線形識別モデル（PRML 第４章）

確率的識別モデル (4.3)，ラプラス近似 (4.4)，ベイズロジスティック回帰 (4.5)

岡崎　直観

東京大学・辻井研究室

2009-03-16

もくじ

確率的識別モデル

識別モデル

- 条件付き確率分布 $p(C_k|x)$ を直接モデル化
- 適応パラメータの数が少なくて済む
- x に関する様々な特徴量を組み込める

生成モデル

- 条件付き確率分布 $p(x|C_k)$ と，クラスの事前確率 $p(C_k)$ をモデル化する
- 事後確率 $p(C_k|x)$ は，モデル化された $p(x|C_k)$ と $p(C_k)$ から，ベイズの定理で計算する

固定基底関数

- 元の入力ベクトル x の特徴を，基底関数ベクトル $\phi(x)$ を用いて，非線形空間に写像する
  - 識別モデルは，特徴空間 $\phi(x)$ において線形の決定境界を構成するが，元の観測空間 x では非線形になる
  - もちろん，$\phi(x) = x$ となる線形の基底関数でも可
  - 基底関数 $\phi_0(x) = 1$ を導入すれば，対応するパラメータ w_0 はバイアスの役割を担う
- ただし，固定基底関数は元の入力ベクトル x における重なりを除去できない
  - C_1 と C_2 に分類される２つの事例が，x の空間上で同じ位置にある時は，$\phi(x)$ の空間でも同じ場所に写像
- 図 4.12 を参照

ロジスティック回帰とは

- **二値分類問題**
  - **入力の特徴空間ベクトル $\phi = \phi(x)$ を，クラス C_1 もしくは C_2 に分類する**
  - **実際は「回帰」ではなくて「分類」を行うが，統計学において「ロジスティック回帰」と呼ばれていた**
  - **線形の** Support Vector Machine (SVM) **と非常に似ていて，主な違いは誤差関数**
- **特徴**
  - **条件付き確率 $P(C_1|\phi)$ 及び $P(C_2|\phi)$ が推定できる**
  - **大規模な学習データにも適用可能**
    - **確率的勾配降下法と組み合わせれば，非常に簡単**
  - **最近でも研究が続けられている** [LLAN06, LWK07]

クラスC_1，C_2の事後確率

事後確率

$$p(C_1|\phi) = y(\phi) = \sigma(w^T\phi)$$
$$p(C_2|\phi) = 1 - p(C_1|\phi) = 1 - y(\phi)$$

解釈

- 入力特徴ϕの重みの和（内積；スコア）を計算
- シグモイド関数を用いて，値域$[-\infty, +\infty]$のスコアを，確率値$[0, 1]$に変換
- $p(C_1|\phi) + p(C_2|\phi) = 1$を満たす

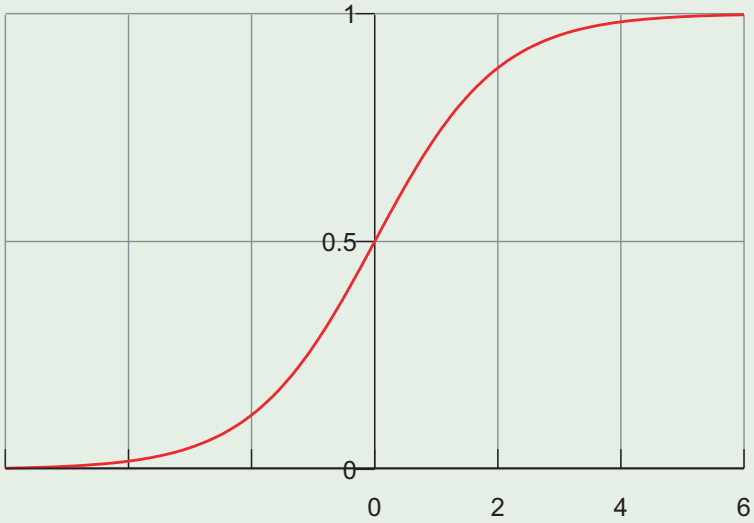
シグモイド関数

シグモイド関数

$$\sigma(a) = \frac{1}{1 + \exp(-a)}$$

- コンピュータで計算するときは，a が負のときにオーバーフローしやすいので注意

シグモイド関数の形



- $a = 0$ に関して対称
- $\lim_{a \to \infty} \sigma(a) = 1$
- $\lim_{a \to -\infty} \sigma(a) = 0$

評判分析（極性判定）の例

- 例文
  - スープ に こく が ある
  - スープ に 調味料 が 入っ て いる
- よい評判（クラス C_1）と，悪い評判（クラス C_2）に分類
- 機能語を除去し，索引語を要素とする特徴ベクトル ϕ を作成:
  (bias, スープ, こく, 調味料, ある, 入る, いる)^T
  - $\phi_1 = (1, 1, 1, 0, 1, 0, 0)^T$
  - $\phi_2 = (1, 1, 0, 1, 0, 1, 1)^T$
- 重みベクトルはすでに学習で求めてあり，
  $w = (-0.5, 0.1, 1.0, -0.7, 0.1, 0.1, 0.0)^T$ だったとする
  - $p(C_1|\phi_1) = \sigma(-0.5 + 0.1 + 1.0 + 0.1) = \sigma(0.7) = 0.668$
  - $p(C_1|\phi_2) = \sigma(-0.5 + 0.1 - 0.7 + 0.1) = \sigma(-1.0) = 0.269$

