Machine Learning Excercise

情報科学類　佐久間 淳

April 2019

演習の解答が必要な方は佐久間に直接メールしてください. 解答を差し上げることができない場合もあります.

1　第一回

1.1　最急降下法で単回帰の平均二乗誤差最小化

1. 平均二乗誤差 E(w) のパラメータ w, b での微分を求めなさい
2. 最急降下法のステップ 2 の更新式を求めなさい. 初期パラメータを (w^0, b^0), t 回目の更新パラメータを (w^t, b^t), ステップサイズパラメータを η とすること.
3. η は適度に調整する必要がある。η が大きすぎるとどのようなことが起こるか。η が小さすぎるとどのようなことが起こるか。

1.2　Adult データ

Adult データの属性は以下の通り. a) 年齢、b)workclass (民間、自営業、連邦政府、州政府、無職, etc), c) 最終学歴 (博士、修士、学士、高校卒業,etc), d) 教育年数、e) 婚姻状態 (文民と婚姻、軍人と婚姻、離婚、未婚、etc.)、f) 職種 (技術補佐、農業漁業、セールス、etc.)、g) 家族、h) 人種、i) 性別、j) 資本利得、k) 資本損失、l) 週あたり労働時間、m) 出身国

1. a-m のそれぞれの属性は、数値属性、順序属性、カテゴリカル属性のどれに相当するか.
2. これらのデータからその人物の年収が 5 万ドル以上か以下かを予測するとした時に、1-of-k 表現した方が良いと考えられる属性はどれか

1.3　不動産価格の予測

あなたはつくば市の不動産業者です。住居 (家屋, マンション) を仕入れ、転売するのが仕事です。住居がいくらで買ってもらえるかは、様々な要因で決まります。地価予測に使えそうな指標 (=特徴) を多数あげてください。特徴は、「観測可能な値」でなければなりません。例えば、便利な場所、は観測できないので NG です。最寄り駅からの距離 (m) は観測可能なので OK です。

1. 「不動産の値段の予測に使えそうな特徴」を 5 つ記入してください (上記以外)
2. あげられた特徴は、数値属性、順序属性、カテゴリカル属性か、分類しなさい

1.4 wine データの回帰

winequalityred.csv を読み込み、wine_linearRegression_scaling.ipynb を実行して以下の問いに答えなさい。

1. quality = 0.100 * fixed acidity + 5.63 と表示される
   (a) 特徴量はどれか
   (b) 目標値はどれか
   (c) モデルパラメータはどれか
   (d) バイアスはどれか
2. 最も良い予測を与えるモデルに使われている特徴はどれか
3. Quality の予測に最も正の影響が強い特徴はどれか
4. Quality の予測に最も負の影響が強い特徴はどれか
5. なぜどのモデルもバイアスは 5.636 なのか
6. スケーリングされていない特徴について線形モデルを学習した場合、問題 3 や問題 4 のような分析をしてはならない。なぜか。

1.5 行列計算

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}, \boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \boldsymbol{A} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix} \tag{1}$$

とする。

1. $\boldsymbol{x}^T\boldsymbol{y} = \boldsymbol{y}^T\boldsymbol{x} = x_1y_1 + \ldots + x_ny_n$ を示せ。
2. $\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}$ を成分 $(x_i, a_{i,j})$ で表せ。
3. $\frac{\partial}{\partial \boldsymbol{x}}\boldsymbol{a}^T\boldsymbol{x} = \boldsymbol{a}$ を示せ。
4. $\frac{\partial}{\partial \boldsymbol{x}}(\boldsymbol{a}^T\boldsymbol{x} + b)^2$ を求めよ。

2 第二回

2.1 凸関数の最適化

$f(\boldsymbol{x}) = 2x_1^2 + x_1x_2 + x_2^2 - 5x_1 - 3x_2 + 4$ とする。$f(\boldsymbol{x})$ が凸関数であることは既知とする。

1. f の勾配 ∇f を求めよ
2. $(0,0), (1,2), (1,0.5), (1,1)$ における f の勾配を求めよ
3. f を最小にする $\boldsymbol{x}$ とその時の $f(\boldsymbol{x})$ を求めよ

2.2 線形回帰の導出

$\boldsymbol{x}_i = \begin{pmatrix} 1 \\ x_{i1} \\ \vdots \\ x_{iD} \end{pmatrix}, \boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1^T \\ \boldsymbol{x}_2^T \\ \vdots \\ \boldsymbol{x}_N^T \end{pmatrix}, \boldsymbol{t} = \begin{pmatrix} t_1 \\ t_2 \\ \vdots \\ t_N \end{pmatrix}, \boldsymbol{w} = \begin{pmatrix} w_0 \\ w_1 \\ \vdots \\ w_D \end{pmatrix}$ とする。事例群 $\{(\boldsymbol{x}_i, t_i)\}_{i=1}^N$ を使って線形回帰モデル $t = \boldsymbol{w}^T\boldsymbol{x}$ を求めることを考える。

1. 二乗誤差和 E を $\boldsymbol{w}$ の関数で表せ
2. 二乗誤差和 E の $\boldsymbol{w}$ についての勾配 $\nabla_{\boldsymbol{w}} E$ を求めるために、以下を導出せよ
   (a) $\sum_{i=1}^N t_i\boldsymbol{x}_i = \boldsymbol{X}^T\boldsymbol{t}$
   (b) $\sum_{i=1}^N \boldsymbol{x}_i\boldsymbol{w}^T\boldsymbol{x}_i = \boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w}$
3. 勾配 $\frac{\partial E}{\partial \boldsymbol{w}}$ を $\boldsymbol{x}_i$(あるいは $\boldsymbol{X}$), $\boldsymbol{t}$ の式で示せ。
4. (近似解法) 線形回帰モデルを最急降下法で求めるときの、パラメータの更新式を $\boldsymbol{x}_i$(あるいは $\boldsymbol{X}$), $\boldsymbol{t}$ の式で示せ。初期解を $\boldsymbol{w}^0$, t 回目の更新時の回を $\boldsymbol{w}^t$, ステップサイズパラメータを η とする。
5. (解析解) $\frac{\partial E}{\partial \boldsymbol{w}} = 0$ なる $\boldsymbol{w}$ が $\boldsymbol{w} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{t}$ であることを示せ。

