5.1 ベイジアンフィルタとは

ベイジアンフィルタは、迷惑メールフィルタの仕組みとして広く知られている機械学習処理のアルゴリズムです。膨大な言葉の組合せで表現される自然言語の文書の分類に、その真価が発揮されます。そして、機械学習処理という言葉のとおり、学習量が多ければ多いほど性能が向上します。

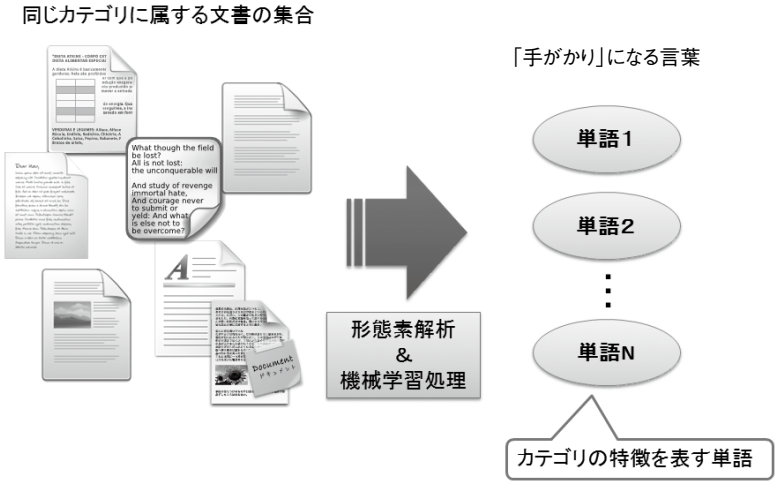
この節では、ベイジアンフィルタがいかなる仕組みで動いているのかを解説しますが、本書で取り上げるベイジアンフィルタは、「ナイーブベイズフィルタ」というベイジアンフィルタです。機械学習処理を実現するベイジアンフィルタの中で最もシンプルなロジックで、それは驚くほど簡単な仕組みになっています。

5.1.1 ベイジアンフィルタの基本的な考え方

【1】ベイジアンフィルタの概要

ベイジアンフィルタの基本的な考え方は、「手がかり」を洗い出すというものです。「手がかり」とは、分類の決め手となる言葉のことです。ベイジアンフィルタは計算処理を通して、自然言語の文章を構成する多数の言葉の中から「手がかり」となる言葉を選び抜きます(**図5.1.1**)。

図5.1.1　手がかりとなる言葉を見つけ出す

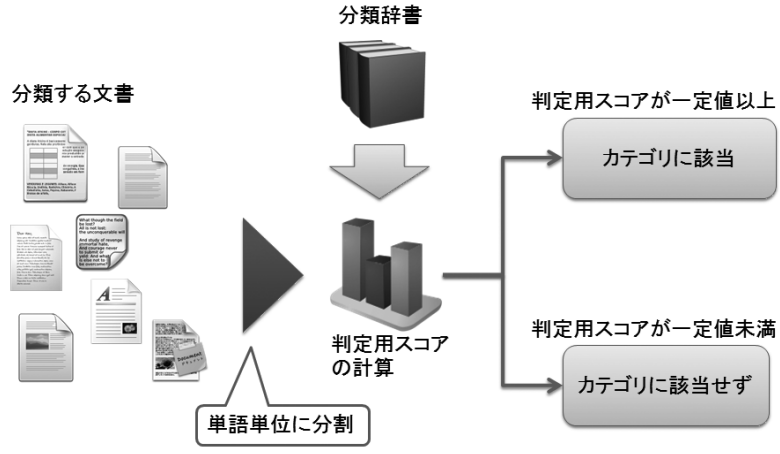


「手がかり」となる言葉を洗い出したら、それらを「分類辞書」に記録します。この「分類辞書」には、「手がかり」となった言葉と、その言葉の「スコア」（決め手としての強さを表す数字）が記録されます。では、ベイジアンフィルタが、どのようにして文書を分類するのか、その概要は次の〔1〕～〔5〕のとおりです（**図5.1.2**）。

〔1〕まず、ターゲットとなる文書を単語に分割します。
〔2〕そして、各単語を「分類辞書」と照らし合わせます。
〔3〕「手がかり」となる単語を見つけたら、「判定用スコア」を加算します。
〔4〕ターゲットとなる文書に含まれる単語全てをチェックした後、「判定用スコア」が一定値を超えていたら分類に該当した（迷惑メールに該当したなど）と判定します。
〔5〕「判定用スコア」を計算することができたら、後は基準値を決めて分類するだけです。

