授業実践

機械学習についての知識・技能を活用する授業実践

埼玉県立川越南高等学校教諭　春日井　優

1．はじめに

2016年度および2017年度の２年間に渡り，国立教育政策研究所教育課程指定校事業の共通教科「情報」として指定を受けて実践研究に取り組みました。研究テーマとして「知識・技能を活用する問題解決型協働学習の指導法及び評価に関する研究」と設定しました。生徒が問題発見を行い，情報科で学習する知識・技能を活用して問題解決を行う協働的な学習活動による授業実践を行いました。2016年度は「モデル化とシミュレーション」，2017年度は「機械学習」の知識・技能を習得し，それらの知識・技能を活用して問題解決に取り組む授業を行いました。

このような授業の展開により実践研究を行った理由として，次期学習指導要領改定に向けて知識・技能を活用することも重要視されるよう指導法を見直すことが示され[1]，情報科の授業の中で得た知識・技能を活用する授業を実践する必要があると考えていたためです。

また，新設される「情報Ⅱ」で「データサイエンス」をはじめ新たな内容が加わり，そのような内容に対しても，知識・技能を活用していくことが必要だと考えたことによるものです。

２年間の実践のうち，本稿では2017年度に実践した機械学習に関連する内容についてまとめます。本稿で紹介する実践内容は，新学習指導要領を予想しながら実践したことを踏まえてお読みいただくようお願いします。

2．授業の概要

実施科目　情報の科学
対象学年　３学年
単　　元　モデル化とシミュレーション
　　（Bag of Wordsモデルとして）
実施時数　14時間（50分授業）

表１　本実践での授業時間数

１時間	形態素解析
２～３時間	tf-idfによる特徴語抽出 Word Cloudによる特徴語可視化
４～６時間	ベイズの定理 単純ベイズ分類器
７～12時間	グループでの問題解決
13時間	グループの成果の発表
14時間	探究に向けた問題の発見 学習全体の振り返り

3．授業の内容について

(1) 形態素解析

この授業の１時間目には，形態素解析についての授業を行いました。機械学習の題材として，日本語の文章を対象に授業を行いたいと考えており，日本語を扱うには形態素解析をして文章を細かく分ける必要があるためです。形態素とは，単語よりも細かい単位のことで，例えば「可能性」という単語は，さらに「可能」+「性」と単語よりも細かく分けることができます。

形態素解析をして結果を示すWebサイトがあるので，生徒に検索させて，Web上で形態素解

析させました。また，プログラムを用いて形態素解析できるように，Pythonで書いたプログラムを配布しました。このプログラムでは，形態素解析したい文章を変数として与えます。生徒には変数にさまざまな文章を代入させ，そのプログラムを実行させて形態素解析した結果を得させました。

(2) tf-idfによる特徴語抽出とWord Cloudによる特徴語可視化

2～3時間目には，tf-idfによる特徴語の抽出，Word Cloudによる特徴語の可視化を行いました。tf-idfとは，文章内での形態素ごとの出現頻度であるtfと，複数の文章の中でどの程度希少性があるかを表すidfにより，形態素の重要性を数値化しようというものです。

表2　tfの計算に用いた表

出現語	出現回数	tf
果物	1	1／5
ケーキ	2	2／5
ビタミン	1	1／5
赤い	1	1／5

出現語［果物｜ケーキ｜ビタミン｜ケーキ｜赤い］

果物を題材にいくつかの形態素を与え，数を数えて表2のような表を完成させながら値を求める指導を行いました。idf，tf-idfについても同様に手作業を交えて指導しました。なお，tf-idf，tf，idfは，次の式で定義しました。

$$tfidf_{i,j} = tf_{i,j} \cdot idf_i$$

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

$$idf_i = \log \frac{|D|}{|\{d : d \ni t_i\}|}$$

ここで，$n_{i,j}$は文書d_jにおける単語t_iの出現回数，$\sum_k n_{k,j}$は文書d_jにおけるすべての単語の出現回数の和，$|D|$は総文書数，$|\{d : d \ni t_i\}|$は単語t_iを含む文の数を表しています。

その後，Pythonで書いたプログラムを配布しました。生徒は，そのプログラムに与える文章をテキストファイルとして保存し，実行することにより，tf-idfを計算した結果がExcelに出力される流れを経験しました。さらに，Excelに出力した結果をグラフにし，重要な形態素を見つける経験をしました（図1）。

また，同じプログラムでWord Cloudも出力するようにしておきました。それにより，図2のような画像が出力され，生徒は出力されたWord Cloudの画像を確認しました。

