言語処理のための機械学習入門

第1章　必要な数学的知識

茨城大学工学部情報工学科
真下飛瑠

1.1　準備と本書における約束事

- 単語分割
  - 与えられた文を単語に分割するタスク
- 品詞タグ付け
  - 文中の単語の品詞を推定するタスク
- 構文解析
  - 文の構文的な構造を推定するタスク
- 文書分類
  - 与えられた文書に対し、いくつかのクラスのうちの適切なものを決定するタスク
- 事例
  - それぞれのタスクでの対象の処理単位
- コーパス
  - 言語の様々な形の用例の集まり

- $n(w,d)$　$n_{w,d}$
  - 文書 dにおける単語 wの出現回数
- $n(w,s)$　$n_{w,s}$
  - 文　sにおける単語 wの出現回数
- $n(w,c)$　$n_{w,c}$
  - クラス cに属する文書群における単語wの出現回数
- $N(w,c)$　$N_{w,c}$
  - クラス cに属する文書のうち wが出現するような文書の数
- $N(c)$　N_{c}
  - クラス cに属する文書数
- $\delta(w,d)$　$\delta_{w,d}$
  - 文書 dにおいて単語 wが出現したら 1
  - そうでなければ 0となる

1.2　最適化問題

最適化問題とは

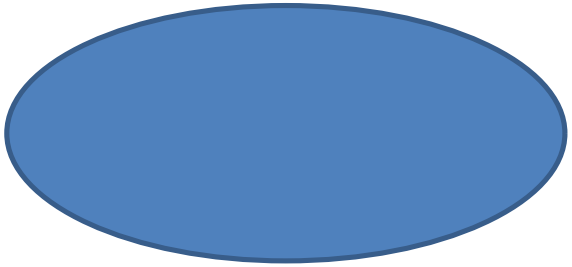
- ある制約のもとで、関数を最大化（または最小化）する変数値とそのときの関数値を求める問題
- それぞれ最大化問題、最小化問題と呼ばれる
- 一般的に次のように書かれる

$$\begin{aligned} \max. \quad & f(x) \\ \text{s.t.} \quad & g(x) \geq 0 \\ & h(x) = 0 \end{aligned}$$

- 実行可能解
  - 制約を満たす解
- 実行可能領域
  - 実行可能解の集合
- 閉形式
  - 「$x_1 =$」などのような解の表し方
- 解析的に解ける
  - 閉形式の解が得られる問題

1.2.1 凸集合と凸関数

- 凸集合とは直感的にいえば
  「へこみのない集合」



（a）凸集合

（b）凸ではない集合

•凸集合の定義

集合$\mathrm{A} \subseteq \mathrm{R}^d$が凸集合であるとは、

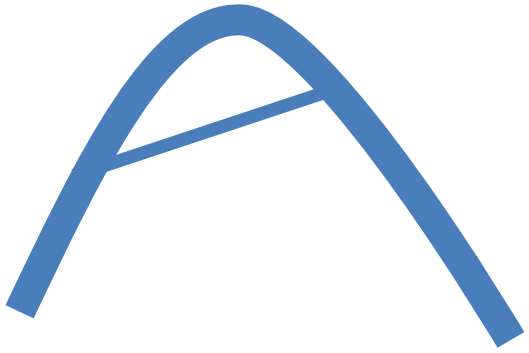
任意の$x^{(1)} \in A$と$x^{(2)} \in A$があったとき、

任意の$t \in [0,1]$に対して

$$tx^{(1)} + (1-t)x^{(2)} \in A$$

が成り立つこと

- 凸関数とは、その値についてへこみがないような関数のこと
- ある関数が上に凸であるとは、その関数を表すグラフ上の任意の2点を結ぶ線分が、常にグラフの下側、もしくは同じ高さにあること



上に凸な関数

- **上に凸であること**

関数$f(x)$が上に凸であるとは、
任意の$x^{(1)}, x^{(2)} \in R^d$に対し、

$$f(tx^{(1)} + (1-t)x^{(2)}) \geq tf(x^{(1)}) + (1-t)f(x^{(2)})$$

が成立すること

- 上に凸な関数であるための1次の条件
  - すべての接線が、その関数のグラフの上側に来ること

$$\text{任意の } x^{(1)}, x^{(2)} \in R \text{について}$$

$$f(x^{(2)}) - f(x^{(1)}) \leq \frac{\partial f(x^{(1)})}{\partial x}(x^{(2)} - x^{(1)})$$