3　第三回

3.1　多項式単回帰の二乗誤差

polynomialRegSquaredError.ipynb を実行して以下の問いに答えなさい。

$y = f(x) + N(0,1)$ に従い $i = 1, \ldots, 6$ の 6 点のデータを生成した。ここで, $f(x) = \sin(x)$ であり, $N(0,1)$ は平均 0 分散 1 の正規分布から生成した正規乱数である。$i = 1, ..., 6$ のサンプル (y_i, x_i) から多項式回帰によって関数 f を推定することを考える。

1. プログラムは、生成された多項式特徴量の値を表示している。多項式特徴量を求める式を示せ
2. 最小の二乗誤差を与える多項式の次数はいくつか
3. 二乗誤差が次数 d=5 でほぼゼロになるのはなぜか？

3.2　多項式単回帰の訓練誤差とテスト誤差

polynomialReg_testErr.ipynb を実行して以下の問いに答えなさい。$y = \cos(1.5\pi x) + 0.1 * N(0,1)$ に従い 30 の訓練事例を生成した。ここで $N(0,1)$ は平均 0 分散 1 の正規分布から生成した正規乱数である。プログラムの実行結果は、これらの訓練事例について、次数 degree を 1 から 20 まで変化させながら、多項式単回帰を行い、訓練誤差 TrainErr. およびテスト誤差 TestErr（とその標準偏差）を表示している。

1. 訓練誤差を最小にする多項式特徴量は何次元のものか
2. テスト誤差を最小にする多項式特徴量は何次元のものか
3. 訓練誤差は小さいが、テスト誤差が大きくなっている状態をなんと呼ぶか
4. もっとも優れた多項式特徴量は何次元か

3.3　バイト君とコンビニ店長

あるコンビニエンスストアの店長は、ビールの売り上げ本数 $y_i, (i = 1, ..., T)$ が毎日変化するため、これを予測し、仕入れの数量を調整したいと考えた。予測のために、i 日の日付 $x_{i1}(1, 2, \ldots, 30)$, 気温 x_{i2}(度)、湿度 x_{i3}(%)、と, その日に売れたビールの本数 y_i を記録した.　$\boldsymbol{x}_i^T = (x_{i1}, x_{i2}, x_{i3})$ と定義する. $(\boldsymbol{x}_i^T, t_i) = (15, 21, 35, 60)$ は、ある月の 15 日、気温 21 度、湿度 35%の時、ビールは 60 本売れたということである。

店長 「バイト君は大学で機械学習の授業取ってるんだって？これまでのデータはあるから毎日のビールの売れ行きを予測してみてくれないか？」

バイト君　「余裕っすよ．これまでのデータ$_1$ をもらえますか?多項式重回帰$_2$ で予測モデルを作成して, 過去のデータでまずどれだけ予測が当たるか$_3$ 評価してみます.」

店長 「結果はどうだい？」

バイト君 「すごいっすよ. 過去のどの日の予測も, ほとんど誤差なく$_4$ 売れ行きを当てることができています. 機械学習すごいっす.」

(店長心の声) 商売はそんなに甘いものではない. 誤差なく当たることなどあるのだろうか. しかしバイト君は機械学習を勉強したといっているし...

店長 「そうか. では明後日 24 日の天気予報では，気温 22 度，湿度 65%だそうだから，明後日のビールの売り上げをさっそく予測してもらおうか.」

バイト君 「多項式特徴量を使った回帰による予測だと, 24 日はバカ売れっす! 10 ケース仕入れても売り切れです!」

しかし 24 日は給料日前日なうえ，急激な冷え込みのため，ビールは全く売れなかった$_5$. 店長は在庫を抱え，バイト君はクビになりそうである.

1. 下線 1 の事例を機械学習では何と呼ぶか
2. 下線 2 において, 4 次の多項式特徴量を使った重回帰モデル $f(\boldsymbol{x})$ を式で示せ
3. 下線 3 で評価した量を何と呼ぶか. $i = 1, \ldots, N$ として、その評価量を式で表せ
4. 下線 4 の現象をなんと呼ぶか
5. 下線 5 の現象が起きた理由について、考えられることを全てあげよ
6. 予測性能を向上させるために考えられるモデリング上の工夫について以下の語句を使って考察しなさい. [テスト誤差, 交差検証].
7. 予測性能を向上させるために考えられる特徴量の設計上の工夫について考えられることを書きなさい

3.4 多項式単回帰の正則化

polynomialRidgeRegression.py を実行して以下の問いに答えなさい。

$y = \cos(1.5\pi x) + 0.1 * N(0, 1)$ に従い、30 の訓練事例を生成した。ここで $N(0, 1)$ は平均 0 分散 1 の正規分布から生成した正規乱数である。プログラムは、正則化係数を [1e-30,1e-20,1e-10,1e-5, 1e-3,1e-2,1e-1, 1,10,100] の範囲で変化させながら、これらの訓練事例を用いて、degree=30 の多項式回帰モデルによる学習を行っている。その出力は訓練誤差 TrainErr. およびテスト誤差 TestErr（とその標準偏差）を表示している。

1. $d = 30$ の多項式特徴量を用いた時に、もっとも小さい訓練誤差を与える正則化パラメータはいくらか

2. d=30 の多項式特徴量を用いた時に、もっとも小さいテスト誤差を与える正則化パラメータはいくらか

3. もっとも適当な正則化パラメータはいくらか

4. 正則化パラメータの選択過程において、k-fold 交差検証を使うメリットはあるか？あるとすればどのようなメリットがあるか。

4 第四回

4.1 リッジ回帰とラッソの比較

wine_ridgeRegression.py と wine_lasso.py を実行して以下の問いに答えなさい。Wine データを訓練事例とテスト事例に分割し、訓練事例で ridge 回帰と LASSO を学習させた。正則化パラメータ (alpha) は $2^{-16}, 2^{-15}, ..., 2^{12}$ まで変化させた。プログラムの実行結果は、正則化パラメータ、各特徴量とそれに対応する係数 (昇順にソート)、訓練誤差およびテスト誤差を表している。

