ISS-P-032

ロジスティック回帰によるデータ分類

川見　純一†　　三浦　孝夫†
†法政大学理工学部創生科学科

1. はじめに

近年、人口知能や統計学の分野において、ある事例がどのクラスに属するかを決定する分類問題は活発に研究されている。そして自然言語処理の分野では最大エントロピー法を利用した研究が散見される。しかし、最大エントロピー法の優位性にも関わらず、この手法を用いた研究者はごく一部である。

本稿では、最大エントロピー法を用いたデータ分類を行い、最大エントロピー法の性能向上を GIS によって行い、単純ベイズと性能を比較し評価する。

2. 提案手法

本研究では、最大エントロピー法を用いたデータ分類を行い、GIS による性能の向上する手法を提案する。

3. 実験手順

本研究では"CD－毎日新聞　2017"の記事 3 か月分を利用する。まず、2017 年の記事の中から安定して出現し、出現数の多いクラスと利用する 3 か月の月を選択する。

選択した各記事から名詞、名詞句を取り出し、単語リストとして利用する。その際、素性として利用する単語を選択する際、Zipf の法則を用いる。Zipf の法則は第 1 法則と第 2 法則があり、以下のように定義される。

第 1 法則

単語の頻度 f と頻度順位 r との積が定数 C になる。

$f \times r = C$

第 2 法則

出現頻度 f の単語数 Ff と頻度1の単語の数 F1 との間に次の関係が成り立つ。

$F1/Ff = f(f+1)/2$

両方の式が同時に成り立つような単語の頻度 f=fk を求めることで、索引語として望ましい中程度の頻度を得る。これを素性とし、GIS を用いてパラメータ推定を行う。

GIS アルゴリズムは以下の反復スケーリング法で行える。

第 1 ステップ

素性に初期値1の重み（パラメータ）を与え、補完定数 C の決定を行う。C はクラスごとによって出現する見出し語数が異なるため、単語を一定にするために用いる。

第 2 ステップ

素性の重み αi を経験期待値とモデル期待値の差分が収束するまで繰り返す。

この GIS を用いて、パラメータ推定を行い、クラスごとの確率分布を作成する。このときの性能向上を単純ベイズと比べ評価を行う。

参考文献

[1]高村大也：自然言語処理のための機械学習入門、コロナ社、2010

[2]新納浩幸（2020）「最大エントロピー法と自然言語処理」