- 上に凸な関数であるための2次の条件
  - 2階微分 $f''(x)$ がつねに負または0であること

1.2.2　凸計画問題

- ある最適化問題が凸計画問題である
  - その目的関数が凸関数
  - かつ実行可能領域が凸集合

制約が与えられていない場合

- 微分して0になる点を求める
- 閉形式で解が求まらない場合、何らかの解を初期値として与え、より良い解へと更新していく（数値解法）
- 数値解法には、最急勾配法やニュートン法がある

等式制約付凸計画問題

- ラクランジュの乗数法で求める

1. ラグランジュ乗数 λ を導入し、ラグランジュ関数L(x, λ)を定義する

$$L(x,\lambda) = f(x) + \lambda g(x)$$

2. 次の連立方程式を解く

$$\nabla_x f(x) + \lambda \nabla_x g(x) = 0$$

$$g(x) = 0$$

不等式制約付き凸計画問題

- 等式制約の場合と同様に、ラグランジュ関数を作る

$$L(x,\lambda) = f(x) + \lambda g(x)$$

  但し、$\lambda \geqq 0$
- 制約を満たすxにおいては、λ g(x)$\geqq$0であるのでL(x, λ)はより大きくなり、最大化の観点からは「お得」
- 制約を満たさないxにおいては、逆に損になる

鞍点

- ラグランジュ関数 $L(x,\lambda)$ について、次のような不等式を満たす点 (x^*,λ^*) のこと

$$L(x,\lambda^*) \le L(x^*,\lambda^*) \le L(x^*,\lambda)$$

- 凸計画問題においては、鞍点が最適解を与えることが知られている

1. まず $L(x,\lambda)$ を最大化するxを求める。一般に、求めたxは λ に依存するので、$x^*(\lambda)$ と表せる
2. $L(x^*(\lambda),\lambda)$ を λ について最小化すると、λ^* が得られる
3. λ^* から 、最適解 x^* が得られる

