これまでのまとめ

ベイズ推論 *(Bayesian inference)* の仕組みをまとめる．
現象や観測に関するデータを $\mathcal{D}$ とする．我々はそれを説明するモデルや仮説を作る．そのモデルを $\mathcal{M}$ と書こう．モデルは我々が作ったものであるから，それに含まれるパラメータや初期条件・境界条件など (まとめてパラメータとよぶことにする) を決めればモデルからデータを生成する，すなわち「シミュレーション」$\mathcal{M} \to \mathcal{D}$ を実行できる．しかし，実際に必要なことは，データからモデルのパラメータについて推論すること，すなわち「逆シミュレーション」$\mathcal{D} \to \mathcal{M}$ である．
そこで以下のようにベイズの定理を用いる．

$$P(\mathcal{M}|\mathcal{D}) = \frac{\mathcal{P}(\mathcal{D}|\mathcal{M})\,\mathcal{P}(\mathcal{M})}{\mathcal{P}(\mathcal{D})} \tag{1}$$

(1) 式の右辺の各因子は以下の意味をもつ．

- 分子の 1 番目の因子 $P(\mathcal{D}|\mathcal{M})$ は「シミュレーション」$\mathcal{M} \to \mathcal{D}$ から計算できるはずだ．この因子は**尤度** *(likelihood)* とよばれる[a]．
- 分子の 2 番目の因子 $P(\mathcal{M})$ はモデルのパラメータについて，我々が事前に知っている情報を組み込むことができるものである．この因子は**事前確率** *(prior probability)* とよばれる．
- 分母の因子 $P(\mathcal{D})$ は規格化 (normalization) 因子，すなわち (1) の両辺を $\mathcal{M}$ に含まれるパラメータの可能なすべての値について積分すると全確率 (total probability) 1 になることを保証するためのものである．したがって

  $$P(\mathcal{D}) = \sum_{\mathcal{M}} \mathcal{P}(\mathcal{D}|\mathcal{M})\,\mathcal{P}(\mathcal{M}) \tag{2}$$

  $P(\mathcal{D})$ は周辺確率であるから，$\mathcal{M}$ を含んでいないので，モデルのパラメータを推定する話には関係がない[b]．

(1) 式の右辺を計算した結果として得られる左辺を**事後確率** *(posterior probability)* とよぶ．

[a]「尤」は「もっともらしい」の意味 (犬ではない) で，尤度は「ゆうど」と読む．
[b]複数のモデルを比較して，どのモデルがもっともらしいかを決めるときにはこの因子 (evidence とよばれる) が有用となる．

藤原 義久 / yoshi.fujiwara@gmail.com

例: 簡単なテキスト解析「森鴎外か夏目漱石か」

ベイズ推論の応用として，ナイーブベイズ (naive Bayes) とよばれる推論の例をやってみよう．ナイーブベイズは，スパムメールのフィルター (spam filter)，文書からのトピックの同定のためのモデル (topic model) など，さまざまな分類器 (classifier) として広く使われている．

付録の文章は森鴎外 (1862–1922, 文久 2 年–大正 11 年) と夏目漱石 (1867–1916, 慶応 3 年–大正 5 年) の作品からの抜粋である．どの作品がどちらの作家によるものか分かるだろうか.

データ $\mathcal{D}$ ：森鴎外と夏目漱石の作品 (文書)．各文書 d は，単語 w からなっていると見なせる．すなわち

$$d = \{w_{d1}, w_{d2}, \cdots, w_{dn_d}\} \tag{3}$$

ここで，w_{di} は文書 d 中で i 番目に出現する単語で，ある語彙 (辞書)V から選ばれている，すなわち $w_{di} \in V$ としよう．文書群を D とすると，各行が $d \in D$，各列が単語 $w \in V$ であり，d 行目かつ w 列目の要素が d に w が出現した頻度であるような巨大な行列としてデータを表現できる．もちろん，この行列は多くの要素が 0 であろう．これを次式で表す.

$$f_{dw} = \text{文書 } d \text{ に単語 } w \text{ が出現する頻度} \qquad (d \in D, w \in V) \tag{4}$$

なお，文書 d に単語 w が出てこなければもちろん $f_{dw} = 0$ である.

