1. はじめに

情報技術の革新と発達により、そこで発生するコミュニケーションもまた様々に発達してきた。特にインターネット上のコミュニケーションの発達は特に目覚ましい。
様々なソーシャルメディアでネット上でのオンラインでのコミュニケーションは今までにない以上に発展しており、実生活上でもその比率は年々に上昇してきている。LINE, Twitter, Instagram, Facebook, Slack などオンラインでの意思疎通を円滑に行うチャットツールやソーシャルメディアは具体例を挙げれば枚挙に暇がない。
さてそういったツールやり取りの中でも欠かせないものの一つとして「顔文字」がある。通常のメッセージとは別にその時々に適した感情を表すものとして利用される。「(。・・。) ｵｯｹｰ♪」、「m(_ _　)m」など表したい気持ちや思いをカジュアルに伝える手段として特に若い人の間で流行ってきた。

今回はその「顔文字」をどのように分類するか、という点を課題として設定した。この問題に対して機械学習の手法を使って予測を行い、どの程度の精度が達成できるか実験を行った。

なぜ今顔文字の分類を行うのか？

分類された感情（「嬉しい」、「悲しい」など）を手掛かりに適切な顔文字（「(≧∇≦)b」，「:-()」）がサジェストできることが達成したい目標である。このような顔文字の補完を機械的に行う必要があり、この点こそまさに顔文字の分類の意義になる。なぜ顔文字が入力する機会が増えているのか、なぜその補完が必要なのか、なぜ分類を機械的に行わなければならないのか、の 3 つの観点に沿ってそれぞれの疑問に対する答えを本項で説明する。

はじめに、の冒頭で述べているようにインターネットや情報技術の発展に伴ってコミュニケーションのあり方の変化がしている。そのためチャットツールでのやり取り中心に顔文字が入力する機会は増えている。
インターネットでのオンライン上のやり取りが増える以前では親しい友人や家族とのやり取りは主に対面の話し言葉で行われていた。この時期の特徴としては距離の制約・口頭でのコミュニケーションの 2 つが特に特徴だった。文章についても本を中心とした推敲を重ねた文章が主であった。それが電話の出現とともに距離の制約がなくなり、ネットの到来とともに口頭からチャットでのコミュニケーションを主軸としたものに変化してきた。
これは今までに経験していなかった変化である。このような変化の中で、顔文字や絵文字などを利用する機会が増えた。

顔文字を入力する機会が増えていると同時に、それを入力する端末も合わせて変化しなければならない。なぜならこのような顔文字は、様々な特殊文字が入り組んでいるため覚えにくく、特殊文字を組み合わせて構成した表意文字のため打ちにくい。
そのような性質から顔文字を入力する際に入力デバイスに何らかの補完が必要になる。一般的に利用されているスマートフォンの iPhone や Android でも標準でこのような顔文字に対する補完が備わっている。
それでは既にカテゴライズしているにもかかわらず、分類器を今新しく作る必要はあるのだろうか。

標準で備わっている補完機能については 2 つの欠点が存在する。一つは数の制約であり、もう一つは柔軟性に対する制約である。新しく顔文字の分類器そこで提供される顔文字の数で十分であれば問題ないが、これら IME で標準的に備わっている顔文字の数は限られている。この点が数に対する制約である。また、顔文字の大きな特徴として、新しい顔文字の流行や発見が頻繁に発生している点にある。新しい顔文字や流行りで最新の顔文字は標準的に提供されている IME には記載されていない。この点が柔軟性に対する欠点である。
このような欠点が存在するため、既存の IME に標準的に備わっている顔文字の分類では十分でないことは明らかであり、分類器を新しく作る必要性はこの点から考えても明らかであろう。