1. リッジ回帰について、訓練誤差とテスト誤差を最小にする正則化パラメータはいくつか
2. ラッソについて、訓練誤差とテスト誤差を最小にする正則化パラメータはいくつか
3. リッジ回帰と LASSO について、それぞれ適切な正則化パラメータを選択せよ
4. 学習結果として得られたモデルと正則化パラメータ、得られたテスト誤差の観点から、リッジ回帰と LASSO の違いを考察しなさい

4.2 リッジ回帰とラッソの正則化パス

wine_ridgeRegression_solutionPath.py と wine_LASSO_solutionPath.py を実行して以下の問いに答えなさい。Wine データを訓練事例とテスト事例に分割し、 10^6 から 10^{-3} まで正則化パラメータを変化させつつ、訓練事例で ridge 回帰と LASSO を学習させた。プログラムの実行結果は、その正則化パラメータの変化 (正則化パス) を表している。

1. リッジ回帰と LASSO の正則化パスの挙動の最大に違いは何か。その違いがなぜ発生するか説明しなさい

2. 正則化パラメータを大きくした時に、その特徴の予測に与える影響 (係数の絶対値大きさ) が (多くの場合) 減少するのはなぜか
3. 正則化パラメータを大きくした時に、その特徴の予測に与える影響 (係数の絶対値の大きさ) が増加することがあるのはなぜか

4.3 最急降下法と確率的最急降下法

wine_GD.ipynb と wine_SGD.ipynb を実行して以下の問いに答えなさい。プログラムは Wine データを訓練事例とテスト事例に分割し、訓練事例で ridge 回帰を勾配降下法 (GD) と確率的勾配降下法 (SGD) によって最適化した。 いずれも複数のステップサイズパラメータ η で実行した。

1. GD ではステップサイズにかかわらすアルゴリズム停止時にはほぼ同じ誤差を達成しているが、ステップサイズによって停止までにかかるステップ数が異なる。それはなぜだと考えられるか。

2. GD では目的関数が単調減少するのに対して、SGD ではそうならないのはなぜか

3. GD は t 回目の更新における誤差と、$t+1$ 回目の更新における誤差の差分が ϵ より小さくなったときに収束したと判定し、アルゴリズムを停止するように設定されているが、SGD では GD と同じ収束判定法は使えない。それはなぜか。SGD ではどのような条件で収束判定を使うと良いと考えられるか。アイディアを考えて説明しなさい。

4.4 最急降下法と確率的最急降下法

$\boldsymbol{x}_i = \begin{pmatrix} 1 \\ x_{i1} \\ \vdots \\ x_{iD} \end{pmatrix}, \boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1^T \\ \boldsymbol{x}_2^T \\ \vdots \\ \boldsymbol{x}_N^T \end{pmatrix}, \boldsymbol{t} = \begin{pmatrix} t_1 \\ t_2 \\ \vdots \\ t_N \end{pmatrix}, \boldsymbol{w} = \begin{pmatrix} w_0 \\ w_1 \\ \vdots \\ w_D \end{pmatrix}$ とする。事例群 $\{(\boldsymbol{x}_i, t_i)\}_{i=1}^N$ を使ってリッジ回帰による線形モデルを求めることを考える。

1. リッジ回帰の最急降下法の更新式を示せ

2. リッジ回帰の確率勾配降下法の更新式を示せ

5 第五回

5.1 マージン

超平面 $\boldsymbol{w}^T\boldsymbol{x} = 0$ と点 $\boldsymbol{x}_i$ の距離を求めなさい。(ヒント：点 $\boldsymbol{x}_i$ と超平面 $\boldsymbol{w}^T\boldsymbol{x} = 0$ の距離を d とおき, 点 $\boldsymbol{x}_i$ から超平面 $\boldsymbol{w}^T\boldsymbol{x} = 0$ への垂線を考える)

5.2 確率的劣勾配降下法による SVM の最適化

$\boldsymbol{x}_i = \begin{pmatrix} 1 \\ x_{i1} \\ \vdots \\ x_{iD} \end{pmatrix}, \boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1^T \\ \boldsymbol{x}_2^T \\ \vdots \\ \boldsymbol{x}_N^T \end{pmatrix}, \boldsymbol{t} = \begin{pmatrix} t_1 \\ t_2 \\ \vdots \\ t_N \end{pmatrix}, \boldsymbol{w} = \begin{pmatrix} w_0 \\ w_1 \\ \vdots \\ w_D \end{pmatrix}$ とする。 $t_i \in \{-1, 1\}$ である。事例群 $\{(\boldsymbol{x}_i, t_i)\}_{i=1}^N$ を使って SVM による線形分類モデルを確率的劣勾配降下法を使って求めることを考える。目的関数は以下の通り。

$$E(\boldsymbol{w}) = \frac{1}{C}\|\boldsymbol{w}\|_2^2 + \sum_{i=1}^{N} \max\{0, 1 - t_i(\boldsymbol{w}^T\boldsymbol{x}_i)\} \tag{2}$$

目的関数は損失関数項 (hinge 損失) と正則化項 (L2 正則化) からなる。確率的劣勾配降下法では、パラメータを、一つの事例に対する損失項と正則化項の劣勾配を用いて更新する.