最尤法による学習

- 学習とはパラメータ（特徴量に対する重み）を，データにフィットするように調整すること
- クラス C_1 と C_2 を，２値変数 $t \in \{1, 0\}$ で表す
- N 個の学習事例 $\{\phi_n, t_n\}_{n=1}^{N}$ が与えられる

学習データに対する尤度関数

$$p(t|w) = \prod_{n=1}^{N} y_n^{t_n}(1 - y_n)^{1-t_n}$$

y_n は事例 ϕ_n がクラス C_1 に分類される確率

誤差関数

交差エントロピー誤差関数

$$
\begin{aligned}
E(\boldsymbol{w}) &= -\ln p(\boldsymbol{t}|\boldsymbol{w}) \\
&= -\ln \left[\prod_{n=1}^{N} y_n^{t_n} (1-y_n)^{1-t_n} \right] \\
&= -\sum_{n=1}^{N} \{ t_n \ln y_n + (1-t_n) \ln(1-y_n) \}
\end{aligned}
$$

- ここで，$y_n = \sigma(a_n)$，$a_n = \boldsymbol{w}^T \boldsymbol{\phi}_n$
- y_n は $\boldsymbol{\phi}_n$ がクラス C_1（$t_n = 1$）に属する確率

シグモイド関数の微分

ロジスティック回帰モデルのパラメータを最尤法を用いて決定する際，シグモイド関数の微分が必要になる．

演習 4.12.

$$\sigma(a) = \frac{1}{1+\exp(-a)}$$

$$\frac{d\sigma}{da} = (-1) \cdot \frac{1}{\{1+\exp(-a)\}^2} \cdot \exp(-a) \cdot (-1)$$

$$= \frac{1}{1+\exp(-a)} \cdot \frac{\exp(-a)}{1+\exp(-a)} = \sigma(1-\sigma)$$

□

誤差関数の勾配

演習 4.13.

$$\nabla E(\boldsymbol{w}) = \sum_{n=1}^{N} \frac{\partial E}{\partial y_n} \frac{\partial y_n}{\partial a_n} \frac{\partial a_n}{\partial \boldsymbol{w}} = \sum_{n=1}^{N} (y_n - t_n)\boldsymbol{\phi_n}$$

$$\frac{\partial E}{\partial y_n} = -\left(\frac{t_n}{y_n} + \frac{1 - t_n}{1 - y_n} \cdot (-1) \right) = \frac{y_n - t_n}{y_n(1 - y_n)}$$

$$\frac{\partial y_n}{\partial a_n} = \frac{\partial \sigma(a_n)}{\partial a_n} = \sigma(a_n)\{1 - \sigma(a_n)\} = y_n(1 - y_n)$$

$$\frac{\partial a_n}{\partial \boldsymbol{w}} = \boldsymbol{\phi_n}$$

□

誤差関数の勾配の解釈

$\nabla E(w)$ の解釈

$$-\nabla E(w) = \sum_{n=1}^{N} (t_n - y_n)\phi_n$$

- y_n と t_n が一致するのであれば，特徴ベクトル ϕ_n の重みを修正する必要がない
- $t_n = 1$ のとき，$y_n < 1$ であれば，その誤差に応じて，特徴ベクトル ϕ_n の重みを増やす
- $t_n = 0$ のとき，$0 < y_n$ であれば，その誤差に応じて，特徴ベクトル ϕ_n の重みを減らす

確率的勾配降下法（SGD）で w を求める

オンライン学習

- 一つずつ取り出した事例に対して，重み w の更新則を適用する

SGD によるロジスティック回帰の学習

1. $n \in \{1, ..., N\}$ に対して，以下の処理を繰り返す
   1. 特徴ベクトル ϕ_n を取り出す
   2. 誤差 $\nabla E(w) = (t_n - y_n)\phi_n$ を計算する
   3. パラメータベクトルを更新: $w^{\text{new}} = w^{\text{old}} - \eta \nabla E(w)$
2. 必要があればステップ 1 に戻る

Python による実装例

```
N = 17997       # Change this to present the number of training instances.
eta0 = 0.1      # Initial learning rate; change this if desired.

def update(W, X, l, eta):
    # Compute the inner product of features and their weights.
    a = sum([W[x] for x in X])
    # Compute the gradient of the error function (avoiding +Inf).
    g = ((1. / (1. + math.exp(-a))) - l) if -100. < a else (0. - l)
    # Update the feature weights by Stochastic Gradient Descent (SGD).
    for x in X:
        W[x] = W[x] - eta * g

def train(fi):
    t = 1
    W = collections.defaultdict(float)
    # Loop for instances.
    for line in fi:
        fields = line.strip('\n').split('\t')
        update(W, fields[1:], float(fields[0]), eta0 / (1 + t / float(N)))
        t += 1
    return W
```