- 2の問題を、双対問題と呼ぶ

1.3 確率

基礎的な知識

- 事象
  - 起こりうる結果
- 事象空間
  - 考えている事象の集合
- 確率変数
  - 0~1の値をとり、すべての確率変数の値を足すと1になる

$$P(X = x)$$

は確率変数Xがある値xをとるという出来事が起こる確率を表す

- 確率分布
  - 確率変数のどの値がどの程度の確率を持っているかを記述したもの
- 確率関数
  - 確率分布を表した関数
  - 次のように定義される

$$p(x) \equiv P(X = x)$$

  - 単純にP(x)済ませることも多い

1.3.1 期待値、平均、分散

- 確率変数Xの期待値、もしくは平均mxとは、

$$mx = \sum_{x} xP(X = x)$$

  で定義される。
- すべての値についてその確率を掛け、和をとったもの

- 確率変数Xを引数とする関数f(X)についても、次のように計算できる

$$m_{f(X)} = \sum_{x} f(x) P(X = x)$$

- 分散σ^2_xは、確率変数の平均からの離れ具合を表す

$$\sigma_x^2 = \sum_{x} (x - mx)^2 P(X = x)$$

- 標本平均
  - 観測されたデータDに対し次のように定義される

$$\overline{X} = \frac{1}{|D|} \sum_{x^{(i)} \in D} x^{(i)}$$

- 標本分散
  - 標本平均と同様に、次のように定義される

$$S^2 = \frac{1}{|D|} \sum_{x^{(i)} \in D} (x^{(i)} - \overline{X})^2$$

1.3.2　結合確率と条件付き確率

- 確率変数Xがxに、かつYがyとなる確率

$$P(X = x, Y = y)$$

をxとyの結合確率と呼ぶ

- Y=yであることがわかっているときのX=xの確率を条件付き確率と呼ぶ

$$P(X = x \mid Y = y)$$

- 「xとyが同時に起こる」ということは「yが起こって、さらにyが起こったという条件のもとでxが起こる」ということである。つまり次の関係式が成り立つ

$$P(x, y) = P(x \mid y)P(y)$$

- これは引数が増えた場合も、条件が与えられた同時確率についても同様な関係式が成り立つ

ベイズの定理

xは確率変数Xの任意の値であり、

yは確率変数Yの任意の値であるとする

$P(x) \neq 0$のとき次の等式が成り立つ

$$P(y \mid x) = \frac{P(y)P(x \mid y)}{P(x)}$$

- これは条件が増えても同様に成り立つ

1.3.3 独立性

- 2つの確率変数XとYについて考える。XとYそれぞれの任意の値xとyについて

$$P(X = x, Y = y) = P(X = x)P(Y = y)$$

が成り立つとき、XとYは独立であるという

- 条件式を変形すると以下のようになる

$$P(X = x, Y = y) = P(X = x)P(Y = y)$$

$$\Leftrightarrow P(x \mid y)P(y) = P(x)P(y)$$

$$\Leftrightarrow P(x \mid y) = P(x)$$

1.3.4　代表的な離散確率分布

- 確率関数で記述される
- 種類（確率関数の形）とパラメータの値を与えることで、一意に決まる
- パラメータpを明記する場合は、$P(x;p)$ のように、セミコロンのあとに書く
- ある確率分布における値がとりうる範囲のことを標本空間という

ベルヌーイ分布

- とりうる値が2つ(0,1)であるような確率変数を記述する
- 確率変数Xはベルヌーイ分布に従い、確率pでa、確率1-pでbの値をとるとすると、

$$P(X = x; p) = \delta(x, a)p + \delta(x, b)(1 - p)$$
$$= \delta(x, a)p + (1 - \delta(x, a))(1 - p)$$
$$= p^{\delta(x,a)}(1 - p)^{1-\delta(x,a)}$$

と表せる

二項分布

- ベルヌーイ分布と同様に2つの値をとる
- n回の試行でx回が1の確率を与える
- 確率pの事象がx回起こる確率は、

$$P(x;p,n)={}_nC_x p^x(1-p)^{n-x}$$

但し、

$${}_nC_x=\frac{n!}{x!(n-x)!}$$

多項分布

- m個の値をとりうる確率変数があり、各値にはそれぞれp_iの確率が与えられているとする
- 1回の試行でm個の値のうち一つが起こる
- これをn回試行した場合、各値がそれぞれk_m回起こる確率は、

$$\frac{n!}{\prod_i k_i!}\prod_i p_i^{ki}$$

ポアソン分布

- 0以上の整数を値とする確率変数の確率分布
- パラメータμ(>0)として、整数x(≧0)が起こる確率は、

$$P(x|\mu) = \frac{\mu^x}{x!} e^{-\mu}$$

で与えられる

- ポアソン分布は二項分布のある種の極限として得られる

1.4　連続確率変数

- 離散変数以外に、とびとびでない値をとる連続変数がある
- 連続変数であり確率変数であるような変数を連続確率変数と呼ぶ
- 連続の場合、各値をとる確率は一般に限りなく0に近い。よって各値での実数値は確率というより確率のつまり具合を表す

- 確率密度関数
  - 実数値を与える関数
  - p(x)のように小文字で表す
  - 確率として意味を持たせるには、ある範囲で積分する必要がある

$$P(a \leq x \leq b) = \int_a^b p(x)dx$$

1.4.1 平均、分散

- 連続確率変数Xの平均m、分散σ^2は次のように定義される

$$m = \int xp(x)dx$$

$$\sigma^2 = \int (x-m)^2 p(x)dx$$

1.4.2 連続確率分布の例

- 正規分布
  - ガウス分布とも呼ばれる
  - 1次元の正規分布の確率密度関数は

$$p_{gauss}(x;m,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-m)^2}{2\sigma^2}\right)$$

  - で定義される。mとσ²はそれぞれ平均と分散である

- ディリクレ分布

$x_i \geq 0, \sum_i x_i = 1$であるような

$x = (x_1, \cdots, x_n)$に対して確率を与える分布

$$p(x;\alpha) = \frac{1}{\int \prod_i x^{\alpha_i - 1}{}_i dx} \prod_i x^{\alpha_i - 1}{}_i$$

- 分母は積分すると1になるよう導入されている
- ディリクレ分布に従う確率変数は一般に極端な値をとりにくい
- 多項分布のパラメータの確率分布を表すことができ、実際にその目的でよく使われる