この「判定用スコア」を計算するまでが、ベイジアンフィルタの肝心な部分となります。

図5.1.2　分類辞書を使って文書を仕分ける



【2】ベイジアンフィルタの具体的プロセス

では、ベイジアンフィルタの各プロセスを詳しく見ていきましょう。まず、「手がかり」となる言葉の洗い出しです。どのようにすれば洗い出すことができるのでしょうか？ ベイジアンフィルタは、確率の違いを用いて「手がかり」となる言葉を見つけ出します。具体的には、「ベイジアンフィルタで分類したいカテゴリ」（迷惑メールなど）に属する文書に特有の言葉を、確率を使って見つけ出すのです。例えば、ある単語Aについて考えてみましょう。

- 「ベイジアンフィルタで分類したいカテゴリ」では、単語Aが10%の確率で出現したと仮定します。

- 「ベイジアンフィルタで分類したいカテゴリ」以外では、単語Aが0.5%の確率で出現したと仮定します。

この時、単語Aのスコアは**図5.1.3**のように求めることができます。

図5.1.3　単語Aのスコアの計算

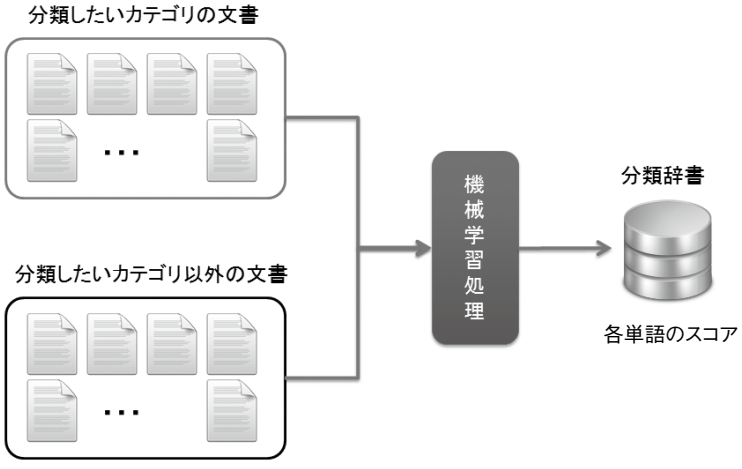
$$\text{単語Aのスコア} = \frac{10\%}{10\% + 0.5\%} = \frac{10\%}{10.5\%} = 0.952381 \fallingdotseq 0.952$$

この計算例では、単語Aのスコアは、約0.952となりました。では、このスコアはどのような意味を持つのでしょうか。無作為に取り出した文書の中に単語Aが含まれている場合、約0.952という非常に高い確率で「ベイジアンフィルタで分類したいカテゴリ」ということを意味します。

ここで例として挙げた単語Aは、スコアが高い方の単語です。実際には、スコアが高い単語、スコアが低い単語が混じり合っています。

単語のスコアを計算するには、たくさんの文書を必要とします。このたくさんの文書を「ベイジアンフィルタで分類したいカテゴリ」に属する文書と、それ以外の文書にグループを分けておきます。この分けられた文書が、「学習用データ（または教師用データ）」と呼ばれるものです。学習用データを機械学習処理にかけると、各単語のスコアが計算されていきます（**図5.1.4**）。

図5.1.4　学習用データと機械学習処理



このようにして、様々な単語のスコアを求めて、データベースに保存しておきます。そのデータベースが先ほど述べた「分類辞書」です。

ベイジアンフィルタが文書を分類する時、この分類辞書を参照しながら、文書を構成する各単

語のスコアを求めます。そして、文書全体のスコアである判定用スコアを計算するのです。

以上は、ベイジアンフィルタによる文書の分類の基本的な仕組みです。次項では、迷惑メールフィルタを題材にもうすこし具体的に解説しましょう。

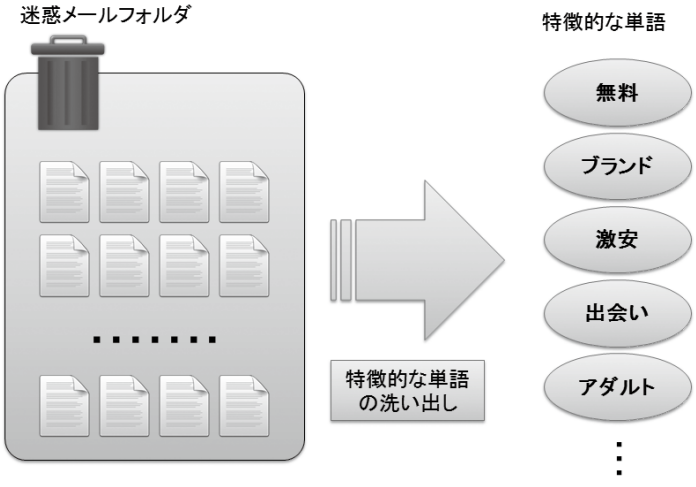
5.1.2 ベイジアンフィルタの実用例：迷惑メールフィルタ

前項では、ベイジアンフィルタの基本的な仕組みを解説しました。この項では、具体的なベイジアンフィルタの実用例として「迷惑メールフィルタ」を取り上げ、その基本的な仕組みについて解説します。