最尤推定と過学習

演習 4.14

- データ集合が線形分離可能となる条件
  - $t_n = 1$ となるすべての n に対して，$w\phi_n > 0$
  - $t_n = 0$ となるすべての n に対して，$w\phi_n < 0$
- すべての n に対して $p(C_k|x_n) = 1$ にしたい
  - $t_n = 1$ となるすべての n に対して，$a_n \to +\infty$
  - $t_n = 0$ となるすべての n に対して，$a_n \to -\infty$
- これを実現するには，w の大きさを無限大に発散
  - $w = \alpha v$ に分解し，$\alpha \to \infty$ とすれば，すべての a_n が原点から $+\infty$ もしくは $-\infty$ に向かっていく
- 従って，線形分離可能なデータ集合に対して，その最尤解は w の大きさが ∞ の時に得られる

過学習を防止する方法

- 重み w の事前分布を仮定し，MAP 推定を行う
- 誤差関数に正則化項を付加する

重みベクトルの二乗和による正則化

$$E(\boldsymbol{w}) = -\sum_{n=1}^{N} \{t_n \ln y_n + (1 - t_n) \ln(1 - y_n)\} + \frac{C}{2} \boldsymbol{w}^T \boldsymbol{w}$$

$$\nabla E(\boldsymbol{w}) = \sum_{n=1}^{N} (y_n - t_n) \boldsymbol{\phi}_n + C\boldsymbol{w}$$

反復再重み付け最小二乗

- ニュートン—ラフソン法でロジスティック回帰の学習を行う
- 簡単な例として，二乗和誤差関数による線形回帰モデルに，ニュートン—ラフソン法を適用する
- 次に，ロジスティック回帰モデルにニュートン—ラフソン法を適用し，反復重み付き最小二乗法（IRLS: iterative reweighted least squares method）を得る

ニュートン—ラフソン法

$f(x) = 0$ となる x を近似的に求める方法．
ある点 x_0 の周りで，$f(x)$ をテーラー展開する．

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2 + \ldots$$

２次以上の項は無視して，$f(x)$ を近似する

$$f(x) \approx f(x_0) + f'(x_0)(x - x_0)$$

近似した $f(x)$ に対して，$f(x) = 0$ とすれば，

$$x = x_0 - \frac{f(x_0)}{f'(x_0)}$$

誤差関数 $E(w)$ の最小化

- 誤差関数が凸関数ならば，唯一の最小解を持つ
  - ロジスティック回帰の誤差関数は，凸関数である（後述）
- $\nabla E(w) = 0$ となる w を求める
  - $x \leftarrow w$，$f \leftarrow \nabla E$，$f' \leftarrow H$（ヘッセ行列）と置換

ニュートン法による $E(w)$ 最小化

$$w^{(new)} = w^{(old)} - H^{-1}\nabla E(w)$$

線形回帰モデルにニュートン法を適用 (1/2)

$E(\boldsymbol{w})$ から，勾配 $\nabla E(\boldsymbol{w})$ と二階微分 H を求める

$$E(\boldsymbol{w}) = \frac{1}{2}\sum_{n=1}^{N}\left(t_n - \boldsymbol{w}^T\boldsymbol{\phi}_n\right)^2$$

$$\nabla E(\boldsymbol{w}) = \sum_{n=1}^{N}\left(\boldsymbol{w}^T\boldsymbol{\phi}_n - t_n\right)\boldsymbol{\phi}_n = \Phi^T\Phi\boldsymbol{w} - \Phi^T\boldsymbol{t}$$

$$H = \nabla\nabla E(\boldsymbol{w}) = \sum_{n=1}^{N}\boldsymbol{\phi}_n\boldsymbol{\phi}_n^T = \Phi^T\Phi$$

ただし，Φ は $N \times M$ の行列

$$\Phi = \begin{pmatrix}\boldsymbol{\phi}_1^T \\ \cdots \\ \boldsymbol{\phi}_N^T\end{pmatrix} = \begin{pmatrix}\phi_0(\boldsymbol{x}_1) & \cdots & \phi_{M-1}(\boldsymbol{x}_1) \\ \cdots & \cdots & \cdots \\ \phi_0(\boldsymbol{x}_N) & \cdots & \phi_{M-1}(\boldsymbol{x}_N)\end{pmatrix}$$

線形回帰モデルにニュートン法を適用 (2/2)

ニュートン—ラフソン法による更新式は，

$$
\begin{aligned}
\boldsymbol{w}^{(\mathrm{new})} &= \boldsymbol{w}^{(\mathrm{old})} - H^{-1}\nabla E(\boldsymbol{w}) \\
&= \boldsymbol{w}^{(\mathrm{old})} - (\Phi^T\Phi)^{-1}\left\{\Phi^T\Phi\boldsymbol{w}^{(\mathrm{old})} - \Phi^T\boldsymbol{t}\right\} \\
&= (\Phi^T\Phi)^{-1}\Phi^T\boldsymbol{t}
\end{aligned}
$$

$w^{(\mathrm{old})}$ が消え，反復回数１回で正確な解が求まる．これは，解析的に求めた解 w_{ML}（式 3.15）と同じである．

ロジスティック回帰の勾配とヘッセ行列

誤差関数の勾配 $\nabla E(w)$ とヘッセ行列 H は，

$$\nabla E(\boldsymbol{w}) = \sum_{n=1}^{N} (y_n - t_n)\boldsymbol{\phi_n} = \Phi^T(\boldsymbol{y} - \boldsymbol{t})$$

$$H = \sum_{n=1}^{N} y_n(1 - y_n)\boldsymbol{\phi_n}\boldsymbol{\phi_n}^T = \Phi^T R \Phi$$

ここで，R は $N \times N$ の対角行列で，その要素 R_{nn} は，

$$R_{nn} = y_n(1 - y_n)$$

ロジスティック回帰の誤差関数の凸関数性

演習 4.15.