最後に機械的に分類を行う意義について説明する。
機械的に分類を行わなければならないのは人手でのコストがかかるからである。当然のことように思われるが、この人手でのコストという点を少し深く考えたい。人手でのコストがかかるのは大量の顔文字が存在し、新しい顔文字も日々増えているからである。この点に限って考えれば機械で分類を行うことのメリットは大きい。なぜなら、大量の顔文字を短時間で処理できる、新しい顔文字の追加などのメンテナンスが用意、のためである。
大量の顔文字が存在する理由としては、インターネットの発達だけではなく、ユニコードで登録されている文字が増えたため今まででは表現できなかった多用な顔文字が出現していることも理由の一つである。また、表意文字であるため少しの表現を変えることは比較的容易であり、日々新しい顔文字が簡単に作り出されている。
このように、仮にある時点の顔文字を人手で全て分類したとしてもその後新しい未知の顔文字は出現し対処しなければならず現時点でも大量の顔文字が存在する。この点が人手で分類することに対する大きなコストになっており、機械で行うことの大きなメリットになっている。

顔文字において用いた手法と期待する結果

解いた手法

A. 主ポイント
多変数のロジスティック回帰を用いて解いた

B. (A の手法に対して）なぜ?

期待していること・結論の想定

A. 主ポイント
顔文字の感情が正しく分類できること

手法

A. 主ポイント
手法としてはロジスティック回帰を利用した

B. （Aの手法に対してなぜ？）

1. 識別モデルであるため出現確率も含めて出力したかった ⇒ 識別関数の手法も検討したが利用しなかった
2. 多変数クラス分類で最も標準的に使われている手法がロジスティック回帰であるため
3. 実装が比較的シンプルに行える

期待する結果

A. 主ポイント
期待する結果としては顔文字が一定以上の程度できちんと分類できる

まとめ

2. 問題設定とモデルの解説

特徴量の構成方法

特徴量は以下で計算する

$$tf\_idf(t_i, d_j) = tf(t_i, d_j) * idf(t_i) \tag{1}$$

$$tf(t_i, d_j) = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{2}$$

$$idf(t_i) = \frac{|D|}{\sum_{j=1}^{|D|} n^0_{k,j}} \tag{3}$$

$n_{k,j}$ は単語 t_k が文章 d_j で出現する数を表している

ロジスティック分布の導出

確率モデルで分類問題を表す

以下では一般的な形でモデリング化するため、入力の変数を $\mathbf{x} \in \mathbf{R^D}$ として出力のラベルを $C \in \{C_1, C_2, \ldots, C_K\}$ とする。今回は確率モデルとして定式化するため、ある入力が与えられたときの特定のラベルが出現する確率を考えればよい。

これは条件付き確率 $P(C|\mathbf{x})$ として表せる。したがって、ある入力値 $\mathbf{x}$ が与えられときの最適なラベルは、この条件付き確率を最大化させるようなラベルである。
この最適なラベルを $C^\star$ とすれば以下のようにして

$$C^\star = \underset{k}{\operatorname{argmax}}\, P(C_k|\mathbf{x}) \tag{4}$$

未知の入力値に対するラベルの予測を与えることができる。

以後は一般化も踏まえて、入力ベクトルを固有の特徴を抽出する変換 $\boldsymbol{\phi}(\mathbf{x}) \in \mathbf{R^K}$ を加えたものを前提に

考える。

最大エントロピー原理

さて、上記の議論はあくまで何らかのパラメーター ω を用いて上記の条件付き確率を制限して $P(C|\mathbf{x};\omega)$ のように表せないと、議論がこれ以上先に進めない。
そのため最大エントロピー原理を用いて、トレーニングデータが与えられときに、尤もらしい確率分布がどのように表せるかを考えたい。