読者の皆さんは、普段から迷惑メールをよく目にしている人も多いのではないでしょうか。迷惑メール用に作成したメールフォルダを開くと、うんざりするような見出しのメールが大量に詰まっていることでしょう。このうちのいくつかを抜き出して文面を読んだとします。すると、迷惑メールでよく用いられる単語が浮かび上がってくると思います。

例えば、「無料」「ブランド」「激安」「出会い」「アダルト」といった単語を挙げることができます。これらの単語が、迷惑メールを分類する上での具体的な「手がかり」となるのです(**図5.1.5**)。

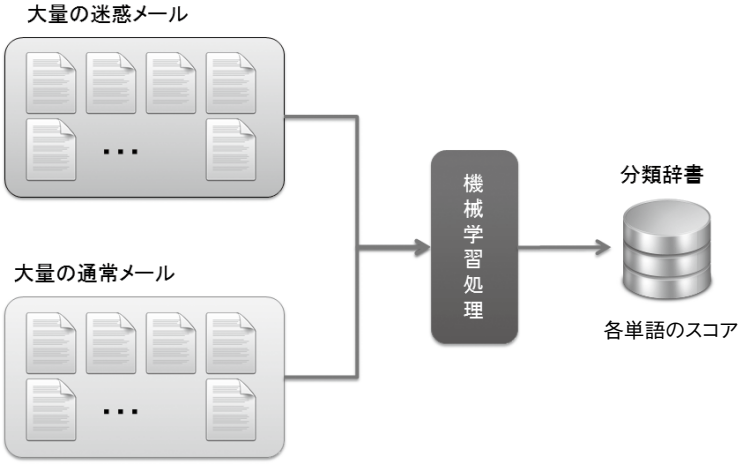
図5.1.5　迷惑メールの特徴となる単語



こうした、分類する上での「手がかり」になる単語を、高速に見つけ出すのが機械学習処理ですが、実際に迷惑メールフィルタを作成するためには、大量の学習用データが必要です。すなわち、「大量の迷惑メール」と「大量の通常メール」です。この２つの大量データに対して、機械学習処理では、次の〔1〕～〔5〕のプロセスで処理していきます(**図5.1.6**、**図5.1.7**)。

〔1〕まず「大量の迷惑メール」と「大量の通常メール」のそれぞれの単語の分布を計算します。
〔2〕続いて、各単語が迷惑メールの中で出現する確率と、通常メールの中で出現する確率を計算します。
〔3〕そして、各単語のスコアを求め、分類辞書に記録していきます。

図5.1.6　迷惑メールフィルタにおける機械学習処理



例えば、学習用データとして迷惑メールを2万件、通常メールを2万件、機械学習処理に投入したとします。そして、「出会い」という単語が、迷惑メールの中では4,000件のメールで見つかったとします。出現確率は、20%となります。一方、通常メールでは、20件のメールで見つかったとします。出現確率は、0.1%となります。これらから「出会い」という単語のスコアは、図5.1.7のように求めることができます。

図5.1.7　単語「出会い」のスコアの計算

$$\text{単語「出会い」のスコア} = \frac{20\%}{20\% + 0.1\%} = \frac{20\%}{20.1\%} = 0.995025 \cong 0.995$$

上記の例では、「出会い」という単語のスコアが、約0.995となりました。非常に高い確率で、迷惑メールであることを意味するスコアの高さです。

同様にして、「出会い」以外の単語についても、迷惑メールにおける出現確率と、普通のメールにおける出現確率を計算します。こうして、迷惑メールを分類するための分類辞書ができあがります。

〔4〕分類辞書ができあがったら、それを用いてメールの文面の判定用スコアを計算します。
　迷惑メールの場合、必然的に判定用スコアは高い値となります。判定用スコアが一定値以上のメールを、迷惑メールと判定することになります。
〔5〕後は、性能が最善になるように判定用スコアの閾値を調節します。

迷惑メールフィルタの基本的な仕組みは以上のとおりです。いかがでしたか？ 意外とシンプルでしょう？ このシンプルさと、なおかつ多くの場面で安定して性能を発揮するのがベイジアンフィルタの優れた点です。

次節では、各単語のスコアや、判定用スコアを計算する処理の実装方法について解説します。