モデル $\mathcal{M}$ ：森鴎外か夏目漱石かを未知のパラメータ $\theta = 0, 1$ とする，$\mathcal{D}$ を生成するモデルを作りたい．モデルにはいろいろな可能性があるだろう[1]．ここでは以下の簡単なモデルを考えよう．作家が文章を書くとき，言葉，単語を選ぶが，その選択は作家ごとで違うはずだ．ある作家は「きらびやかな」をよく用いるが,「壮麗な」はめったに使わないかもしれない．一方，単語が文章中に出現する順番や組み合わせも重要であるかもしれないが，それは無視することにして，一つの文書は単語をバラバラにして一つの袋に詰めたようなものであると見なすことにしよう．これを「袋仮説」とでも名付けよう.

尤度は以下のように考えれば経験的なやり方で計算できる．もし

$$P(w|\theta) = \text{作家 } \theta \text{ が分かったとき，その作家が単語 } w \text{ を使う確率} \tag{5}$$

を構成することができれば，それにしたがって文章を生成すればよい．しかし，(5) はどのように構成するか．いま，モデルを作るための訓練データ (training data) として，いくつかの作品はその作家が分かっているとしよう．訓練データから使われている単語の頻度を調べることができれば，その経験的な値をもって (5) を構成することができるだろう.

頻度 (4) を用いると，(5) は次のように書き直せることが分かる.

$$P(w|\theta) = \frac{\sum_{d \in \mathrm{train}(\theta)} f_{dw}}{\sum_{w \in V} \sum_{d \in \mathrm{train}(\theta)} f_{dw}} \tag{6}$$

ここで，$d \in \mathrm{train}(\theta)$ は，作家 θ に対応する訓練データの全文書 d を意味する.

ベイズ推論 $\mathcal{D} \to \mathcal{M}$ ：いよいよ，作家が未知の作品，すなわちモデルをテストしたいデータ (test data)d があったとして，θ を推定してみよう.

まず，どちらの作家か見当もつかないとすれば，事前確率は

$$P(\theta = 0) = P(\theta = 1) = 1/2 \tag{7}$$

[1]読者自らまず考えてみよ．実にいろいろなモデル，すなわち妄想を考えることができるはずだ.

藤原 義久 / yoshi.fujiwara@gmail.com

とするのが自然だろう．(1) はいまの場合，以下のように書ける．

$$P(\theta|d) = \frac{P(d|\theta)\,P(\theta)}{P(d)} \tag{8}$$

新しい文書 d が与えられたとき，作家が θ である条件付確率が計算できるので，$P(\theta = 0|d) > P(\theta = 1|d)$ であれば $\theta = 0$ が，逆の場合は $\theta = 1$ であろうと推論できる．なお，分母 $P(d)$ は前述の通り推定とは無関係であるから以下では無視しよう．

「袋仮説」から単語の出現は統計的に独立であるから

$$P(d|\theta) = \prod_{w_{d\cdot}} P(w_{d\cdot}|\theta) \tag{9}$$

と計算できる．ここで $w_{d\cdot}$ は，出現の順番はともかく，d に含まれる単語を表す．(9) の両辺の対数をとってから，頻度 (4) を用いると，積の対数はそれぞれの対数の和であるから

$$\log P(d|\theta) = \sum_{w_{d\cdot}} \log P(w_{d\cdot}|\theta) = \sum_{w \in V} f_{dw} \log P(w|\theta) \tag{10}$$

と書き直せることが分かる．二つ目の和は語彙 V のすべての単語 w について取っていることに注意しよう．右辺の $P(w|\theta)$ は (6) でモデルとして構成されている．

以上により，事後確率を計算することができた．ただし，新しい文書 d に未知の単語 w が出現した場合，$f_{dw} = 0$ であるから，(6) をそのまま使うと $P(w|\theta) = 0$ となってしまう，という技術的な問題はある．一つの方法は「まだ観測はされてないが，作家は未使用の単語も使う可能性がわずかだがある」として，すべての頻度を 1 だけ水増しして，(6) の代わりに以下の式を適用することである．

$$P(w|\theta) = \frac{1 + \sum_{d \in \mathrm{train}(\theta)} f_{dw}}{\sum_{w \in V}(1 + \sum_{d \in \mathrm{train}(\theta)} f_{dw})} \tag{11}$$

実際の計算例では (11) を使った．

参考文献

[1] 石田 基広「R によるテキストマイニング入門 (第 2 版)」(森北出版, 2017).
「森鴎外か夏目漱石か」という例題はここで取り上げられているが，本稿は独自にプログラミング言語 Python で最初から書き上げたもので，引用文献よりももっと初歩的な取扱いである．