$r_n = y_n(1-y_n)$，$\rho_n = \frac{1}{\sqrt{r_n}}\phi_n$ とおくと，

$$H = \sum_{n=1}^{N} y_n(1-y_n)\phi_n\phi_n^T = \sum_{n=1}^{N} \rho_n\rho_n^T$$

ゼロではない任意のベクトル u に対して，

$$u^T H u = \sum_{n=1}^{N} u^T \rho_n \rho_n^T u$$

ϕ_n のいずれかが 0 でなければ，$0 < r_n < 1$ より，ρ_n もゼロではない．よって，$u^T H u > 0$ が成り立ち，ヘッセ行列 H は正定値行列であるから，$E(w)$ は w の凸関数であり，大域的最小解を持つ．□

ニュートン法によるロジスティック回帰の学習

ニュートン—ラフソン法による更新式は，

$$
\begin{aligned}
\boldsymbol{w}^{(\mathrm{new})} &= \boldsymbol{w}^{(\mathrm{old})} - H^{-1}\nabla E(\boldsymbol{w}) \\
&= \boldsymbol{w}^{(\mathrm{old})} - (\Phi^T R \Phi)^{-1}\Phi^T(\boldsymbol{y} - \boldsymbol{t}) \\
&= (\Phi^T R \Phi)^{-1}\left\{\Phi^T R \Phi \boldsymbol{w}^{(\mathrm{old})} - \Phi^T(\boldsymbol{y} - \boldsymbol{t})\right\} \\
&= (\Phi^T R \Phi)^{-1}\Phi^T R \boldsymbol{z}
\end{aligned}
$$

ただし，z は以下を要素とする N 次元ベクトル

$$
\boldsymbol{z} = \Phi \boldsymbol{w}^{(\mathrm{old})} - R^{-1}(\boldsymbol{y} - \boldsymbol{t})
$$

多クラスロジスティック回帰とは

- **多クラス分類問題**
  - **入力の特徴空間ベクトル $\phi = \phi(x)$ を，K 個のクラス $C_1, ..., C_K$ のいずれかに分類する**
  - **自然言語処理の分野では，最大エントロピー法（Maximum Entropy）と呼ばれる**
  - **入力と出力に構造を持たせ，より一般化させたものが条件付き確率場（CRF: Conditional Random Field）**
- **特徴**
  - **条件付き確率 $P(C_k|\phi)$ が推定できる**
  - **大規模な学習データにも適用可能**
  - **確率をよく使う NLP ではポピュラーな分類器**

クラスC_kの事後確率

事後確率

$$p(C_k|\phi) = y_k(\phi) = \frac{\exp(a_k)}{\sum_j \exp(a_k)}$$

$$a_k = \boldsymbol{w}_k^T \phi$$

解釈

- 入力特徴ϕと，クラスkに関する重みw_kの内積（スコア）を計算し，a_kとする
- すべてのクラスに関するスコアa_jの和$\sum_j a_j$が1になるように，a_kを正規化して，確率値に変換

単語の品詞推定の例

英語の名詞を単数形（C_1）か複数形（C_2）に分類
単語を３次元の特徴ベクトル ϕ で表す:

(**バイアス項**, **単語の末尾が** 's' **で終わる**, **単語の末尾が** 'us' **で終わる**)

'book", "books", "virus" **に対して，特徴ベクトルはそれぞれ，**

$$\phi_1 = (1, 0, 0), \phi_2 = (1, 1, 0), \phi_3 = (1, 1, 1)$$

クラス C_1 と C_2 の重みベクトル，w_1 と w_2 を学習で求めたとする

$$w_1 = (0.7, -1.5, 1.7), w_2 = (-0.1, 1.5, -0.7)$$

与えられた単語が複数形である確率 $P(C_2|\phi)$ を計算すると，

$$a_{1,1} = w_1^T \phi_1 = 0.7, \quad a_{1,2} = w_2^T \phi_1 = -0.1, \quad P(C_2|\phi_1) = \frac{e^{-0.1}}{e^{0.7} + e^{-0.1}} = 0.310$$

$$a_{2,1} = w_1^T \phi_2 = -0.8, \quad a_{2,2} = w_2^T \phi_2 = 1.4, \quad P(C_2|\phi_2) = \frac{e^{1.4}}{e^{-0.8} + e^{1.4}} = 0.900$$

$$a_{3,1} = w_1^T \phi_3 = 0.9, \quad a_{3,2} = w_2^T \phi_3 = 0.7, \quad P(C_2|\phi_3) = \frac{e^{0.7}}{e^{0.9} + e^{0.7}} = 0.450$$

事後確率 y_k の a_j に関する偏微分

演習 4.17.