1. 一つの事例に対する損失 $\ell_i(\boldsymbol{w}) = \max\{0, 1 - t_i(\boldsymbol{w}^T\boldsymbol{x}_i)\}$ の劣勾配を求めなさい

2. 正則化項の勾配を求めなさい

3. 確率的劣勾配降下法による SVM の最適化における更新式を導出しなさい。ただし、ステップサイズパラメータを η とする。

5.3 ヒンジ損失と二乗損失

hinge_and_squared_loss.ipynb と hinge_and_squared_loss_withOutlier.ipynb を実行して以下の問いに答えなさい。この二つのプログラムは、ランダムに生成された 40 点の 2 クラス分類問題のための訓練事例について、ヒンジ損失と二乗損失で線形分類モデルを学習し、その決定境界を表示している。前者は外れ値を含まない訓練事例、後者は外れ値を含む訓練事例で学習させた結果である。両プログラムとも、一つ目のセクションではヒンジ損失を用いて、二つ目のセクションでは二乗損失を用いて分類のための超平面を学習させている。以下の問いに答えよ。

1. 外れ値を含まない場合の、ヒンジ損失と二乗損失で学習させた場合の線形モデルの違いについて（違いがあれば）説明せよ

2. 外れ値を含む場合の、ヒンジ損失と二乗損失で学習させた場合の線形モデルの違いについて（違いがあれば）説明せよ

3. 問題 1 および問題 2 において、ヒンジ損失と二乗損失で違いが生じた場合、なぜその違いが生じたのか説明せよ

5.4 ROC 曲線の作成と正則化パラメータのチューニング

IRIS データを SVM で分類するプログラム svm_iris_roc_visualization_auc_tuning.ipynb を実行して以下の問いに答えなさい。ここでは, 与えられたあやめの特徴から、それが virginica か (true)、そうでないか (false) を予測するモデルを作成している。損失関数にはヒンジ損失を用いて、C=0.01 と固定する (C はここでは $L2$ 正則化パラメータの逆数)。切片を変化させながら、ROC 曲線を作成し、AUC を評価する。以下の問いに答えよ。

1. この分類モデルにおける false positive とはどのような状況か

2. virginica のデータを分類モデルで予測した時に、実際に virginica と予測する確率をなんと呼ぶか

3. セクション ”Next, we will learn a SVM classifier, and visualize the ROC curve ”のコードを実行して以下の問いに答えよ。このプログラムで学習した分類器において, あやめデータを予測した時に、virginica でないにも関わらず、virginica と予測してしまう確率を 0.2 以下に抑えた時に、virginica のデータを virginica と正しく予測する確率は最大でどの程度か

4. セクション ”Now, we learn how to tune the parameter using the AUC scores” のコードを実行して以下の問いに答えよ。$C = 10^{-1}$ におけるモデルの AUC は 0.945 より大きく、それ以外の C におけるモデルの AUC は 0.945 より小さい。このとき、$C = 10^{-1}$ と設定した場合、どのような false positie rate においても、そのモデルの true positive rate を他のモデルの true positive rate 以上にすることが可能か？理由とともに答えなさい。

6 第六回

6.1 多項式カーネル

交互作用項を含む多項式特徴量と多項式カーネルについて以下の問題に答えなさい。

1. 2 次元ベクトルの 3 次多項式特徴量を考える。また 3 次多項式カーネルを $k(\boldsymbol{x}, \boldsymbol{x}') = (\boldsymbol{x}^T\boldsymbol{x}' + 1)^3$ と定義する。二次多項式特徴量同士の内積は、カーネルを通じて $k(\boldsymbol{x}, \boldsymbol{x}') = \phi(\boldsymbol{x})^T\phi(\boldsymbol{x}')$ と計算できることを示せ。

2. 多項式次元数の増加につれて交互作用項の数は指数的に増加するため, 計算量的な意味で扱いが厄介であるが、予測においては重要な役割を果たすことがある。例えば、spam メール判定ルールとして、x_1:実行可能ファイルが添付ファイルに含まれる (false=0/true=1), x_2:同一メールが K 以上の異なるドメインのメールアドレスに送信されている (false=0/true=1)、という二つの特徴量を用いて、両者が真である場合のみスパムメールと判定する $(x_1 \wedge x_2 = 1)$ によってスパム判定することを考えよう。このような分類は、交互作用項 x_1x_2 を使って自然に表現できる $(x_1x_2 \geq 1)$。この例にならって、交互作用項が分類の予測に役立つ特徴量と分類器の例をあげなさい。

6.2 カーネルと SVM

svm_with_kernels.ipynb を実行して以下の問いに答えなさい。

1. We first create the data では、プログラムは 3 種類のデータを生成します。線形モデルでの分類性能が、最も良いと考えられるデータセットはどれか。最も悪いと考えられるデータセットはどれか。

2. Linear kernel では、特徴量やカーネルを用いない線形モデルの SVM を学習する。ここで、パラメータ C は、SVM の定義で現れたパラメータ C である。この C を大きくすると (あるいは小さくすると) どのような挙動が期待できるか？実際にパラメータを変化させたとき、何が起こったか考察しなさい。

3. Gaussian kernel ではガウシアンカーネル (円形基底カーネル, RBF カーネル) における 3 つのデータの分類性能と決定境界を表示する。パラメータ s は、スライドにおけるガウシアンカーネルのパラメータ σ^2 に相当する。s を変化させた時に、分類器の挙動はどのように変化するか 考察しなさい。なお、このプログラムではパラメータ C は [1,30000] の区間で, 1000 刻みで最適な値（最もテスト誤差を小さくする値）が選ばれている。

6.3 ロジスティックシグモイド関数

ロジスティックシグモイド関数 $\sigma(a) = \frac{1}{1+\exp(-a)}$ について以下の問いに答えよ。

1. $\sigma(-a) = 1 - \sigma(a)$ を示せ.

2. $\frac{d}{da}\sigma(a) = \sigma(a)(1-\sigma(a))$ を示せ.

3. $a \in \mathbb{R}$ のとき, $\sigma(a) \in [0,1]$ を示せ.