[2] 岩田 具治「トピックモデル」(講談社, 2015).
分かりやすい教科書．数学的な説明やアルゴリズムなどを豊富に含む．

藤原 義久 / yoshi.fujiwara@gmail.com

付録

以下の 8 つの文章はいずれも冒頭のみの抜粋[2]

01.txt

古い話である。僕は偶然それが明治十三年の出来事だと云うことを記憶している。どうして年をはっきり覚えているかと云うと、その頃僕は東京大学の鉄門の真向いにあった、上条と云う下宿屋に、この話の主人公と壁一つ隔てた隣同士になって住んでいたからである。その上条が明治十四年に自火で焼けた時、僕も焼け出された一人であった。その火事のあった前年の出来事だと云うことを、僕は覚えているからである。

上条に下宿しているものは大抵医科大学の学生ばかりで、その外は大学の附属病院に通う患者なんぞであった。大抵どの下宿屋にも特別に幅を利かせている客があるもので、そう云う客は第一金廻りが好く、小気が利いていて、お上さんが箱火鉢を控えて据わっている前の廊下を通るときは、きっと声を掛ける。時々はその箱火鉢の向側にしゃがんで、世間話の一つもする。部屋で酒盛をして、わざわざ肴を拵えさせたり何かして、お上さんに面倒を見させ、我儘をするようでいて、実は帳場に得の附くようにする。先ずざっとこう云う性の男が尊敬を受け、それに乗じて威福を擅にすると云うのが常である。然るに上条で幅を利かせている、僕の壁隣の男は頗る趣を殊にしていた。

この男は岡田と云う学生で、僕より一学年若いのだから、とにかくもう卒業に手が届いていた。岡田がどんな男だと云うことを説明するには、その手近な、際立った性質から語り始めなくてはならない。それは美男だと云うことである。色の蒼い、ひょろひょろした美男ではない。血色が好くて、体格ががっしりしていた。僕はあんな顔の男を見たことが殆ど無い。強いて求めれば、大分あの頃から後になって、僕は青年時代の川上眉山と心安くなった。あのとうとう窮境に陥って悲惨の最期を遂げた文士の川上である。あれの青年時代が一寸岡田に似ていた。尤も当時競漕の選手になっていた岡田は、体格ではかに川上なんぞに優っていたのである。

02.txt

朝小間使の雪が火鉢に火を入れに来た時、奥さんが不安らしい顔をして、「秀麿の部屋にはゆうべも又電気が附いていたね」と云った。

「おや。さようでございましたか。先っき瓦斯煖炉に火を附けにまいりました時は、明りはお消しになって、お床の中で煙草を召し上がっていらっしゃいました。」

雪はこの返事をしながら、戸を開けて自分が這入った時、大きい葉巻の火が、暗い部屋の、しんとしている中で、ぼうっと明るくなっては、又微かになっていた事を思い出して、折々あることではあるが、今朝もはっと思って、「おや」と口に出そうであったのを呑み込んだ、その瞬間の事を思い浮べていた。「そうかい」と云って、奥さんは雪が火を活けて、大きい枠火鉢の中の、真っ白い灰を綺麗に、盛り上げたようにして置いて、起って行くのを、やはり不安な顔をして、見送っていた。邸では瓦斯が勝手にまで使ってあるのに、奥さんは逆上せると云って、炭火に当っているのである。

電燈は邸ではどの寝間にも夜どおし附いている。しかし秀麿は寝る時必ず消して寝る習慣を持っているので、それが附いていれば、又徹夜して本を読んでいたと云うことが分かる。それで奥さんは手水に起きる度に、廊下から見て、秀麿のいる洋室の窓の隙から、火の光の漏れるのを気にしているのである。

03.txt

石田小介が少佐参謀になって小倉に着任したのは六月二十四日であった。

徳山と門司との間を交通している蒸汽船から上がったのが午前三時である。地方の軍隊は送迎がなかなか手厚いことを知っていたから、石田はその頃の通常礼装というのをして、勲章を佩びていた。故参の大尉参謀が同僚を代表して桟橋まで来ていた。

雨がどっどと降っている。これから小倉までは汽車で一時間は掛からない。川卯という家で飯を焚かせて食う。夜が明けてから、大尉は走り廻って、切符の世話やら荷物の世話やらしてくれる。