y_k を a_k と a_l（ただし，$k \neq l$）で微分した後，まとめる

$$\frac{\partial y_k}{\partial a_k} = \frac{e^{a_k}\sum_i e^{a_i} - e^{a_k}e^{a_k}}{\left(\sum_i e^{a_i}\right)^2} = y_k - y_k^2,$$

$$\frac{\partial y_k}{\partial a_l} = \frac{0 \cdot \sum_i e^{a_i} - e^{a_k}e^{a_l}}{\left(\sum_i e^{a_i}\right)^2} = -y_k y_l,$$

$$\therefore \frac{\partial y_k}{\partial a_j} = y_k(I_{kj} - y_j)$$

□

尤度関数

- クラス $C_1, ..., C_K$ に対して，1-of-K 表記法を用いて，目的変数ベクトル t を表す
  - 例えば $K = 5$ の場合，C_2 は $t = (0, 1, 0, 0, 0)^T$
- N 個の学習事例 $\{\phi_n, t_n\}_{n=1}^N$ が与えられる
  - 目的変数ベクトル t_n を N 個まとめて，t_{nk} を要素とする $N \times K$ 行列 T で表す

学習データに対する尤度関数

$$p(T|w_1, ..., w_K) = \prod_{n=1}^{N} \prod_{k=1}^{K} p(C_k|\phi_n)^{t_{nk}} = \prod_{n=1}^{N} \prod_{k=1}^{K} y_{nk}^{t_{nk}}$$

誤差関数

交差エントロピー誤差関数

- 尤度の負の対数をとったもの

$$E(\boldsymbol{w}_1, ..., \boldsymbol{w}_K) = -\ln p(T|\boldsymbol{w}_1, ..., \boldsymbol{w}_K) = -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk} \ln y_{nk}$$

誤差関数の勾配

演習 4.18.

$$
\begin{aligned}
\frac{\partial E(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_K)}{\partial \boldsymbol{w}_j} &= -\sum_{n=1}^{N}\sum_{k=1}^{K} \frac{\partial}{\partial y_{nk}} (t_{nk} \ln y_{nk}) \frac{\partial y_{nk}}{\partial a_{nj}} \frac{\partial a_{nj}}{\partial \boldsymbol{w}_j} \\
&= -\sum_{n=1}^{N}\sum_{k=1}^{K} \frac{t_{nk}}{y_{nk}} \cdot y_{nk} \left(I_{kj} - y_{nj} \right) \cdot \phi_n \\
&= -\sum_{n=1}^{N} \phi_n \left\{ \sum_{k=1}^{K} t_{nk} I_{kj} - \sum_{k=1}^{K} t_{nk} y_{nj} \right\} \\
&= -\sum_{n=1}^{N} \phi_n (t_{nj} - y_{nj}) = \sum_{n=1}^{N} (y_{nj} - t_{nj}) \phi_n
\end{aligned}
$$

□

誤差関数の最小化

- 誤差関数の勾配には，二乗和誤差関数，及びロジスティック回帰モデルと同様に，誤差 $(y_{nj} - t_{nj})$ と基底関数 ϕ_n の積が現れる
- ロジスティック回帰ど同様に，確率的勾配降下法を用いて逐次学習を行ってもよい
- ヘッセ行列 H を求めて，ニュートン—ラフソン法に基づくバッチ学習を構成することもできる
  - ヘッセ行列 H のサイズは $MK \times MK$ になり，H 自体の計算や，逆行列 H^{-1} の計算が大変になるので，L-BFGS 法などで H^{-1} を近似することが多い

プロビット回帰とは

- 活性化関数は，標準ガウス分布の累積分布関数
  - 累積分布関数の逆関数はプロビット関数と呼ばれる
  - ロジスティック回帰では，シグモイド関数を用いた
- ロジスティック回帰の結果と似る傾向がある
  - ロジスティック回帰のベイズ的な扱いにおいて，プロビット関数が近似として用いられる
- 外れ値に対して敏感
  - 外れ値とは，入力ベクトル x の測定誤差や，目的変数値 t の間違ったラベル付け

標準ガウス分布の累積分布関数

標準ガウス分布の累積分布関数

$$\Phi(a) = \int_{-\infty}^{a} \mathcal{N}(\theta|0,1)d\theta$$

- 平均が 0，分散が 1 の標準ガウス分布で与えられる確率密度を $[-\infty, a]$ の区間で累積する
- シグモイド関数と似ている（図 4.9 参照）
- 一般的なガウス分布を利用しても，線形係数 w のリスケーリングと等価で，モデルの形が変化することはない

累積分布関数を erf 関数で表す

演習 4.21.

$$\Phi(a) = \int_{-\infty}^{a} \mathcal{N}(\theta|0,1)d\theta = \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\theta^2} d\theta$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{a}{\sqrt{2}}} e^{-t^2} \cdot \sqrt{2} dt$$

$\theta = \sqrt{2}t$ **として置換積分**

$$= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{0} e^{-t^2} dt + \frac{1}{\sqrt{\pi}} \int_{0}^{\frac{a}{\sqrt{2}}} e^{-t^2} dt$$

積分の区間を分解

$$= \frac{1}{\sqrt{\pi}} \cdot \frac{\sqrt{\pi}}{2} + \frac{1}{2} \mathrm{erf}\left(\frac{a}{\sqrt{2}}\right)$$