以後は計算を簡単にするために、出力のラベルはすべて 1 of K 符号化で表されているとしよう。1 of K 符号化でラベルがエンコードされる場合、正解データが C_k とすると L 次元ベクトル $\mathbf{t} \in \mathbf{R^L}$ として表せる。このとき $t_i \in \{0,1\}$ かつ $t_i = \delta_{i,k}\,(i = 1,2,\ldots,L)$ が成り立っている。
デルタ関数の定義は以下の通りである。

$$\delta_{i,k} = \begin{cases} 1 & (i = k) \\ 0 & (i \neq k) \end{cases}$$

定義がややこしそうだが、結局ラベルが L 個あったら L 次元ベクトルとして表し、C_k が正解データであれば k 番目の要素を 1 としてそれ以外を 0 とするベクトルである。

ここで、トレーニングデータとそのラベルをそれぞれ

$$(\boldsymbol{\phi}(\mathbf{x^{(1)}}), \mathbf{t^{(1)}}), (\boldsymbol{\phi}(\mathbf{x^{(2)}}), \mathbf{t^{(2)}}), \ldots\ (\boldsymbol{\phi}(\mathbf{x^{(N)}}), \mathbf{t^{(N)}}) \tag{5}$$

が与えられたとする。
このとき

$$\sum_{k=1}^{K} P(C_k|\mathbf{x^{(n)}}) = 1 \qquad (n = 1,2,\ldots,N) \tag{6}$$

$$\sum_{n=1}^{N} P(C_k|\mathbf{x^{(n)}})\,\boldsymbol{\phi}(\mathbf{x^{(n)}}) = \sum_{n=1}^{N} t_k^{(n)} \boldsymbol{\phi}(\mathbf{x^{(n)}}) \qquad (k = 1,2,\ldots,K) \tag{7}$$

が満たさなければならないと仮定しよう。

(6) は確率の定義より明らかに満たさなければならない。
(7) についてはいわゆる $P(C_k|\mathbf{x^{(n)}})$ が十分 $t_k^{(n)}$ をよく表さなければならない、という制約である。条件付きエントロピーは $-\sum_{k=1}^{L} P(C_k|\mathbf{x^{(n)}}) \ln P(C_k|\mathbf{x^{(n)}})$ より、これを (6), (7) の制約の元で最大化すればよい。
$P_k^{(n)} = P(C_k|\mathbf{x^{(n)}})$ のように簡易的に表すことにすれば、ラグランジュの未定乗数法より

$$H(p) = \sum_{n=1}^{N}\sum_{k=1}^{K} -P_k^{(n)} \ln P_k^{(n)} \tag{8}$$
$$+ \sum_{n=1}^{N} \lambda^{(n)} \{ \sum_{k=1}^{K} P_k^{(n)} - 1 \}$$
$$+ \sum_{k=1}^{K} {\omega_{\mathbf{k}}}^t \{ \sum_{n=1}^{N} \phi(\mathbf{x^{(n)}})(P_k^{(n)} - t_k^{(n)}) \}$$

を最大にするような $P_k^{(n)}$ を求めればよいことがわかる。
ここでスラッグ変数 $\lambda^{(n)}$ と ${\omega_{\mathbf{k}}}^t$ を導入した。

$P_k^{(n)}$ を求める

式の定式化までは行えたのであとは $H(p)$ を単純に $P_d^{(m)}$ で微分すればよい。

$$\begin{aligned}\frac{\partial H(p)}{\partial P_d^{(m)}} &= \sum_{n,k} \left\{ -\frac{\partial P_k^{(n)}}{\partial P_d^{(m)}} \{\ln P_d^{(m)} + 1\} + \lambda^n \{ \frac{\partial P_k^{(n)}}{\partial P_d^{(m)}} \} + \omega_{\mathbf{k}}^{\mathbf{t}} \phi(\mathbf{x^{(n)}}) \frac{\partial P_k^{(n)}}{\partial P_d^{(m)}} \right\} \\ &= -lnP_d^{(m)} - 1 + \lambda^{(m)} + \omega_{\mathbf{d}}^{\mathbf{t}} \phi(\mathbf{x^{(m)}})\end{aligned}$$