4. $\log\left(\frac{P(t=1|\boldsymbol{x})}{1-P(t=1|\boldsymbol{x})}\right) = \boldsymbol{w}^T\boldsymbol{x}$ は $P(t=1|\boldsymbol{x}) = \sigma(\boldsymbol{w}^T\boldsymbol{x})$ と等価であることを示せ.

6.4 ロジスティック回帰による予測

LR_prediction.ipynb を実行して以下の問いに答えなさい。プログラムは 10 点の二次元のラベル付きサンプル (ラベルは 0 or 1) を生成します。生成された特徴とラベルは以下の通り

	1	2	3	4	5	6	7	8	9	10
特徴 x_1	1.5	-0.5	1.0	1.5	0.5	1.5	-0.5	1.0	0.0	0.0
特徴 x_2	-0.5	-1.0	-2.5	-1.0	0.0	-2.0	-0.5	-1.0	-1.0	0.5
ラベル t	1	0	0	1	1	1	0	1	0	0

このデータを超平面 $\boldsymbol{w}^T\boldsymbol{x} = 0$ に基づくロジスティック回帰 $y = \sigma(\boldsymbol{w}^T\boldsymbol{x})$ で分類することを考えます。ここでは三つの超平面 $\boldsymbol{w}_1^T = (6,3,-2)$, $\boldsymbol{w}_2^T = (4.6,1,-2.2)$, $\boldsymbol{w}_3^T = (1,-1,-2)$ を考えます。エクセルやプログラムなどを用いて、以下の問いに答えなさい。この課題については、結果は manaba からダウンロードしたエクセルシートに結果を記入して提出すること。

(参考) プログラムは三つのモデルに基づくロジスティック回帰で予測された $P(t=1|\boldsymbol{x})$ の等高線を表示しています。回答の参考にしてください。

1. $\boldsymbol{w}_i^T\boldsymbol{x}_j \quad (i = 1,2,3, j = 1,\ldots,10)$ を求めよ

2. $\sigma(\boldsymbol{w}_i^T\boldsymbol{x}_j) \quad (i = 1,2,3, j = 1,\ldots,10)$ を求めよ

3. モデル $\sigma(\boldsymbol{w}_i^T\boldsymbol{x})$ による $\boldsymbol{x}_j$ の分類結果 (確率ではなく分類結果) を求めよ

4. $\boldsymbol{w}_1, \boldsymbol{w}_2, \boldsymbol{w}_3$ で実現されるロジスティック回帰モデルの、観測値 $(x_1,\ldots,x_{10})$ に対する正解率を求めよ

6.5 ベルヌーイ分布の尤度

A 君がコインを 7 回投げたところ, その観測値として (表, 裏, 表, 裏, 裏, 裏, 裏) を得た。以下の問いに答えなさい。

1. このコインの出目を $\mu = 0.25$ および $\mu = 0.1$ のベルヌーイ分布としたときの尤度をそれぞれ求めよ. ただしベルヌーイ分布では表に対応する確率変数の値を 1, 裏に対応する確率変数の値を 0 とする。

2. $\mu = 0.25$ のベルヌーイ分布に従うコインと $\mu = 0.1$ のベルヌーイ分布に従うコイン、この観測値においてはどちらがより尤もらしいか？

3. この観測値に対するパラメータ μ におけるベルヌーイ分布の対数尤度を μ の式で表し、最尤推定量を求めよ

6.6　ロジスティック回帰と対数尤度

LR_prediction.ipynb を実行して以下の問いに答えなさい。用いるデータやモデルは 6.4 と同じである。

1. 訓練サンプル $(x_1, \ldots, x_{10})$ について交差エントロピー損失 $E(\boldsymbol{w}_1), E(\boldsymbol{w}_2), E(\boldsymbol{w}_3)$ を求めなさい
2. 最も尤度の高いモデルはどれか
3. $\boldsymbol{w}_1, \boldsymbol{w}_2, \boldsymbol{w}_3$ どれが分類モデルとして適切か, 理由とともに答えよ

7　第七回

7.1　ロジスティック回帰の最急降下法

$\boldsymbol{x}_i = \begin{pmatrix} 1 \\ x_{i1} \\ \vdots \\ x_{iD} \end{pmatrix}, \boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1^T \\ \boldsymbol{x}_2^T \\ \vdots \\ \boldsymbol{x}_N^T \end{pmatrix}, \boldsymbol{t} = \begin{pmatrix} t_1 \\ t_2 \\ \vdots \\ t_N \end{pmatrix}, \boldsymbol{w} = \begin{pmatrix} w_0 \\ w_1 \\ \vdots \\ w_D \end{pmatrix}$ とする。 $t_i \in \{0, 1\}$ である。事例群 $\{(\boldsymbol{x}_i, t_i)\}_{i=1}^N$ を使ってロジスティック回帰モデル $t = \sigma(\boldsymbol{w}^T\boldsymbol{x})$ を最急降下法を使って求めることを考える。目的関数 (交差エントロピー損失) は以下の通り。

$$E(\boldsymbol{w}) = -\log L(\boldsymbol{w}) = -\sum_{n=1}^{N} \{t_n \log \hat{t}_n + (1 - t_n) \log(1 - \hat{t}_n)\} \tag{3}$$

ここで $\hat{t}_n = \sigma(\boldsymbol{w}^T\boldsymbol{x}_n)$ である。

1. 交差エントロピー損失の $\boldsymbol{w}$ についての勾配が以下であることを示せ

$$\nabla E(\boldsymbol{w}) = \sum_{n=1}^{N} (\hat{t}_n - t_n)\boldsymbol{x}_n = \boldsymbol{X}^T(\hat{\boldsymbol{t}} - \boldsymbol{t}) \tag{4}$$

2. ステップサイズパラメータが η であるとき, 最急降下法の更新式を示せ

7.2　ロジスティック回帰のニュートン法

定義と記法は 7.1 節に準じます。

1. 交差エントロピー損失の $\boldsymbol{w}$ についての ヘシアン行列が以下であることを示せ

$$\boldsymbol{H} = \boldsymbol{X}^T\boldsymbol{R}\boldsymbol{X} \tag{5}$$

   ただし, $\boldsymbol{R}$ は $N \times N$ 対角行列で、第 n 対角要素は $\boldsymbol{R}_{nn} = \hat{t}_n(1 - \hat{t}_n)$.