ガウス積分 $\int_{-\infty}^{\infty} e^{-t^2} dt = \sqrt{\pi}$

$$= \frac{1}{2}\left\{1 + \mathrm{erf}\left(\frac{a}{\sqrt{2}}\right)\right\}$$

□

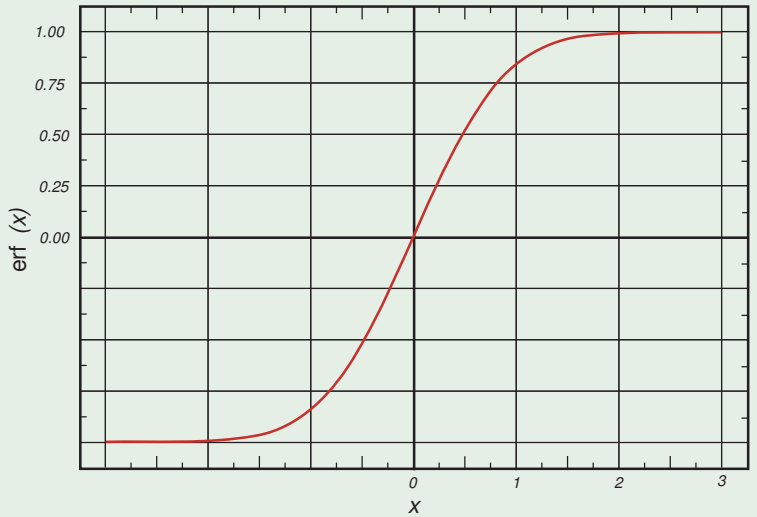
誤差関数（erf 関数）

erf 関数

$$\mathrm{erf}(a) = \frac{2}{\sqrt{\pi}} \int_0^a e^{-\theta^2} d\theta$$

- erf 関数を解析的に計算するのは困難
- C99 では math.h で定義されており，近似計算が容易

erf 関数の形



- $a = 0$ に関して対称
- $\lim_{a \to \infty} \sigma(a) = 1$
- $\lim_{a \to -\infty} \sigma(a) = -1$

誤差関数の最小化

- これまでの最尤推定法に関する議論を拡張すれば，プロビット回帰モデルのパラメータを求めることができる
- 最尤推定の結果は，ロジスティック回帰と似る傾向がある
- プロビット回帰の方が外れ値に対して敏感
  - シグモイド関数では $x \to \infty$ に対して，$\exp(-x)$ で減衰
  - 累積分布関数では $x \to \infty$ に対して，$\exp(-x^2)$ で減衰

正準連結関数

一般化線形モデル

入力変数 ϕ の線形結合に非線形関数 f を適用し，y を得る

$$y = f(w^t \phi)$$

ここで，$f(\cdot)$ を活性化関数（activation function），$f^{-1}(\cdot)$ を連結関数（link function）と呼ぶ．

正準連結関数

パラメータベクトル w に関する事例 n での誤差関数の微分は，誤差 $(y_n - t_n)$ と特徴ベクトル ϕ_n の積で表される．これは，正準連結関数と知られている関数を活性化関数に選び，指数分布族の中から，条件付き確率分布を選択することによる，一般的な結果．

ラプラス近似とは

- 連続変数の集合上に定義される確率密度分布 $p(z)$ に対し，ガウス分布による近似を得る．
- 確率密度分布におけるモード（最頻値；極値）を中心に近似するのが，基本的なアイディア
- 尤度を積分により計算することが困難なとき，積分したい関数をガウス分布で近似するテクニック

１変数 z における分布 $p(z)$ の近似 (1/2)

連続な 1 変数 z における確率密度分布，

$$p(z) = \frac{1}{Z} f(z), Z = \int f(z) dz$$

分布 $p(z)$ のモードを中心とするガウス分布で近似したい．モード z_0 を中心とした $\ln f(z)$ のテイラー展開を考える．

$$\ln f(z) \approx \ln f(z_0) + \frac{\partial \ln f(z_0)}{\partial z}(z - z_0) + \frac{1}{2}\frac{\partial^2 \ln f(z_0)}{\partial z^2}(z - z_0)^2 + \ldots$$

３次以上の項は無視する．$p(z)$ のモード z_0 において，$p'(z_0) = 0$，すなわち $f'(z_0) = 0$ であるから，１次の項は消去できる．

$$\frac{\partial \ln f(z_0)}{\partial z}(z - z_0) = \frac{1}{f(z_0)}\frac{\partial f(z_0)}{\partial z}(z - z_0) = 0$$

１変数 z における分布 $p(z)$ の近似 (2/2)

ここで，

$$a = -\left.\frac{\partial^2}{\partial z^2} \ln f(z)\right|_{z=z_0}$$

とおくと，$\ln f(z)$ の近似は，

$$\ln f(z) \approx \ln f(z_0) - \frac{1}{2}a(z - z_0)^2$$

両辺の指数を取ると，

$$f(z) \approx f(z_0) \exp\left\{-\frac{1}{2}a(z - z_0)^2\right\}$$

ガウス分布の正規化のための標準的な結果 (2.42) を利用すると，

$$f(z) \approx \left(\frac{a}{2\pi}\right)^{1/2} \exp\left\{-\frac{1}{2}a(z - z_0)^2\right\}$$