のように求まるので、以下のように微分をゼロをおけば

$$\frac{\partial H(p)}{\partial P_d^{(m)}} = 0 \tag{9}$$
$$P_d^{(m)} = \exp\left\{ \lambda^{(m)} - 1 + \omega_{\mathbf{d}}^{\mathbf{t}} \phi(\mathbf{x^{(m)}}) \right\} \tag{10}$$

(10) を (6) に代入すれば

$$\exp\left\{ \lambda^{(m)} - 1 \right\} = \exp\left(\omega_{\mathbf{d}}^{\mathbf{t}} \phi(\mathbf{x^{(m)}}) \right) \tag{11}$$

より (11) を (10) に代入して添字を整理すれば、

$$P_k^{(n)} = P(C_k | \mathbf{x^{(n)}}) = \frac{\exp\left(\omega_{\mathbf{k}}^{\mathbf{t}} \phi(\mathbf{x^{(m)}}) \right)}{\sum_{d=1}^{K} \exp\left(\omega_{\mathbf{d}}^{\mathbf{t}} \phi(\mathbf{x^{(m)}}) \right)} \tag{12}$$

と表せる。
このようにして目的であった条件付き確率分布がパラメーター $\omega_{\mathbf{d}}^{\mathbf{t}}$ を用いて表せるところまで求めることができた

(12) は多変数のロジスティック分布である

多変数ロジスティック分布の最尤推定

条件付き確率分布が得られたので (5) のトレーニングデータが与えられたときに負の対数尤度は以下のように表せる。

$$H(\mathbf{W}) = -\ln\left\{\prod_{n=1}^{N}\prod_{k=1}^{K} P(C_k|\mathbf{x^{(n)}})^{t_k^{(n)}}\right\} \tag{13}$$

$$= -\sum_{k,n} t_k^{(n)} \ln P_k^{(n)} \tag{14}$$

この対数尤度を最小化するような $\mathbf{W} = \omega_{\mathbf{k}}^{\mathbf{t}}$ $(k = 1, 2, \ldots K)$ を最急勾配法によって表せればよい。

ただしここで

$$P_k^{(n)} = \frac{\exp\left\{a_k^{(n)}\right\}}{\sum_{d=1}^{K}\exp\left\{a_d^{(n)}\right\}} \tag{15}$$

また

$$\begin{aligned} a_k^{(n)} &= a_k(\mathbf{x^{(n)}}) \\ &= \omega_{\mathbf{k}}^{\mathbf{t}}\phi(\mathbf{x^{(m)}}) \\ &= \sum_{d=1}^{D}\omega_{k,d}\phi_d(\mathbf{x^{(n)}}) \\ &= \sum_{d=1}^{D}\omega_{k,d}\phi_{d,n} \quad (\phi_d(\mathbf{x^{(n)}}) = \phi_{d,n}\text{とした}) \end{aligned}$$

とする。
(12) の関係式を変数の依存関係で分割しただけである。

このとき P_k に対して a_j の微分を考えると

$$\begin{aligned} \frac{\partial P_k}{\partial a_j} &= \frac{\partial}{\partial a_j}\left\{\frac{\exp\{a_k\}}{\sum_{d=1}^{K}\exp\{a_d\}}\right\} \\ &= \left\{\frac{\partial}{\partial a_j}(\exp\{a_k\})\right\}\frac{1}{\sum_{d=1}^{K}\exp\{a_d\}} + \\ &\exp\{a_k\}(-\frac{1}{(\sum_{d=1}^{K}\exp\{a_d\})^2})\frac{\partial}{\partial a_j}\left\{\sum_{d=1}^{K}\exp\{a_d\}\right\} \\ &= \frac{\exp\{a_k\}}{(\sum_{d=1}^{K}\exp\{a_d\})}(\delta_{jk} - \frac{\exp\{a_j\}}{(\sum_{d=1}^{K}\exp\{a_d\})}) \\ &= P_k(\delta_{k,j} - P_j) \end{aligned}$$