汽車の窓からは、崖の上にぴっしり立て並べてある小家が見える。どの家も戸を開け放して、女や子供が殆ど裸でいる。中には丁度朝飯を食っている家もある。仲為のような為事をする労働者の家だと士官が話して聞せた。

田圃の中に出る。稲の植附はもう済んでいる。おりおり蓑を着て手籠を担いで畔道をあるいている農夫が見える。

段々小倉が近くなって来る。最初に見える人家は旭町の遊廓である。どの家にも二階の欄干に赤い布団が掛けてある。こんな日に干すのでもあるまい。毎日降るのだから、こうして曝すのであろう。

がらがらと音がして、汽車が紫川の鉄道橋を渡ると、間もなく小倉の停車場に着く。参謀長を始め、大勢の出迎人がある。一同にそこそこに挨拶をして、室町の達見という宿屋にはいった。

[2]青空文庫 https://www.aozora.gr.jp/

04.txt

金井湛君は哲学が職業である。

哲学者という概念には、何か書物を書いているということが伴う。金井君は哲学が職業である癖に、なんにも書物を書いていない。文科大学を卒業するときには、外道哲学と 前の希臘哲学との比較的研究とかいう題で、余程へんなものを書いたそうだ。それからというものは、なんにも書かない。

しかし職業であるから講義はする。講座は哲学史を受け持っていて、近世哲学史の講義をしている。学生の評判では、本を沢山書いている先生方の講義よりは、金井先生の講義の方が面白いということである。講義は直観的で、或物の上に強い光線を投げることがある。そういうときに、学生はいつまでも消えない印象を得るのである。殊に縁の遠い物、何の関係もないような物を藉りて来て或物を説明して、聴く人がはっと思って会得するというような事が多い。 は新聞の雑報のような世間話を材料帳に留めて置いて、自己の哲学の材料にしたそうだが、金井君は何をでも哲学史の材料にする。真面目な講義の中で、その頃青年の読んでいる小説なんぞを引いて説明するので、学生がびっくりすることがある。

小説は沢山読む。新聞や雑誌を見るときは、議論なんぞは見ないで、小説を読む。しかし若し何と思って読むかということを作者が知ったら、作者は憤慨するだろう。芸術品として見るのではない。金井君は芸術品には非常に高い要求をしているから、そこいら中にある小説はこの要求を充たすに足りない。金井君には、作者がどういう心理的状態で書いているかということが面白いのである。それだから金井君の為めには、作者が悲しいとか悲壮なとかいう積で書いているものが、極て滑稽に感ぜられたり、作者が滑稽の積で書いているものが、却て悲しかったりする。

05.txt

雑煮を食って、書斎に引き取ると、しばらくして三四人来た。いずれも若い男である。そのうちの一人がフロックを着ている。着なれないせいか、メルトンに対して妙に遠慮する傾きがある。あとのものは皆和服で、かつ不断着のままだからとんと正月らしくない。この連中がフロックを眺めて、やあ――やあと一ツずつ云った。みんな驚いた証拠である。自分も一番あとで、やあと云った。

フロックは白い手巾を出して、用もない顔を拭いた。そうして、しきりに屠蘇を飲んだ。ほかの連中も大いに膳のものを突ついている。ところへ虚子が車で来た。これは黒い羽織に黒い紋付を着て、極めて旧式にきまっている。あなたは黒紋付を持っていますが、やはり能をやるからその必要があるんでしょうと聞いたら、虚子が、ええそうですと答えた。そうして、一つ謡いませんかと云い出した。自分は謡ってもようござんすと応じた。

それから二人して東北と云うものを謡った。よほど以前に習っただけで、ほとんど復習と云う事をやらないから、ところどころはなはだ曖昧である。その上、我ながら覚束ない声が出た。ようやく謡ってしまうと、聞いていた若い連中が、申し合せたように自分をまずいと云い出した。中にもフロックは、あなたの声はひょろひょろしていると云った。この連中は元来謡のうの字も心得ないもの共である。だから虚子と自分の優劣はとても分らないだろうと思っていた。しかし、批評をされて見ると、素人でも理の当然なところだからやむをえない。馬鹿を云えという勇気も出なかった。

06.txt

硝子戸の中から外を見渡すと、霜除をした芭蕉だの、赤い実の結った梅もどきの枝だの、無遠慮に直立した電信柱だのがすぐ眼に着くが、その他にこれと云って数え立てるほどのものはほとんど視線に入って来ない。書斎にいる私の眼界は極めて単調でそうしてまた極めて狭いのである。