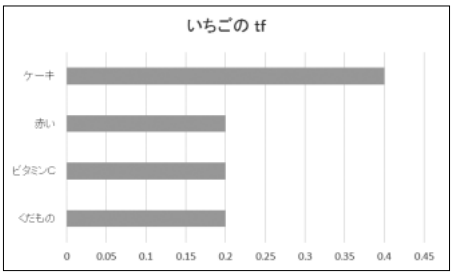

図1　計算結果をもとに作成したグラフ

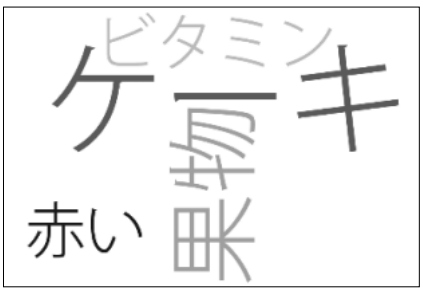

図2　果物を題材に作成したWord Cloud

(3) ベイズの定理

4～6時間目では，ベイズの定理とそれをもとに考えられた単純ベイズ分類器についての授業を行いました。

ベイズの定理とは，条件付き確率をもとに導出される定理で，

$$\mathrm{P}(cat|view) = \frac{\mathrm{P}(cat) \times \mathrm{P}(view|cat)}{\mathrm{P}(view)}$$

と数式では表現されます。確率を学んでから時間が経っているため，説明を簡略化するために，長方形の面積を用いて考え方を説明しました。

ベイズの定理の考え方が理解できるよう，具体例として迷惑メールの分類を題材に，次のような問題を例示して説明しました。

「受け取るメールのうち0.2の確率で迷惑メール

が届く。迷惑メールを分析すると，“振り込め”という語が迷惑メールには0.8の確率で含まれている。迷惑メール以外のメールには“振り込め”という語が0.05の確率で含まれている。今届いたメールには“振り込め”という語が含まれていることがわかった。このメールが迷惑メールである確率を求めよ。」

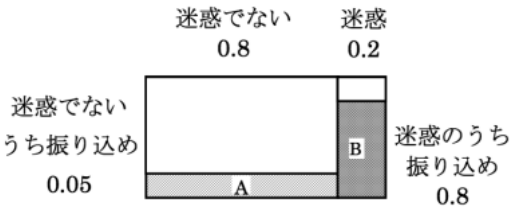

図３　迷惑メールの確率を求めるための図

説明に用いた図は図３です。分子は事前に迷惑だと推測できる確率P(cat)＝0.8と迷惑メールの場合に“振り込め”が含まれる確率P(view|cat)＝0.2の積（図中のBの領域の面積），分母は“振り込め”という語が含まれている確率でP(view)＝0.2（図中のAとBの領域の面積の和）です。これにより，“振り込め”という語を含むメールのうち，迷惑メールの確率(B)は，0.8(＝B/(A＋B))と求められることを説明しました。同様に，“振り込め”という語を含み迷惑メールでない確率(A)は，0.2(＝A/(A＋B))と求められることも説明しました。

この２つの確率の分母は共通で，確率の大小は分子の大小に依存することも確かめました。

(4) 単純ベイズ分類器

次に機械学習のうちの１つの手法である「単純ベイズ分類器」の仕組みを説明しました。先ほどの例で分子の大小を確かめたのは，迷惑メールである確率が大きいかどうかは，分子を比較すれば推測できることを確認するためです。このことを数式で表現すると，

$$P(\text{cat}|\text{view}) \propto P(\text{cat}) \times P(\text{view}|\text{cat})$$

となり，P(cat|view)の大きさは右辺の分子に比例し，この値の大小を用いて分類する手法です。

実際に文章を分類する際には，P(cat)はある分類catについての事前予測の確率で，計算上では文章の行数に比例するものとしました。また，P(cat|view)は，分類cat中に分類したい文章の形態素がどの程度の確率で観測できたかを表すものです。

分類の仕組みを理解するために，簡単な文章を学習データとして与えて，分類したい文章が分類catに含まれる確率P(cat|view)を，表を用いて計算しました。

具体的には，次の学習に用いる文章をもとに分類したい文章がどの果物に分類されるかを計算し説明しました。計算は表３を完成させながら説明しました。

学習データとして与えた文章（括弧内は分類cat）

文章１（いちご）：[果物｜ケーキ｜ビタミン]

文章２（いちご）：[ケーキ｜赤い]

文章３（りんご）：[果物｜ジュース｜ケーキ｜ビタミン｜青森]

文章４（キウィ）：[ビタミン｜毛｜緑｜黄色]

分類したい文章　：[果物｜ケーキ]

