THE UNIVERSITY OF TOKYO

数理情報工学特論第一
【機械学習とデータマイニング】
3章：分類①

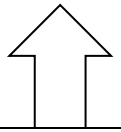
かしま　ひさし
鹿島 久嗣
（数理 6 研）
kashima@mist.i.～

東京大学
THE UNIVERSITY OF TOKYO

DEPARTMENT OF MATHEMATICAL INFORMATICS

「分類」について学びます

- 分類問題の定義
- 分類問題の応用
- 分類のためのモデル：ロジスティック回帰
- 分類問題の定式化
- 学習アルゴリズム：ニュートン法と最急勾配法



- パーセプトロン

分類問題

分類問題とは：
離散的な出力を持つ条件付き確率分布を推定する問題です

- 教師つき学習は、入力 x が与えられた時の出力yの条件付き確率分布$P(y|x)$を、N 個の入出力ペアである訓練データ集合$\{(x^{(i)}, y^{(i)})\}_{i=1}^{N}$をもとに推定する問題
- 出力が実数値 $y \in \mathcal{R}$である場合が「回帰」
- y が離散値、出力の取りうる集合$\mathcal{Y}$が：
  - 2クラス分類：+1 と -1 の2値
  - 多クラス分類：$\{1,2,\ldots,C\}$のC通り

  のような場合が「分類」
- 分類問題における出力を、特別に**クラス**と呼ぶ。

回帰手法でも分類問題を解くことは可能ですが それは必ずしもベストの方法ではありません

- 目標とする出力値を $y^{(i)} \in \{ +1, -1 \}$とすれば、形式的には2クラス分類に回帰手法を適用することは可能
  - 実際は訓練データの出力を $y^{(i)} \in \{+1,-1\}$とするのではなく、出力が+1の訓練データの数を$|N_+|$、出力が-1の訓練データの数を$|N_-|$とすると $y^{(i)} \in \{+1/|N_+|\}, -1/|N_-|\}$としたほうがよい
- 多クラス分類の場合にも、データが C 個のクラスの各々に属するかどうかを{+1, -1}で表し、出力値をベクトル $\mathbf{y}^{(i)} \equiv (y_1^{(i)}, y_2^{(i)}, \ldots, y_C^{(i)})$ で書けば（ベクトル出力に一般化した）回帰手法を適用できる
- しかし、2乗誤差を損失関数とする線形回帰では、確率モデルの仮定として「線形モデルの出力にガウス的なノイズが載る」と仮定
  - これは、出力が{+1,-1}のどちらかであるとする分類モデルの背後にあるべき確率モデルとしては適当ではない

より分類に特化したモデルとして、ロジスティック回帰を中心に紹介します

- ロジスティック回帰：より分類という目的を直接的にモデル化したモデル
- モデル（パラメータ）の学習は回帰のときのように、逆行列1回のようなシンプルではなくなる

分類問題の応用

分類の応用は、結構あります

- 2クラス分類：
  - 購買予測：ある人xが商品を購入する(y= +1)か否(y = -1)か予測
  - 活性予測：ある化合物xが、活性をもつ(y = +1)か否(y = -1)か予測
  - 与信：ある人xが、融資したお金を返済してくれる(y = +1)か否(y = -1)か予測
- 多クラス分類：
  - テキスト分類：ある文書xが、どのカテゴリに属するか（$y \in$ {政治, 経済, スポーツ, …}）を判別
  - 画像認識：ある画像xに映っているものが何か（$y \in$ {自動車, 家, 飛行機, …} ）を識別
  - 行動識別：ある人に取り付けたセンサーデータxからその人の行動（$y \in$ {走っている, 歩いている, …}） を識別

2クラス分類の複雑なケース：関係の予測

- 2つのデータの間の「関係の有無」を予測するような場合も2クラス分類の特殊なケースとして考えられる
- 2つのデータxとx'の間に関係がある(y = +1)か否(y = -1)かを予測
- たとえば：
  - タンパク質の相互作用予測：2つのタンパク質（データ）の間に物理的な相互作用（チームで働くなど）があるかを予測
  - 購買予測：顧客と商品の間の購買関係を予測
- 入力が2データのペアであるような2クラス分類問題になるので、通常の（1データに対する）2クラス分類問題の一般化になっている
- 関係にも複数種類がある場合には、多クラス分類になる