2. ロジスティック回帰のニュートン法によるパラメータ $\boldsymbol{w}$ の更新式を $\boldsymbol{X}$ および $\boldsymbol{t}, \hat{\boldsymbol{t}}, \boldsymbol{R}$ の式で表せ.

7.3　ロジスティック回帰における勾配法とニュートン法の比較

プログラム iris_LR_GD.ipynb および iris_LR_Newton.ipynb を実行して以下の問題に答えよ。IRIS データは分類問題用のデータであり、特徴量はがく片長 (Sepal Length), がく片幅 (Sepal Width), 花びら長 (Petal Length), 花びら幅 (Petal Width) の 4 次元ベクトルである。いずれも連続変数である。目標値はセトナ (setosa)、バーシクル (versicolor)、 バージニカ (virginica) という 3 種類のあやめである。ここでは 2 値分類とするため、virginica か、そうでないか、を予測することとする。

プログラムは Iris データのロジスティック回帰における分類を、勾配法 (GD) とニュートン法 (Newton) で学習し、その学習曲線 (更新回数 vs 目標関数値) を表示する。GD では更新前後のコストの改善が 0.00000001 を下回った時に学習を終了する。GD の一つ目のセクションではステップサイズが 0.0001 の場合に学習曲線を表示している。GD の二つ目のセクションではステップサイズを [0.1, 0.01, 0.008, 0.006, 0.004, 0.003, 0.002, 0.001, 0.0005] と変化させた時の学習曲線を表示している。

1. 勾配法 (GD) において、ステップサイズが 0.003 より大きいとき、目的関数が増加するのはなぜか
2. 勾配法 (GD) において、ステップサイズが 0.002 より小さい場合の、更新回数、実行時間の関係を説明しなさい。ステップサイズをある以上小さくすると、最終的な目的関数値 (final cost) が増加するのはなぜか。
3. 勾配法とニュートン法を、最終的な目的関数値、更新回数、実行時間の観点から比較せよ
4. ニュートン法はステップサイズパラメータを持たないがこのデータにおいて最適化は適切に実行されている。なぜか。
5. 最適化においてニュートン法が適しているケースと最急降下法が適しているケースについて考察しなさい。

7.4 softmax 回帰

プログラム mnist_softmax.ipynb を実行して以下の問題に答えよ。プログラムは Mnist データ (0-9 の手書き文字分類) を読み込み、10 クラス分類の交差エントロピーを最小化による softmax 回帰を学習する。

1. 3 番目のセクションでは、任意のテストデータのインデックスを指定し、その画像と、softmax 回帰への入力の出力を見ることができる。index=6 の画像が, 4 と識別される確率と, 8 と識別される確率をそれぞれ求めよ。ここで u_k は softmax 回帰における $\boldsymbol{w}^T\boldsymbol{x}$ の値を示している。
2. 4 番目のセクションでは、クラスを指定したときに、学習した softmax 回帰において、テストデータの画像とそのデータが指定したクラスとに識別される確率を示している。条件付き確率が閾値 θ を超えた時 ($\Pr(t=5|\boldsymbol{x}) > \theta$), 画像 $\boldsymbol{x}$ をクラス t として認識するものとする。プログラム実行結果の「5」の認識において、精度を 0.9 以上にするもっとも小さい θ はいくらか。小数点以下第 4 位までの範囲で答えよ

7.5 ロジスティック損失とクロスエントロピー損失

定義と記法は 7.1 節に準じる。ただし混乱を防ぐためにラベルが $\{0,1\}$ で与えられる場合 (クロスエントロピー損失) には t_i, ラベルが $\{-1,1\}$ で与えられる場合

(ロジスティック損失) には y_i と表記する。クロスエントロピー損失に基づく目的関数は

$$E_{\text{cross}}(\boldsymbol{w}) = \sum_{i=1}^{N} -t_i \log \hat{t}_i - (1 - t_i) \log(1 - \hat{t}_i) \tag{6}$$

ロジスティック損失に基づく目的関数は

$$E_{\text{logistic}}(\boldsymbol{w}) = \sum_{i=1}^{N} \log(1 + e^{-y_i \boldsymbol{w}^T \boldsymbol{x}_i}) \tag{7}$$

で与えられる。両者を最適化した解は一致することを示せ。

(ヒント) ラベル $t_i = 0$ と $y_i = -1$, $t_i = 1$ と $y_i = 1$ がそれぞれ対応している。

8 第八回

8.1 k-means clustering の更新

以下の 4 つのデータを $k = 2$ で k-means クラスタリングすることを考える。

$$\boldsymbol{x}_1 = \begin{pmatrix} 3 \\ -2 \end{pmatrix}, \boldsymbol{x}_1 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \boldsymbol{x}_2 = \begin{pmatrix} 5 \\ -4 \end{pmatrix}, \boldsymbol{x}_4 = \begin{pmatrix} 2 \\ 2 \end{pmatrix}$$

ただし初期のクラスタ中心は以下で与えるものとする。

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 2 \\ -2 \end{pmatrix}, \boldsymbol{\mu}_2 = \begin{pmatrix} 3 \\ 2 \end{pmatrix}$$

以下の問いに答えよ。

1. 初期状態における J を求めよ
2. 1 回目の更新における r_{nk} を求めよ
3. (1) で求めた r_{nk} に基づき、$\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ を更新し、この時の J を求めよ
4. 2 回目の更新では J が変化せず、アルゴリズムが収束することを示せ