M 次元空間上で定義される分布 $p(z)$ の近似

ラプラス近似を M 次元空間上で定義される分布 $p(z)$ に拡張する．
１変数の時と同様に，対数を取った近似は，

$$\ln f(z) \approx \ln f(z_0) - \frac{1}{2}(z - z_0)^T A(z - z_0)$$

ここで，A は $M \times M$ のヘッセ行列

$$A = -\nabla\nabla \ln f(z)|_{z=z_0}$$

両辺の指数を取ると，

$$f(z) \approx f(z_0) \exp\left\{-\frac{1}{2}(z - z_0)^T A(z - z_0)\right\}$$

ガウス分布の正規化のための標準的な結果 (2.43) を利用すると，

$$q(z) = \frac{|A|^{1/2}}{(2\pi)^{M/2}} \exp\left\{-\frac{1}{2}(z - z_0)^T A(z - z_0)\right\} = \mathcal{N}(z|z_0, A^{-1})$$

ラプラス近似の実用上の問題

- ラプラス近似を適用するには，モード z_0 を見つけ，その上でヘッセ行列 H を評価する必要がある
  - モードは何らかの数値最適化アルゴリズムで求める
  - 真の正規化係数 Z は求めなくてもよい
- 現実に現れる多くの分布は多峰的であり，どのモードで近似するかで結果が異なる
  - ガウス分布でうまく近似できなければならない
  - 真の分布のある１点における局面に基づいてしまうため，全体的な特性を捉えられないことがある

正規化係数 Z の近似

正規化係数 Z を近似した $f(z)$ から求めてみる．

$$
\begin{aligned}
Z &= \int f(z)dz \\
&\approx f(z_0) \int \exp\left\{-\frac{1}{2}(z-z_0)^T A(z-z_0)\right\} dz \\
&= f(z_0)\frac{(2\pi)^{M/2}}{|A|^{1/2}}
\end{aligned}
$$

ここで，正規化ガウス分布に対する積分の結果，及び式 (2.43) を利用した．

モデルエビデンス (1/2)

データ集合 $\mathcal{D}$ に対し，パラメータ $\{\theta_i\}$ を持つモデルの集合を $\{\mathcal{M}_i\}$ を考える．モデルのエビデンスは，式 (3.68) より，

$$p(\mathcal{D}|\mathcal{M}_i) = \int p(\mathcal{D}|w, \mathcal{M}_i)p(w|\mathcal{M}_i)dw$$

この式の右辺は，$f(\theta) = p(\mathcal{D}|w, \mathcal{M}_i)p(w|\mathcal{M}_i)$ とすれば，正規化係数 Z を求めた式 (4.135) と同じ形をしている．通常，事後確率 $p(w|\mathcal{D}, \mathcal{M}_i) \propto p(\mathcal{D}|w, \mathcal{M}_i)p(w|\mathcal{M}_i)$ は，MAP 解 w_{MAP} を中心に急なピークを示すので，式 (4.135) の z_0 を w_{MAP} とすると，

$$p(\mathcal{D}|\mathcal{M}_i) \approx p(\mathcal{D}|w_{\mathrm{MAP}}, \mathcal{M}_i)p(w_{\mathrm{MAP}}|\mathcal{M}_i)\frac{(2\pi)^{M/2}}{|A|^{1/2}}$$

両辺の対数を取り，表記の簡略化のため $\mathcal{M}_i$ を省略すれば，

$$\ln p(\mathcal{D}) \approx \ln p(\mathcal{D}|w_{\mathrm{MAP}}) + \ln p(w_{\mathrm{MAP}}) + \frac{M}{2}\ln(2\pi) - \frac{1}{2}\ln|A|$$

モデルエビデンス (2/2)

ここで，θ_{MAP} は事後確率分布のモードでの θ の値，A は事後確率の負の対数のヘッセ行列

$$A = -\nabla\nabla \ln p(\mathcal{D}|\boldsymbol{w}_{\mathrm{MAP}}) \ln p(\boldsymbol{w}_{\mathrm{MAP}}) = -\nabla\nabla \ln p(\boldsymbol{w}_{\mathrm{MAP}}|\mathcal{D})$$

また，(4.137) の右辺第１項は最適なパラメータを使用したときの対数尤度であり，残りの３つの項はモデルの複雑さにペナルティーを科す「Occam 係数」．
これまでの話は，MacKay の本 [Mac03] に詳しい．

ベイズ情報基準

モデルエビデンスをさらに近似すると，

$$\ln p(\mathcal{D}) \approx p(\mathcal{D}|\boldsymbol{\theta}_{\mathrm{MAP}}) - \frac{1}{2} M \ln N$$

ここで，N はデータ数，M は $\boldsymbol{\theta}$ に含まれるパラメータ数である．右辺に -2 をかけたものが，ベイズ情報量規準 (BIC: Bayesian Information Criterion) である．

$$\mathrm{BIC} = -2 p(\mathcal{D}|\boldsymbol{\theta}_{\mathrm{MAP}}) + M \ln N$$

式 (1.73) の AIC と比較すると，BIC はモデルの複雑さに，より重いペナルティーを科している．

ロジスティック回帰のベイズ的な取り扱い

- ロジスティック回帰にベイズ推論を適用
  - w の事後確率分布は，ラプラス近似で得る
  - 予測分布の計算を解析的に行うため，シグモイド関数をプロビット関数の逆関数 $\Phi(a)$ で置き換える
- 分類決定境界はロジスティック回帰と同じ
  - 与えられた事例 ϕ を C_1 もしくは C_2 に分類するだけであれば，通常のロジスティック回帰と結果は同じ