多値分類の極端なケースとしては、構造データのラベルづけ問題などがあります

- 構造データのラベルづけ問題：自然言語処理における品詞付け問題
  - 文（単語列）x に含まれる単語のそれぞれに対して、その品詞（{名詞, 動詞, 副詞, …}など複数ありうる）を割り当てるタスク
  - 他、固有表現抽出、DNAからの遺伝子発見の問題等色々
- それぞれの単語を独立のものと考えれば、単語の多クラス値分類問題となるが、通常、品詞は互いに依存関係にあり（例えば、名詞の後には動詞が来やすいなど）独立に扱うのは適切ではない
- そこで、文に含まれる単語全てに対する品詞の組み合わせ（1単語目が「名詞」で2単語目が「動詞」など）を1つのクラスとして考えると、クラスの数は $(\text{可能な品詞数})^{(\text{文中の単語数})}$ となり、非常に多くのクラスを持つ（しかも、文長に依存してその数が変化する）ような分類問題になる

分類のためのモデル：ロジスティック回帰

ロジスティック回帰モデル（2クラスの場合）

- 2クラス{+1,-1}の場合のロジスティック回帰モデル：

$$P(y = +1|x; \mathbf{w}) = \sigma(\mathbf{w}^\top \boldsymbol{\phi}(x) + b)$$

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

- 関数 σ は、シグモイド関数（もしくはロジスティック関数）と呼ばれ、実数値を (0,1) の間の値に変換する（＝確率化する）
- σ の中身は、ちょうど線形回帰と同じ形
- 「線形回帰モデルの出力する実数値を確率値に変換している」
- モデルパラメータは線形回帰のときと同じく$(\mathbf{w}, b)$のD+1個
- $\mathbf{w}^\top \boldsymbol{\phi}(x)$ が大きいほど $P(y=+1|x; \mathbf{w})$も大きいので、$\mathbf{w}$の各次元は$\boldsymbol{\phi}(x)$の各次元（が正の値をもつこと）が、クラス+1への所属にどの程度貢献しているかを表す

線形回帰のときと同じくbは無かったことにしてしまいます（簡単にするために）

- 線形回帰の時と同じく b を $\mathbf{w}$ に含めてしまう：

$$\mathbf{w} \equiv (\mathbf{w}^\top, w_{D+1})^\top$$
$$\boldsymbol{\phi}(x) \equiv (\boldsymbol{\phi}(x)^\top, 1)^\top$$

  - 併せて特徴ベクトル$\boldsymbol{\phi}(x)$も変更

- モデルは $P(y = +1|x; \mathbf{w}) = \sigma(\mathbf{w}^\top \boldsymbol{\phi}(x))$ になる
- ロジスティック回帰を用いた予測は、$P(y = +1|\, x; \mathbf{w}) > 0.5$ であれば +1、$P(y = +1|\, x; \mathbf{w}) < 0.5$ であれば -1 と予測すればよい
  - それぞれ $\mathbf{w}^\top \boldsymbol{\phi}(x) > 0$ と $\mathbf{w}^\top \boldsymbol{\phi}(x) < 0$ に相当する
- つまり、D次元の超平面である $\mathbf{w}^\top \boldsymbol{\phi}(x) = 0$を境に、クラス+1と-1が分割されていることになる
  - この超平面のことを**分割超平面**と呼ぶ