8.2 k-means clustering の収束

$\mathbb{R}^D$ のデータ $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ の k-means クラスタリングを考える。クラスタ数は K, $\boldsymbol{x}_n$ が k 番目のクラスタに割り当てられていることを示す変数を r_{nk} と表記する。k 番目のクラスタの中心点を $\boldsymbol{\mu}_k$ とする。k-means クラスタリングの目的関数を以下のように定義する。

$$J(\mu_k, r_{nk}) = \sum_{k=1}^{K} \sum_{n=1}^{N} r_{nk} \|\boldsymbol{x}_n - \boldsymbol{\mu}_k\|_2^2 \tag{8}$$

$\boldsymbol{\mu}_k$ および r_{nk} を k-means クラスタリングアルゴリズムで更新した時について、以下の問いに答えよ。

1. 時刻 t のステップ 2 において, $J(\boldsymbol{\mu}_k^t, r_{nk}^t) \geq J(\boldsymbol{\mu}_k^t, r_{nk}^{t+1})$ を示せ
2. 時刻 t のステップ 2 において割り当て変数が変更された時, ステップ 3 において $J(\boldsymbol{\mu}_k^t, r_{nk}^{t+1}) > J(\boldsymbol{\mu}_k^{t+1}, r_{nk}^{t+1})$ となることを示せ

8.3 k-means clustering の挙動

プログラム k_means.ipynb を実行して以下の問題に答えよ。プログラムは三つの異なる種類のデータについて $k = 3$ の k-means クラスタリングを実行した結果を表している。同じ色のデータ点は同じクラスタに割り当てられたことを表している。以下の問いに答えよ。

1. 2 番目のデータは 1 番目のデータに比べてクラスタリングが正しく実行できているとは言えない。1 番目のデータと 2 番目のデータの違いは何か。またその違いに着目して、2 番目のデータのクラスタリングが正しく実行されない理由について考察しなさい。

2. 3 番目のデータは 1 番目のデータに比べてクラスタリングが正しく実行できているとは言えない。1 番目のデータと 3 番目のデータの違いは何か。またその違いに着目して、3 番目のデータのクラスタリングが正しく実行されない理由について考察しなさい。

8.4 正規直交基底

$\mathbb{R}^3$ 上の基底系

$$\{\boldsymbol{u}_1, \boldsymbol{u}_2, \boldsymbol{u}_3\} = \left\{ \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, \frac{1}{\sqrt{3}} \begin{pmatrix} -1 \\ 1 \\ 1 \end{pmatrix}, \frac{1}{\sqrt{6}} \begin{pmatrix} 1 \\ -1 \\ 2 \end{pmatrix} \right\} \tag{9}$$

を考える。

1. $\{\boldsymbol{u}_1, \boldsymbol{u}_2, \boldsymbol{u}_3\}$ が正規直交基底系であることを示しなさい

2. $\boldsymbol{u}_1, \boldsymbol{u}_2$ からなる二次元の正規直交基底系を考える。$\boldsymbol{x}^T = (0, \sqrt{2}, \sqrt{3})$ をこの正規直交基底系に射影し、その座標を求めなさい

8.5 主成分の導出 (1)

$\boldsymbol{x}_i \in \mathbb{R}^D$ $(i = 1, \ldots, N)$, $\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1^T \\ \boldsymbol{x}_2^T \\ \vdots \\ \boldsymbol{x}_N^T \end{pmatrix}$ とする。$\boldsymbol{x}_1, \ldots \boldsymbol{x}_N$ の第一主成分を求めるための準備として以下を導きなさい。

1. $\boldsymbol{u}_1 \in \mathbb{R}^D$ とする。$\boldsymbol{x}_1, \ldots \boldsymbol{x}_N$ をそれぞれ $\boldsymbol{u}_1$ に射影した座標を $\boldsymbol{z}_1^T = (z_{11}, \ldots, z_{1N})$ とする。$\boldsymbol{z}_1 = \boldsymbol{X}\boldsymbol{u}_1$ であることを示せ。

2. $\boldsymbol{x}_1, \ldots \boldsymbol{x}_N$ が中央化されているならば、$\bar{z}_1 = \frac{1}{N} \sum_{i=1}^N z_{1i} = 0$ であることを示せ

3. $\boldsymbol{x}_1, \ldots \boldsymbol{x}_N$ が中央化されているならば、$z_{11}, \ldots, z_{1N}$ の分散は $\sigma_{z_1}^2 = \boldsymbol{u}^T \boldsymbol{\Sigma} \boldsymbol{u}$ であることを示せ。ここで $\boldsymbol{\Sigma} = \frac{1}{N} \boldsymbol{X}^T \boldsymbol{X}$ である

4. $\boldsymbol{X}$ が中央化されているとき, $\boldsymbol{\Sigma} = \frac{1}{N} \boldsymbol{X}^T \boldsymbol{X}$ は $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ の分散共分散行列であることを示せ

8.6 主成分の導出 (2)

記法は 8.5 に準じる。等式制約を持つ最適化におけるラグランジュ緩和を用いれば、第一主成分方向 $\boldsymbol{u}_1$ は以下を満たす方向であることがわかる。

$$\nabla_{\boldsymbol{u}_1}\left[\boldsymbol{u}_1^T\boldsymbol{\Sigma}\boldsymbol{u}_1 - \lambda(\boldsymbol{u}_1^T\boldsymbol{u}_1 - 1)\right] = \boldsymbol{0} \tag{10}$$

$$\boldsymbol{u}_1^T\boldsymbol{u}_1 = 1 \tag{11}$$

1. 式 (10) は, $(\boldsymbol{\Sigma} - \lambda\boldsymbol{I})\boldsymbol{u}_1 = \boldsymbol{0}$ と等価であることを示せ。
2. $\sigma_{z_1}^2 = \lambda$ であることを示せ。
   (以降は解説) よって、主成分方向は、分散共分散行列の固有ベクトルの方向であることがわかる。分散を最大にする固有ベクトルの方向は、