事後確率分布

w の事前分布としてガウス分布を仮定する．

$$p(w) = \mathcal{N}(w|m_0, S_0)$$

ここで，m_0 と S_0 はある固定のハイパーパラメータである．
w の事後確率分布は，

$$p(w|t) \propto p(w)p(t|w)$$

ここで，$t = (t_1, ..., t_N)^T$ である．両辺の対数をとり，事前分布の仮定と対数尤度の式 (4.89) を用いると，w の事後確率分布は，

$$\ln p(w|t) = -\frac{1}{2}(w - m_0)^T S_0^{-1}(w - m_0) + \sum_{n=1}^{N} \{t_n \ln y_n + (1 - t_n) \ln(1 - y_n)\} + (\text{定数})$$

ここで，$y_n = \sigma(w^T \phi_n)$ である．

ラプラス近似

事後確率分布に対するガウス分布で近似するため，まず事後確率分布を最大化して，MAP（最大事後確率）解 w_{MAP} を求め，ガウス分布の平均とする．
共分散は，負の対数尤度におけるヘッセ行列の逆行列により，次式のように与えられる．

$$S_N^{-1} = -\nabla\nabla \ln p(w|t) = S_0^{-1} + \sum_{n=1}^{N} y_n(1-y_n)\phi_n\phi_n^T$$

最終的に，事後確率分布のガウス分布による近似は，

$$q(w) = \mathcal{N}(w|w_{\mathrm{MAP}}, S_N)$$

予測分布 (1/3)

新たな特徴ベクトル $\phi(x)$ が与えられたとき，クラス C_1 に対する予測分布 $p(C_1|\phi, t)$ を求める．

$$p(C_1|\phi, t) = \int p(C_1|\phi, w)p(w|t)dw \approx \int \sigma(w^T\phi)q(w)dw$$

$a = w^T\phi$ と表すと，$\sigma(w^T\phi)$ は以下のように書ける．

$$\sigma(w^T\phi) = \int \delta(a - w^T\phi)\sigma(a)da$$

ここで，$\delta(x)$ はディラックのデルタ関数である．

$$\int f(x)\delta(x)dx = f(0)$$

予測分布 (2/3)

このことから，クラス C_1 に対する分布予測の右辺は，

$$\int \sigma(w^T\phi)q(w)dw = \int \sigma(a)p(a)da,$$

$$p(a) = \int \delta(a - w^T\phi)q(w)dw$$

ガウス分布である $q(w)$ の周辺分布は，ガウス分布であるから，$p(a)$ もガウス分布．その平均は，

$$\mu_a = \int p(a)ada = \int q(w)w^T\phi dw = w_{\mathrm{MAP}}^T\phi$$

共分散は，

$$\sigma_a^2 = \int p(a)\left\{a^2 - \mu_a^2\right\}da = \phi^T S_N\phi$$

予測分布 (3/3)

最終的に，予測分布は，

$$p(C_1|t) = \int \sigma(a)p(a)da = \int \sigma(a)\mathcal{N}(a|\mu_a, \sigma_a^2)da$$

この積分は，シグモイド関数でのガウス分布の畳み込み積分であり，解析的に評価することが難しい．

プロビット回帰での近似

予測分布の式に出てくるシグモイド関数 $\sigma(a)$ を，プロビット関数の逆関数 $\Phi(\lambda a)$ で近似する．ここで，λ は，$a = 0$ において $\sigma(a)$ と $\Phi(a)$ の傾きをそろえるためのパラメータで，$\lambda^2 = \pi/8$ である．
予測分布の式の畳み込み積分は，次のように表現できる．

$$\int \phi(\lambda a)\mathcal{N}(a|\mu_a, \sigma_a^2)da = \Phi\left(\frac{\mu}{(\lambda^{-2} + \sigma^2)^{1/2}}\right)$$

この式の両辺において，$\Phi(a)$ を $\sigma(a)$ を元に戻すと，

$$\int \sigma(a)\mathcal{N}(a|\mu_a, \sigma_a^2)da \approx \sigma\left(\kappa(\sigma^2)\mu\right),$$

$$\kappa(\sigma^2) = (1 + \pi\sigma^2/8)^{-1/2}$$

最終的に，次のような近似予測分布が得られる．

$$p(C_1|\phi, t) = \sigma\left(\kappa(\sigma_a^2)\mu_a\right)$$

参考文献 I

Su-In Lee, Honglak Lee, Pieter Abbeel, and Andrew Y. Ng, *Efficient l1 regularized logistic regression*, Proceedings of the Twenty-first National Conference on Artificial Intelligence (AAAI-06), 2006, pp. 1–9.

Chih-Jen Lin, Ruby C. Weng, and S. Sathiya Keerthi, *Trust region newton methods for large-scale logistic regression*, ICML '07: Proceedings of the 24th international conference on Machine learning, 2007, pp. 561–568.

David MacKay, *Information theory, inference and learning algorithms*, Cambridge University Press, UK, 2003.