分類問題の定式化

我々の目的は、将来のデータに対して、その正解に高い確率（の対数）を与えるモデルを得ることです

- 我々の目的は（一応）条件付き確率分布$P(y|x; \mathbf{w})$を推定すること
- その推定の良さをどのように定義したらよいだろうか？
- 与えられた訓練データ集合$\{(x^{(i)}, y^{(i)})\}_{i=1}^{N}$に対してでなく、出力未知の入力$x^{(N+1)}$に対して、対応する正しい出力$y^{(N+1)}$を出力すること
- いいかえると、入力$x^{(N+1)}$に対する$y^{(N+1)}$の確率（の、なぜか対数）

$$\log P(y^{(N+1)}|\phi(x^{(N+1)}); \mathbf{w})$$

を大きくすることと言える

  - 確率0.1と0.2の違いと、0.8と0.9の違いはどちらも0.1の差であるが我々は前者の差を重く見る（対数をとる＝比で考える）
  - この値の符号を逆転したものは、情報理論における符号化の文脈において、データを符号化するのに必要な記述長を表す。
    - 記述長は短いほうが良い＝効率良く符号化できるほうが良い

我々の（最大化すべき）目的関数は、未知データの対数尤度の、真のデータ分布による期待値です

- $\log P(y^{(N+1)}|\phi(x^{(N+1)}); \mathbf{w})$を大きくするのが目的とするとして、我々は、次にどのような$(x^{(N+1)}, y^{(N+1)})$が来るかは知らないので、これを直接大きくすることはできそうにない
  - 訓練データ集合$\{(x^{(i)}, y^{(i)})\}_{i=1}^{N}$を使うしかない
- $(x^{(N+1)}, y^{(N+1)})$を生み出している確率分布$Q(x^{(N+1)}, y^{(N+1)})$を考えてみる
- 我々が最大化したいのは、$\log P(y^{(N+1)}|\phi(x^{(N+1)}); \mathbf{w})$ の $Q(x^{(N+1)}, y^{(N+1)})$ についての期待値であるといえる

$$E_{Q(x^{(N+1)},y^{(N+1)})}\left[\log P(y^{(N+1)}|\phi(x^{(N+1)}); \mathbf{w})\right]$$
$$\equiv \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} Q(x^{(N+1)}, y^{(N+1)}) \log P(y^{(N+1)}|\phi(x^{(N+1)}); \mathbf{w})$$

未知データの対数尤度の、真のデータ分布による期待値は、訓練データの対数尤度で近似されます

- 訓練データおよび将来のデータ$\{(x^{(i)}, y^{(i)})\}_{i=1}^{N+1}$は、すべて同一の分布 Q から独立に取り出されたものとすると
  - 大数の法則：独立なサンプルの平均は、サンプル数を大きくすると、その期待値に近づく

  から

$$\frac{1}{N}\sum_{i=1}^{N} \log P(y^{(i)}|\boldsymbol{\phi}(x^{(i)}); \mathbf{w})$$
$$\xrightarrow{n\to\infty} E_{Q(x^{(N+1)},y^{(N+1)})}\left[\log P(y^{(N+1)}|\boldsymbol{\phi}(x^{(N+1)}); \mathbf{w})\right]$$

- 期待値の代わりに対数尤度の和で代用し、これを最大化するようにパラメータを決定する ＝ つまり最尤推定

なお、事後確率最大化（MAP推定）からは L_2正則化項が出てくるのでした

- 最尤推定によってパラメータを決定すれば：

$$\mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{argmax}} \frac{1}{N} \sum_{i=1}^{N} \log P(y^{(i)} | \boldsymbol{\phi}(x^{(i)}); \mathbf{w})$$

- 事後確率最大化（MAP）推定では：

$$\mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{argmax}} \frac{1}{N} \sum_{i=1}^{N} \log P(y^{(i)} | \boldsymbol{\phi}(x^{(i)}); \mathbf{w}) + \log P(\mathbf{w})$$

- 事前分布が正規分布である場合には：

$$\mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{argmax}} \frac{1}{N} \sum_{i=1}^{N} \log P(y^{(i)} | \boldsymbol{\phi}(x^{(i)}); \mathbf{w}) - \lambda \| \mathbf{w} \|_2^2$$

  - L_2正則化項が出てくる
  - 以降はMAP推定／正則化を前提として話をすすめる

学習アルゴリズム：ニュートン法と最急勾配法

分類問題の最適解は閉じた形で求まらないことが多いので パラメータを少しづつ改善する方法をとります

- MAP推定の目的関数を最大化する方法を考える

$$L(\mathbf{w}) \equiv \frac{1}{N}\sum_{i=1}^{N} \log P(y^{(i)}|\boldsymbol{\phi}(x^{(i)});\mathbf{w}) - \lambda \parallel \mathbf{w} \parallel_2^2$$

