集合知プログラミング
第6章　ドキュメントフィルタリング（1）

05T4007T　江口晃

はじめに

- ドキュメントフィルタリングとは
  - -ドキュメントを内容に応じて分類すること
- ドキュメントフィルタリングの利用
  - -スパムメールの削除
    - -ウェブサイトのスパムへの対策
    - -メッセージの内容による自動分類

スパムフィルタリング

- 初期のスパムフィルタリング
  - ルールを基に選り分ける
    例：大文字の過度の使用、薬に関連する単語、HTMLの色
- 初期のスパムフィルタリングの欠点
  - スパマーにすぐにルールを見破られる
  - 小文字に変換できない人からのメッセージがスパムに分類されてしまう
  - 特定のユーザや場所ではスパムであるキーワードでも、ほかの場面では全く問題のない場合がある

ドキュメントと単語

- 特徴の抽出
  - -ドキュメントを分類するために単語を特徴として用いる
  - -スパムに頻繁に出現し十分にありふれたもの
  - -すべてのドキュメントに出現するようなものではいけない

分類器のトレーニング

- 分類器
  - -異なるドキュメントを単語の特徴を利用して分類する
  - -トレーニングを重ねるにつれ分類の仕方を学習する
  - -正答例を読み取ることで学習する
  - -学習するにつれ正確性を増していく

確率を計算する

- 単語が出現する確率
  - それぞれのカテゴリに属するドキュメント数、それぞれの単語が存在する数を保持する
  - 特定のカテゴリに属するドキュメントの中にその単語が存在する数を、そのカテゴリに属するドキュメントの総数で割る
  - 与えられたカテゴリに特定の単語が出現する確率となる

    例："quick”という単語はgoodというカテゴリに分類された三つのドキュメントのうち、二つに出現する。

    Pr(quick|good) = 0.666(2/3)

推測を始める（1）

- 実例のみの確率の欠点
  -トレーニングの初期や、まれにしか出現しない単語のデータ不足

- 仮の確率
  -情報がほとんどない場合に利用する仮の確率を決める
  -仮の確率にどの程度の重みを持たせるか決める
  -実例を基に算出した確率と仮の確率の平均に重みをつけて返す

推測を始める（2）

- 仮の確率を使った計算式

  (仮確率の重み*仮確率+単語が存在する数*実例のみの確率)/(仮確率の重み+単語が存在する数)

- 仮の確率の具体例

  ”money”という単語はbadに分類された一つのドキュメントにしか出現せず、badに分類されているのはこのドキュメント一つである。

  -過去の情報のみの場合

  Pr(money|bad) = 1.0

  -仮の確率を用いた場合

  仮の確率を0.5、重みを1とする

  Pr(money|bad) = (1*0.5+1.0)/(1.0+1.0) = 0.75

単純ベイズ分類器

- 単純ベイズ分類器

  -ドキュメントに含まれている個々の単語の確率をまとめて、ドキュメントが与えられたカテゴリに属する確立を取得する

  -組み合わせた確率は互いに独立していると見なす（実際は独立性の仮定は不正確）

  -完璧とはいえない仮定だが、分類する方法としては効果的

ドキュメント全体の確率

- ドキュメント全体の確率

  -単純ベイズ分類器を使うには、ドキュメント全体の与えられたカテゴリでの確率を決める必要がある

  -ドキュメント全体の確率は、文章中の独立した単語の確率を掛け合わせることで計算できる

ベイズの定理の簡単な紹介

- ベイズの定理

  -ベイズの定理は以下のように表記される

  Pr(A|B)=Pr(B|A)*Pr(A)/Pr(B)

  -本章で当てはめると以下のようになる

  Pr(カテゴリ|ドキュメント)=Pr(ドキュメント|カテゴリ)*Pr(カテゴリ)/Pr(ドキュメント)

  - Pr(カテゴリ|ドキュメント)はドキュメントが特定のカテゴリにぞくする確率
  - Pr(ドキュメント|カテゴリ)はドキュメント全体の確率
  - Pr(カテゴリ)はランダムに選ばれたドキュメントが特定のカテゴリに属する確率
  - Pr(ドキュメント)はどのカテゴリについても同じなのでここでは無視する

カテゴリの選択（1）

- 単純なカテゴリ選択
  -ドキュメントがそれぞれのカテゴリに属する確率を求める
  -もっとも確率の高いカテゴリを選ぶ

- 単純なカテゴリ選択の欠点
  -多くのアプリケーションでは、カテゴリはすべて同等であるという風にはは考えられない
  -アプリケーションによっては、確率がわずかに高いカテゴリを選ばない方がよい場合がある
    - 例：スパムメールの分類

カテゴリの選択（2）

- しきい値を用いたカテゴリ選択

  -ドキュメントがそれぞれのカテゴリに属する確率を求める

  -それぞれのカテゴリに最低限のしきい値を設定する

  -ドキュメントがそのカテゴリに属する確率が、他のカテゴリに属する確率より、しきい値以上に高くなければいけない