平成30年度　課題研究

迷惑メールフィルタリングソフト開発

目　的

大学入学後に学ぶ機械学習に触れるため

概　要

機械学習に適したプログラミング言語Pythonを用いて、サポートベクターマシン（SVM）を実装し、迷惑メールフィルタリングソフトを開発する。

説　明

形態素解析、SVMを用いて迷惑メールフィルタリングを行う。

形態素解析

形態素解析とは...
文章を品詞ごとに分解します。
EX：今日も元気です。
　　今日/も/元気/です/。

図1．形態素解析のソースコード

今回はメールを形態素解析し、名詞のみを取り出しています。
取り出した名詞を出現回数が多い順に100個取り出し、出現回数を0～1の数値に変換しています。
作成した辞書（コーパス）をもとに分類を行います。

図2．スパメールのコーパス

サポートベクターマシン（SVM）

サポートベクターマシンとは...
境界線を引いてデータを分類するアルゴリズムです。線形SVMと非線形SVMがあります。

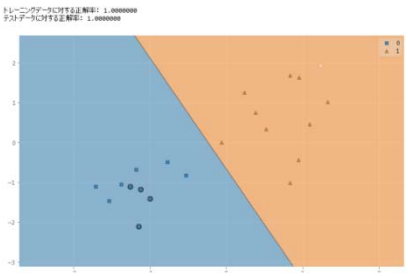
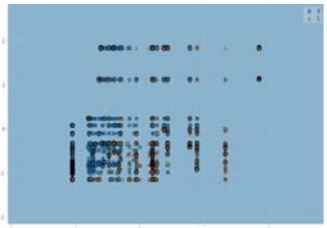


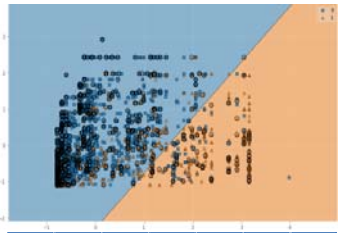
図3．線形SVMによる分類結果の一例

結　果



回数	1回目	2回目	3回目	4回目	5回目	平均
正答率	86.79	87.79	86.25	87.52	88.51	87.37

図4．線形SVM改善前



回数	1回目	2回目	3回目	4回目	5回目	平均
正答率	90.89	90.79	90.80	90.89	92.32	91.13

図5．線形SVM改善後

図4のような見えない位置に境界線が引かれてしまった。そのため値の計算式の改変、文章の正規化を改善した結果、図5のようにデータが分散し、境界線も見える位置に引かれるようになった。

考　察

形態素解析では、名詞以外の品詞も用いることで、精度向上ができると見込める。

感　想

活動を通して、初歩的ではあるが機械学習について学び活用することができた。
グループ活動で個々の役割を分担することにより、作業を効率的に行えたためグループ活動の大切さを再確認することができた。