  - これが最大になるような $\mathbf{w}$ を求めたい

- 回帰の場合とは異なり、これを $\mathbf{w}$で微分し $\mathbf{0}$ と置いても、連立方程式のような簡単な形（閉じた形の解）は得られない

- そこで、$\mathbf{w}$を少しづつ改善していくステップを繰り返す

  - 現時点でのパラメータを $\mathbf{w}^{(t)}$ とすると、これを、目的関数が改善するような $\mathbf{w}^{(t+1)}$ に更新する

$$\mathbf{w}^{(t+1)} \leftarrow \mathbf{w}^{(t)} + \mathbf{d}^{(t)}$$

更新の差分$\mathbf{d}^{(t)}$について目的関数を最大化します

- 更新の差分$\mathbf{d}^{(t)}$はどのように求めたらよいであろうか？

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} + \mathbf{d}^{(t)}$$

- 我々が実現したいことは $L(\mathbf{w}^{(t+1)})$ すなわち

$$L(\mathbf{w}^{(t)} + \mathbf{d}^{(t)}) = \frac{1}{N} \sum_{i=1}^{N} \log P(y^{(i)} | \phi(x^{(i)}); \mathbf{w}^{(t)} + \mathbf{d}^{(t)}) - \lambda \| \mathbf{w}^{(t)} + \mathbf{d}^{(t)} \|_2^2$$

を$\mathbf{d}^{(t)}$について最大化すること、つまり、

$$\mathbf{d}^{(t)} = \mathrm{argmax}_{\mathbf{d}} \quad L(\mathbf{w}^{(t+1)} + \mathbf{d}^{(t)})$$

を解くことである

目的関数を現在のパラメータの周りで展開してみます

- $L(\mathbf{w}^{(t+1)} + \mathbf{d}^{(t)})$ はそのままでは扱いづらいので、
  これを現在のパラメータ$\mathbf{w}^{(t)}$の周りでテーラー展開してみると：

$$L(\mathbf{w}^{(t)} + \mathbf{d}^{(t+1)}) = L(\mathbf{w}^{(t+1)}) + \mathbf{d}^{\top}\nabla(\mathbf{w}^{(t)}) + \frac{1}{2}\mathbf{d}^{\top}\mathbf{H}(\mathbf{w}^{(t)})\mathbf{d} + O(\mathbf{d}^3)$$

  - $\nabla(\mathbf{w}^{(t)})$ は $L(\mathbf{w}^{(t+1)})$ の $\mathbf{w} = \mathbf{w}^{(t)}$ における勾配ベクトル：

$$\nabla(\mathbf{w}^{(t)}) \equiv \left(\left.\frac{\partial L(\mathbf{w})}{\partial w_1}\right|_{w=w_1^{(t)}}, \left.\frac{\partial L(\mathbf{w})}{\partial w_2}\right|_{w=w_2^{(t)}}, \cdots, \left.\frac{\partial L(\mathbf{w})}{\partial w_D}\right|_{w=w_D^{(t)}} \right)^{\top}$$

  - $\mathbf{H}(\mathbf{w}^{(t+1)})$はヘッセ行列とよばれる $D \times D$ 行列

$$[\mathbf{H}(\mathbf{w}^{(t)})]_{i,j} \equiv \left.\frac{\partial^2 L(\mathbf{w})}{\partial w_i \partial w_i}\right|_{w_i=w_i^{(t)}, w_j=w_j^{(t)}}$$

  - 最後の$O(\mathbf{d}^3)$は3次以上の項を表すとする

- 勾配は目的関数$L(\mathbf{w})$ の$\mathbf{w}$=$\mathbf{w}^{(t)}$の周りでの傾き具合を、ヘッセ行列はさらにその「曲がり具合」を表す

