言語処理のための機械学習入門

2.4 文書に対する前処理とデータスパースネス問題
2.5 単語のベクトル表現
2.6 文書や単語の確率分布による表現

河野和平

ストップワード

- 話題の種類と関連性を持たない単語
  - the,is,haveなどはどんな文書にでも出現する。
- 文書のクラスタリングなどの場合
  - ストップワードに関する情報は重要でない。
  - ストップワードを削除してベクトル化を行う。

ステミング

- 派生語などを同一の素性とみなす作業
  - run,ran,rannnerなどは同じような話題を指している。

- ポーターのステマー
  - 語尾のedを除去する (walked → walk)
  - 語尾のateを除去する (operate → oper)
  - 語尾のationalを除去する (operational → oper)

見出し語化

- ある単語の変形は同一の素性であるとみなす。
  - run,runs,ranはrunの変形であり、似た話題を指す。
- すべての単語を基本形に戻す。
- 周囲に出現する単語などの文脈を考慮する。

日本語の前処理

- 単語分割
  - 日本語は単語がスペースで区切られていない。
  - 品詞タグ付けと同時に行われる。(形態素解析)
- ステミング
  - 通常は用いない。
- 見出し語化
  - “走ら”ない，“走り”たい　→“走る”
- 品詞のタグ付け
  - “騒ぎ”を動詞と名詞で区別する。

データスパースネス問題

- ある国語辞典の収録単語を要素としてベクトルを作成する。
  - 収録単語数：50,000語
  - 新聞記事の出現単語:100語

  ⇒ 少なくとも49,900個の要素は0　(疎である)
- 記事をたくさん用意しても、ある単語がどれくらい出現しやすいかわかりずらい。

  (データを処理するための統計値が十分でない)

単語のベクトル表現

- 文書や文と違い、内部に別の単語を含んでいない。

- 目的の処理で単語が含む文字が重要な場合
  - 単語が含む文字を用いてベクトルを表現する。

- 文脈ベクトル表現
  - 単語wの直前直後に出現する単語群を用いる。

単語トークンの文脈ベクトル表現

「高く　跳ぶ　に　は　まず　屈め」

- “跳ぶ”のベクトル

$$x_{\text{跳ぶ}} = \left(n(\text{"高く"}), n(\text{"に"}), n(\text{"は"}), n(\text{"まず"}), n(\text{"屈め"})\right)$$
$$= (1,1,0,0,0)$$

- 文脈窓
  - ベクトルに考慮する対象単語の前後数トークン
  - その大きさを文脈窓幅という。

単語トークンの文脈ベクトル表現

- 位置による区別

  「危険　を　恐れ　ず　攻め　よ」

$$x_{\text{恐れ}} = \left(n(\text{"危険"}_{-2}), n(\text{"危険"}_{-1}), n(\text{"危険"}_{+1}), n(\text{"危険"}_{+2}), \cdots\right)$$
$$= (1,0,0,0,\cdots)$$

- 構文的な情報を用いたベクトル表現

  動詞 w をベクトル表現

  w の主語や目的語として出現する単語を要素とする。

  より詳細に構文的な振る舞いを表現可能。

単語タイプの文脈ベクトル表現

- 複数の文脈窓内で単語が何回出現したかを素性とする。

「Nothing ventured, notihng gained.」

$$
\begin{aligned}
& x_{\text{nothing}} \\
&= (n("ventured"_{+1}), n("ventured"_{-1}), \\
&\quad n("gained"_{+1}), n("gained"_{-1}), n(","_{+1}), n(","_{-1})) \\
&= (1,0,1,0,0,1)
\end{aligned}
$$

確率分布による表現

- 各単語がどの程度出現しやすいか $P(W|d)$

W: 各単語を値とする確率変数　d: 文書

- 単語タイプ w の確率分布
  - w の周囲の単語タイプの出現確率
  - ある単語 w が出現したという条件のもとで、周囲に他の単語 v が出現する確率の確率分布 $P(V|w)$

確率分布による表現

「Nothing ventured, notihng gained.」

nothingの確率表現

$$P(V = "ventured"_{+1}|w) = 0.33$$
$$P(V = ","_{-1}|w) = 0.33$$
$$P(V = "gained"_{+1}|w) = 0.33$$