表３　単純ベイズ分類器の計算に用いた表

	いちご	りんご	キウィ
P(cat)	2/4	1/4	1/4
単語数	5	5	4
P(果物|cat)	1/5	1/5	0
P(ケーキ|cat)	2/5	1/5	0
P(果物 ケーキ|cat)	2/25	1/25	0
P(cat)×P(果物 ケーキ|cat)	4/100	1/100	0

表３では，P(cat)×P(果物 ケーキ|cat)が最大になるのは「いちご」のときです。そのことから，単純ベイズ分類器での分類結果として「いちご」が得られるという仕組みを説明しました。

その後，プログラムを配布して，プログラムの要点を説明し，生徒は実際に動作させてプログラムが文章を推測して回答する様子を確認しました。

tf-idfと単純ベイズ分類器を組み合せた理由は，どちらのアルゴリズムも形態素の出現頻度に着目

し，共通していることが多いためです。

(5) グループで問題解決をする授業

これまでの授業で得た知識・技能を活用できるよう，グループで問題解決に取り組む授業を行いました。3～4人のグループで次の課題に取り組みました。

課題：「tf-idfと単純ベイズ分類器を，社会における問題発見や解決に用いる方法を考えて，社会における問題解決を提案しなさい」

生徒一人一人が問題を考え，それをもとにグループで取り組む問題を決め，解決に取り組みました。グループごとに取り組んだ問題は，表4のようなものです。

表4　生徒がグループで取り組んだ問題

- 旅行でどこに行けばよいか（観光地の特徴）
- どの商品を買えばよいか（商品の特徴）
- どの店に行けばよいか（店の特徴）
- 花粉症の原因はどの花か（症状の特徴）
- 料理に合う芋の品種はなにか（品種の特徴）

表4に示したような問題をtf-idfや単純ベイズ分類器で扱うため，生徒は学習データとなる文章をWebから収集しました。その後，Pythonのプログラムの一部を修正してデータを与え，プログラムを実行して出力された結果を分析し，問題を解決するための解決策を検討しました。その後，グループごとに問題の解決に向けた提案をまとめ，クラス内で発表を行いました。その際のスライドの一部を図4に示します。

この授業により，次のような効果があったと考えています。

・人工知能を支える技術の1つである機械学習について生徒が知ることができた
・知識・技能を身の回りの問題に活用できた
・コンピュータによる出力結果の根拠を考え，批判的に考えることができた

4．おわりに

本稿では，tf-idfとWord Cloudを用いた特徴語の抽出および単純ベイズ分類器による機械学習についての授業実践を紹介しました。「情報Ⅰ」の「情報通信ネットワークとデータの活用」では，単語の重要度やタグクラウドという内容が学習指導要領解説に示されているので，tf-idfの考え方が参考にできると思います。

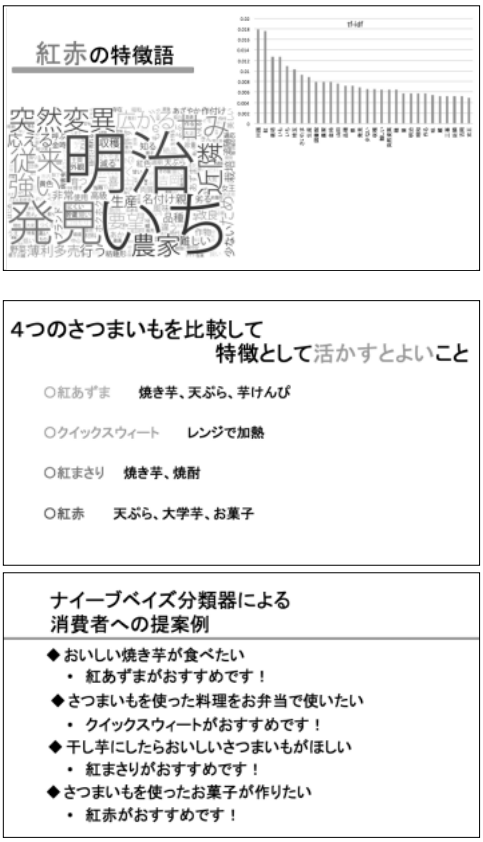

図4　生徒の成果物

「情報Ⅱ」については，難易度が高いため開講を考えていない学校もあるかもしれません。しかし，2022年に入学する高校生が，社会の中核を担う38歳の時にシンギュラリティがくると言われている2045年を迎えます。人工知能が人間を追い越すかどうかはわかりませんが，膨大なデータが存在し，それらのデータをうまく活用することが求められる時代は確実にやってくると思います。

生徒が生きる時代を考え，「情報Ⅱ」がより多く開講されることを願っています。

参考文献

[1] 文部科学省，“初等中等教育における教育課程の基準等の在り方について（諮問）”，(2014)