされるのではなく，データの類似性でクラスタリングされるのではないのでしょうか？

先生　これもよくある誤解です．今考えているのは，クラスタリングの構造の事前分布です．構造の事前分布とはデータを観測する前に，もしn個のデータを観測したとしたら，それらn個のデータをどのようにクラスタリングするのが確率分布として妥当かを議論しています．ですので，先に説明したとおり，実際にデータを観測した際，この事前分布と観測データに対するモデルとをベイズの定理で1つにまとめて事後分布を計算して，事後分布が最大となるクラスタリングを求めることになります．データの特徴の類似性などは観測データに対するモデルの中で反映され，先のCRPはクラスタ構造に関する事前分布に過ぎません．

学生　よくわかりました．つまり，CRPはデータの分割の事前分布を与え，この事前分布とデータの確率モデルと融合することで，データの特徴とデータ数に応じて，事後分布最大化という観点でベイズ的に最適な個数のクラスタリングが実現できるということですね！また，CRPのルールから，データ数が増えるにつれてクラスタ数も増大するというのは自然ですね．

先生　そのとおり！以上が直観的に理解できればノンパラメトリックベイズの本質は理解できたといえます．実際の学習アルゴリズムなどは数式をごちゃごちゃ計算するだけです．また，ディリクレ過程の実現例としてCRP以外にもホップの壺モデルや棒折り過程などがありますが，本質は同じです．また，通常のクラスタリングでは1つのデータは1つのクラスタのみに属すると仮定しますが，複数のクラスタに属する多重クラスタリングも考えられます．例えば，趣味ではAさんと同じグループに属すが，仕事ではBさんと同じグループに属すなど．このような多重クラスタリングの事前分布としてインド料理過程と呼ばれる確率過程があります．これもノンパラメトリックベイズの一種です．

学生　また比喩ですね．インド料理店のランチではバイキング形式でいくつかの料理を同時に選びますが，皿をデータ，皿上の料理をクラスタに各々対応させれば，多重クラスタリングが実現できそうですね．

先生　感が鋭いですね．その通りです．この確率過程も含め，ノンパラメトリックベイズの参考書などをあとで教えておきます．

学生　興味がわいてきました．勉強してみます．どうもありがとうございました．

（2016年1月29日受付）

参　考　文　献

1）T.S. Ferguson: "A Bayesian analysis of some nonparametric problems", The Analas of Statistics, 1, 1, pp.209-230（1973）
2）上田修功，山田武士：“ノンパラメトリックベイズモデル”，日本応用数理学会，17，3，pp.196-214（1997）
3）石井健一郎，上田修功：“続・わかりやすいパターン認識”，オーム社（2014）

上田（うえだ）　修功（なおのり）　1982年，大阪大学工学部通信工学科卒業．1984年，同大学院通信工学専攻修士課程修了．1984年，日本電信電話公社（現NTT）横須賀電気通信研究所入所．1991年，NTTコミュニケーション科学研究所主任研究員．1993年～1994年，米国Purdue大学客員研究員．1998年，NTTコミュニケーション科学研究所特別研究員．2003年，同所知能情報研究部長．2006年，同所協創情報研究部長．2010年～2013年，同所所長．2013年，同所機械学習・データ科学センタ代表．現在，NTTフェロー，特別研究室長．国立情報学研究所客員教授，京都大学大学院情報学研究科連携教授．博士（工学）．