その上私は去年の暮から風邪を引いてほとんど表へ出ずに、毎日この硝子戸の中にばかり坐っているので、世間の様子はちっとも分らない。心持が悪いから読書もあまりしない。私はただ坐ったり寝たりしてその日その日を送っているだけである。

しかし私の頭は時々動く。気分も多少は変る。いくら狭い世界の中でも狭いなりに事件が起って来る。それから小さい私と広い世の中とを隔離しているこの硝子戸の中へ、時々人が入って来る。それがまた私にとっては思いがけない人で、私の思いがけない事を云ったり為たりする。私は興味に充ちた眼をもってそれらの人を迎えたり送ったりした事さえある。

私はそんなものを少し書きつづけて見ようかと思う。私はそうした種類の文字が、忙がしい人の眼に、どれほどつまらなく映るだろうかと懸念している。私は電車の中でポッケットから新聞を出して、大きな活字だけに眼を注いでいる購読者の前に、私の書くような閑散な文字を列べて紙面をうずめて見せるのを恥ずかしいものの一つに考える。これらの人々は火事や、泥棒や、人殺しや、すべてその日その日の出来事のうちで、自分が重大と思う事件か、もしくは自分の神経を相当に刺戟し得る辛辣な記事のほかには、新聞を手に取る必要を認めていないくらい、時間に余裕をもたないのだから。――彼らは停留所で電車を待ち合わせる間に、新聞を買って、電車に乗っている間に、昨日起った社会の変化を知って、そうして役所か会社へ行き着くと同時に、ポッケットに収めた新聞紙の事はまるで忘れてしまわなければならないほど忙がしいのだから。

藤原 義久 / yoshi.fujiwara@gmail.com

07.txt

ようやくの事でまた病院まで帰って来た。思い出すとここで暑い朝夕を送ったのももう三カ月の昔になる。その頃は二階の廂から六尺に余るほどの長い葭簀を日除に差し出して、熱りの強い縁側を幾分か暗くしてあった。その縁側に是公から貰った楓の盆栽と、時々人の見舞に持って来てくれる草花などを置いて、退屈も凌ぎ暑さも紛らしていた。向に見える高い宿屋の物干に真裸の男が二人出て、日盛を事ともせず、欄干の上を危なく渡ったり、または細長い横木の上にわざと仰向に寝たりして、ふざけまわる様子を見て自分もいつか一度はもう一遍あんな逞しい体格になって見たいと羨んだ事もあった。今はすべてが過去に化してしまった。再び眼の前に現れぬと云う不慥な点において、夢と同じくはかない過去である。

病院を出る時の余は医師の勧めに従って転地する覚悟はあった。けれども、転地先で再度の病に罹って、寝たまま東京へ戻って来ようとは思わなかった。東京へ戻ってもすぐ自分の家の門は潜らずに釣台に乗ったまま、また当時の病院に落ちつく運命になろうとはなおさら思いがけなかった。

08.txt

こんな夢を見た。

腕組をして枕元に坐っていると、仰向に寝た女が、静かな声でもう死にますと云う。女は長い髪を枕に敷いて、輪郭の柔らかな瓜実顔をその中に横たえている。真白な頬の底に温かい血の色がほどよく差して、唇の色は無論赤い。とうてい死にそうには見えない。しかし女は静かな声で、もう死にますと判然云った。自分も確にこれは死ぬなと思った。そこで、そうかね、もう死ぬのかね、と上から覗き込むようにして聞いて見た。死にますとも、と云いながら、女はぱっちりと眼を開けた。大きな潤のある眼で、長い睫に包まれた中は、ただ一面に真黒であった。その真黒な眸の奥に、自分の姿が鮮に浮かんでいる。

自分は透き徹るほど深く見えるこの黒眼の色沢を眺めて、これでも死ぬのかと思った。それで、ねんごろに枕の傍へ口を付けて、死ぬんじゃなかろうね、大丈夫だろうね、とまた聞き返した。すると女は黒い眼を眠そうにたまま、やっぱり静かな声で、でも、死ぬんですもの、仕方がないわと云った。

じゃ、私の顔が見えるかいと一心に聞くと、見えるかいって、そら、そこに、写ってるじゃありませんかと、にこりと笑って見せた。自分は黙って、顔を枕から離した。腕組をしながら、どうしても死ぬのかなと思った。

藤原 義久 / yoshi.fujiwara@gmail.com