高次の項を打ち切って目的関数を近似し、最大化すると、ニュートン法の更新式が得られます

- 3次以降の項を省略して目的関数の近似をつくる：

$$L(\mathbf{w}^{(t)} + \mathbf{d}^{(t)})$$

$$\approx \tilde{L}(\mathbf{w}^{(t)} + \mathbf{d}^{(t)}) \equiv L(\mathbf{w}^{(t+1)}) + \mathbf{d}^{\top}\boldsymbol{\nabla}(\mathbf{w}^{(t)}) + \frac{1}{2}\mathbf{d}^{\top}\mathbf{H}(\mathbf{w}^{(t)})\mathbf{d}$$

- これの偏微分を計算してみると：

$$\frac{\partial \tilde{L}(\mathbf{w}^{(t)} + \mathbf{d}^{(t)})}{\partial \mathbf{d}^{(t)}} = \boldsymbol{\nabla}(\mathbf{w}^{(t)}) + \mathbf{H}(\mathbf{w}^{(t)})\mathbf{d}^{(t)}$$

- これを**0**と置いて解けば、解は $\mathbf{d}^{(t)} = -\mathbf{H}(\mathbf{w}^{(t)})^{-1}\boldsymbol{\nabla}(\mathbf{w}^{(t)})$

  のように求まるので、更新式は：

$$\mathbf{w}^{(t+1)} \leftarrow \mathbf{w}^{(t)} - \mathbf{H}(\mathbf{w}^{(t)})^{-1}\boldsymbol{\nabla}(\mathbf{w}^{(t)})$$

- この更新式によってパラメータ更新を行う最適化法を**ニュートン法**と呼ぶ

近似は必ずしも正しくないので、パラメータ更新の方向だけを採用し、更新量は学習率として別途決めることにします

- もしも近似式が正しければ、$\mathbf{w}^{(t+1)}$ は目的関数を最大化する解となる
  - そもそもこの近似が正しければ、更新を繰り返す必要は無い（= 2次関数）
- しかし、実際にはこの近似式は「近似」であるので、ここで求まる $\mathbf{w}^{(t+1)}$ は、あくまで目的関数を2次近似したときの最適解である。
- そこで、パラメータの更新分 $\mathbf{d}^{(t)} = -\mathbf{H}(\mathbf{w}^{(t)})^{-1}\nabla(\mathbf{w}^{(t)})$ の方向だけを採用し、その方向にパラメータをどれだけ動かすかは別途決めるという考え方もできる
- 更新のステップ幅を決定する正の定数 $\eta^{(t)}$ を導入すると：

$$\mathbf{w}^{(t+1)} \leftarrow \mathbf{w}^{(t)} - \eta^{(t)}\mathbf{H}(\mathbf{w}^{(t)})^{-1}\nabla(\mathbf{w}^{(t)})$$

  - $\eta^{(t)}$ は**学習率**などと呼ばれる

学習率 $\eta^{(t)}$ の決定方法：線形探索などを行います

- $\eta^{(t)}$ の簡単な決定方法：
  - 十分に小さい定数に取る
  - $\eta^{(t)} = 1/t$などとする（ステップ幅が次第に小さくなっていく）
- もう少しきちんと決めたい場合には $\eta^{(t)}$の線形探索を行う
  つまり、1変数最適化問題を解く

$$\eta^{(t)} = \underset{\eta \geq 0}{\operatorname{argmax}} \; L(\mathbf{w}^{(t)} + \eta \mathbf{d}^{(t)})$$

- $\eta^{(t)}$ の簡便な探索方法としては、η を適当な初期値から初めて：
  1. もし $L(\mathbf{w}^{(t)}+\eta\mathbf{d}) > L(\mathbf{w}^{(t)})$ となるならば、その η を $\eta^{(t)}$ として採用
  2. そうでないならば η を1/2倍してステップ1へ

最急勾配法：もっとも目的関数が増加する方向にパラメータを更新する簡便な方法です

- （線形探索付きの）ニュートン法の更新式

$$\mathbf{w}^{(t+1)} \leftarrow \mathbf{w}^{(t)} - \eta^{(t)} \mathbf{H}(\mathbf{w}^{(t)})^{-1} \nabla(\mathbf{w}^{(t)})$$

