自然言語処理勉強会

1.1 ~ 1.2

Agenda

1. 必要な数学的知識

1. 準備と本書における約束事

2. 最適化問題

1.1 準備と本書における約束事

自然言語処理タスクの概観

- 単語分割 (word segmentation)
- 品詞タグ付け (part-of-speech tagging)
- 構文解析 (syntactic parsing)
- 文書分類 (text classification)

など

単語分割 (word segmentation)

与えられた文を単語に分割するタスク。

英語では不要（日本語だと分かち書きに当たる）

品詞タグ付け (part-of-speech tagging)

文中の単語の品詞を推定する。
同じ文字でも品詞が異なることがあり、
高精度なタグ付けは難易度が高い。

構文解析 (synthetic parsing)

文章の構文的な構造を推定する。

文書分類 (text classification)

与えられた文章のクラスを予測する問題

クラスわけは、文章の内容や著者など様々ある

他にも照応解析、質問応答、機械翻訳などなどある。

それぞれのタスクにおいて、
処理の対象となるものをinstanceと呼ぶ。

機械学習は、データから分類規則などを学ぶ枠組みである。NLPの場合は、コーパスと呼ばれる言語データを用いて、言語データを数値化して機械学習を行う。コーパスとは、言語のさまざまな用例を集まりを指す。

コーパス≠辞書　∵ 辞書は用例ではないため

概念の確認

文書dにおける単語wの出現回数　　$n(w,d) \quad or \; n_{w,d}$

文sにおける単語wの出現回数　　$n(w,s) \quad or \; n_{w,s}$

クラスcにおける単語wの出現回数　　$n(w,c) \quad or \; n_{w,c}$

クラスcに属する文書数　　$N(c) \quad or \; N_c$

指示関数

$$\delta(w,d) = \begin{cases} 1 & \textbf{if } w \in d \\ 0 & \textbf{otherwise} \end{cases}$$

1.2 最適化問題

最適化問題とは、ある制約下で関数を最適にする変数値と関数値を求める問題である。一般に以下のように書く。

$$max\ f(x)$$
$$subject\ to\ g(x) \geq 0,\ h(x) = 0$$

最大化したい関数f(x)を目的関数、
最適値を与える変数値を最適解という。

また、g(x)の制約を不等式制約、
h(x)の制約を等式制約と呼ぶ

制約を満たす解のことを実行可能解と呼び、
実行可能解の集合を実行可能領域と呼ぶ。

また、解を「x = 」のように表す形式を閉形式と呼び、
閉形式で問題の解が得られる時、解析的に解けると呼ぶ。