より

$$\frac{\partial P_k}{\partial a_j} = P_k(\delta_{k,j} - P_j) \tag{16}$$

が成り立つため (14) を $\omega_{m,j}$ に対して微分すると

$$\begin{aligned}
\frac{\partial H(\mathbf{W})}{\partial \omega_{m,j}} &= \sum_{n=1}^{N}\sum_{k=1}^{K}\sum_{l=1}^{K} t_k^{(n)} \left\{ \frac{\partial}{\partial a_l^{(n)}} ln P_k^{(n)} \right\} \frac{\partial a_l^{(n)}}{\partial \omega_{m,j}} \\
&= -\sum_{n,k,l} t_k^{(n)} P_k(\delta_{k,l} - P_j) \frac{\partial}{\partial \omega_{m,j}} \left\{ \sum_{d=1}^{D} \omega_{l,d}\phi_{d,n} \right\} \\
&= -\sum_{n,k,l} t_k^{(n)} P_k(\delta_{k,l} - P_j)\delta_{m,l}\phi_{j,n} \\
&= \sum_{n=1}^{N} \left\{ P_m^{(n)} - t_m^{(n)} \right\} \phi_{j,n}
\end{aligned}$$

添字を差し替えて

$$\frac{\partial H(\mathbf{W})}{\partial \omega_{k,d}} = \sum_{n=1}^{N} \left\{ P_k^{(n)} - t_k^{(n)} \right\} \phi_{d,n} \tag{17}$$

と与えられることがわかる。

最急降下法の規則まとめ

微分が得られたので最急降下法で停留解を得ることが可能になる。
これまでの議論をまとめると、規則は以下で与えられる

$$\begin{aligned}
\omega_{k,d}^{(new)} &= \omega_{k,d} - \eta \frac{\partial H}{\partial \omega_{k,d}} \\
&= \omega_{k,d} - \eta \sum_{n=1}^{N} \left\{ P_k^{(n)} - t_k^{(n)} \right\} \phi_{d,n} \\
P_k^{(n)} &= \frac{\exp\left\{ a_k^{(n)} \right\}}{\sum_{d=1}^{K} \exp\left\{ a_d^{(n)} \right\}} \\
a_k^{(n)} &= \sum_{d=1}^{D} \omega_{k,d}\phi_{d,n} \\
\phi_{d,n} &= \phi_d(\mathbf{x^{(n)}})
\end{aligned}$$

またこのようにして求められた最適解 $\omega_{d,k}^{\star}$ にを用いて (4) ラベルの予測を行えることができる

おまけ（二次の微小量）

ちなみに最急降下法ではなく二次の微小量を用いてニュートンラフソン法を使うことも可能である

(17) をさらに $\omega_{s,t}$ で微分すると

$$
\begin{aligned}
\frac{\partial^2 H}{\partial \omega_{s,t} \partial \omega_{k,d}} &= \sum_{n=1}^{N} \frac{\partial P_k^{(n)}}{\partial \omega_{s,t}} \phi_{d,n} \\
&= \sum_{n=1}^{N} \sum_{l=1}^{K} \frac{\partial P_k^{(n)}}{\partial a_l^{(n)}} \frac{\partial a_l^{(n)}}{\partial \omega_{s,t}} \phi_{d,n} \\
&= \sum_{n=1}^{N} \sum_{l=1}^{K} P_k^{(n)} \left\{ \delta_{k,l} - P_l^{(n)} \right\} \delta_{s,l} \phi_{t,n} \phi_{d,n} \\
&= \sum_{n=1}^{N} P_k^{(n)} \left\{ \delta_{k,s} - P_s^{(n)} \right\} \phi_{t,n} \phi_{d,n}
\end{aligned}
$$

と得られるため