  を実行するためには、以下の2つの計算が必要：

  - 現在のパラメータ$\mathbf{w}^{(t)}$における勾配ベクトル$\nabla(\mathbf{w}^{(t)})$
    - これは比較的易しい（少ない計算量）
  - ヘッセ行列の逆行列 $\mathbf{H}(\mathbf{w}^{(t)})^{-1}$
    - これをパラメータ更新の度に計算することは大変

- そこで、単純な近似として、ヘッセ行列の逆行列を$\mathbf{H}(\mathbf{w}^{(t)})^{-1} \equiv -\mathbf{I}$ と置いてしまうことにすると、更新式は：

$$\mathbf{w}^{(t+1)} \leftarrow \mathbf{w}^{(t)} + \eta^{(t)} \nabla(\mathbf{w}^{(t)})$$

- 目的関数が最も大きい方へ変化する方向$\nabla(\mathbf{w}^{(t)})$ に向かってパラメータを更新するので **最急勾配法**と呼ぶ

パラメータの更新式における勾配$\nabla(\mathbf{w}^{(t)})$を具体的に計算してみます

- 目的関数を $\mathbf{w}$ で偏微分すると、

$$\frac{\partial L(\mathbf{w})}{\partial \mathbf{w}} = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{P(y^{(i)}|x^{(i)};\mathbf{w})} \frac{\partial P(y^{(i)}|x^{(i)};\mathbf{w})}{\partial \mathbf{w}} - 2\lambda\mathbf{w}$$

- モデルとしてロジスティック回帰を用いることにする、
  つまり $P(y^{(N+1)} = +1|\, \boldsymbol{\phi}(x^{(N+1)});\mathbf{w}) \equiv \sigma(\mathbf{w}^\top \boldsymbol{\phi}(x^{(i)}))$とする

- 以下の2つの事実：
  - $$\frac{\partial P(y^{(i)}|x^{(i)};\mathbf{w})}{\partial \mathbf{w}} = \frac{\partial P(y^{(i)}|x^{(i)};\mathbf{w})}{\partial z}\frac{\partial z}{\partial \mathbf{w}}$$
  - シグモイド関数の微分：$\partial\sigma(z)/\partial z = 1 - \sigma(z)$

より偏微分が計算できる：

$$\frac{\partial L(\mathbf{w})}{\partial \mathbf{w}} = \frac{1}{N}\sum_{i=1}^{N} \left(\frac{P(y^{(i)} = -1|x^{(i)};\mathbf{w})}{P(y^{(i)} = +1|x^{(i)};\mathbf{w})}\right)^{y(i)} \boldsymbol{\phi}(x^{(i)}) - 2\lambda\mathbf{w}$$

若干複雑になりますが、ヘッセ行列も同様に求まります

- 勾配$\nabla(\mathbf{w}^{(t)})$ をもう一度微分して、ヘッセ行列を求めておくと：

$$H(\mathbf{w}) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{P(y^{(i)}|x^{(i)};\mathbf{w})}\mathbf{H}_P^{(i)}(\mathbf{w}) - \frac{1}{P(y^{(i)}|x^{(i)};\mathbf{w})^2}\frac{\partial P(y^{(i)}|x^{(i)};\mathbf{w})}{\partial \mathbf{w}}\left(\frac{\partial P(y^{(i)}|x^{(i)};\mathbf{w})}{\partial \mathbf{w}}\right)^{\top}\right) - 2\lambda\mathbf{I}$$

- なお、$\mathbf{H}_P(\mathbf{w})$ は、$P(y^{(N+1)}|\phi(x^{(N+1)});\mathbf{w})$に対するヘッセ行列
  - $\mathbf{H}_P(\mathbf{w})$ の (k, l)要素は：

$$[\mathbf{H}_P^{(i)}(\mathbf{w})]_{k,\ell} \equiv \frac{\partial^2 P(y^{(i)}|x^{(i)};\mathbf{w})}{\partial w_k \partial w_\ell}$$