8.7 主成分分析の例

プログラム PCA_iris.ipynb を実行して以下の問題に答えよ。プログラムは iris データについて主成分数 4 の主成分分析を実行し、その寄与率 (explained variance ratio) および第 i 主成分と第 j 主成分の二次元空間にデータを射影し、その散布図を表示している。上から下に $i = 1, 2, 3, 4$ の主成分, 左から右に $j = 1, 2, 3, 4$ の主成分の組み合わせで散布図が生成されている。またあやめの種類毎に異なる色の点でプロットされている。

1. 第一主成分, 第二主成分方向の組み合わせによる散布図では、3 種のアヤメは比較的よく分離して分布しているが、より番号の大きい主成分方向の組み合わせ（例えば第三主成分と第四主成分方向）による散布図では、三種のアヤメは混在して分布している。その理由を説明しなさい。
2. このアヤメが setosa かどうかを分類する分類器を構築することを考える。このとき、第一主成分の値のみを使って分類器を構築したとする。このような分類器を、4 次元の全ての特徴を用いた分類器と比較した時のメリットとデメリットを説明せよ。

8.8 総合問題 (過去の期末試験より)

近年, キノコ狩りが大ブームである. しかし野山で採取したキノコは毒キノコが含まれるため, 専門家の判断を仰がずに食べることは危険であり，誤食による事故が後を絶たない. あなたが機械学習を学んだことを聞きつけた T 市の保健所は, 機械学習を用いてキノコの毒性を判定させるプログラムを作ることをあなたに依頼した. あなたは野山へいき, N 本のキノコを採取した.

各キノコから取得可能なデータは以下の通りである. [傘の色 (R,G,B), 茎の色 (R,G,B), 傘の直径と茎の直径の比, 傘の直径と全長の比, 茎の根元の直径, 同位置に生えていた同種のキノコの数]. これらは全て数値データであり, 並べると 10 次元となる. また専門家に依頼し, 各キノコの毒性を鑑別してもらい, 毒性の有無を示すデータを得た. n 番目のキノコの毒性以外のデータを $\boldsymbol{x}_n \in \mathbb{R}^{10}$, 毒性のデータを $t_n \in \{$ 毒性あり, 毒性なし $\}$ とする. このとき，以下の問いに答えよ.

1. あなたは, 毒性の有無を $f(\boldsymbol{x}) = t$ の形式で直接推定するアプローチ A と, 毒性の有無を $f(\boldsymbol{x}) = P(t|\boldsymbol{x})$ といった条件付き確率の形式で直接推定するアプローチ B を検討している. 一般の分類問題におけるそれぞれのアプローチをなんと呼ぶか. またそれぞれのアプローチの代表的な機械学習の手法名をあげなさい。

2. あなたは二種類の分類プログラム P と Q を作成した. 今, 手元には N 個のキノコのデータ $\{(\boldsymbol{x}_n, t_n)\}_{n=1}^N$ があり, これだけを用いて P と Q のどちらの予測精度 (テスト誤差) が高いかを決定し, 良い方だけを保健所に渡したい. テスト誤差の予測の分散を減らすために使われる代表的な手法名を一つ上げなさい.

3. 分類プログラムを保健所に渡したところ, 分類プログラムはなかなか良い精度で毒性判定ができて便利であるが, 毒性ありと毒性なしのキノコのデータ分布 (散布図) を目で見て判別することはできないだろうか, と相談された. キノコデータは 10 次元の数値データであり, 人間が直接目で見ることはできない. あなたは, 毒性ありと毒性なしのキノコの分布の特徴が最もよくあらわれるような低次元空間に射影し可視化すれば, データを目で見て判定することができると考えた. これを実現する代表的な手法名を一つあげなさい

4. 保健所から, 分類プログラムの精度をもっと上げてほしいと頼まれたあなたは, 学習データを追加すれば予測精度の向上が望めると思い, 再び野山へ行き大量のキノコを採取し, 専門家に毒性鑑別を依頼し, キノコを毒ありと毒なしの二つのかごに分けて保管していた. 新しい分類プログラムを作るにあたり, S 教授に相談しようと二つのかごを渡したところ, そそっかしい S 教授はすべってころんでかごをひっくり返してしまい, 全てのキノコが一様に混ざってしまった. やり場のない怒りがこみ上げた. しかし気を取り直して思い出してみると, 毒キノコとそうでないキノコの間には，見かけやサイズ上の明確な違いがあったように思われたため, 混ざってしまった全てのキノコのデータを測定し, データ同士でグループわけすることで、毒性のありなしを判別できるのではないかと考えた. これを実現する代表的な機械学習の手法名を一つあげなさい。

9　第九回

9.1　3 層 NN の逆誤差伝播法の導出

入力ベクトルを $\boldsymbol{x} \in \mathbb{R}^D$, 目標値を $\boldsymbol{d} \in \mathbb{R}^K$ の回帰を考える。ベクトル $\boldsymbol{a}$ の第 i 要素を $\boldsymbol{a}_i$ などと書くことにする。訓練サンプル $(\boldsymbol{x}, \boldsymbol{d})$ とする (学習には訓練サンプルは多数必要であるが、確率的勾配降下法を適用することを考え、ここでは訓練サンプルを一つだけ考える)。

この訓練サンプルについて、三層ニューラルネットワーク (入力層, 中間層, 出力層) による回帰の学習を考える。第 ℓ 層の第 i ノードへの入力を $u_i^{(\ell)}$, 出力を $z_i^{(\ell)}$ と書く。また第 ℓ 層の第 i ノードと第 $\ell+1$ 層の第 j ノードの間の重みパラメータを $w_{ji}^{(\ell+1)}$ と書く。全ての層の重みパラメータの集合を便宜的に $\boldsymbol{W}$ と書く。

このニューラルネットワークの情報の流れは以下の通りである。第一層 (入力層) の第 i ノードへの入力を $u_i^{(1)} = x_i$ とする。入力層の第 i ノードへは、その入力をそのまま出力 $z_i^{(1)} = u_i^{(1)}$ する。第二層の第 j ノードへの入力は $u_j^{(2)} = \sum_i w_{ji}^{(2)} z_i^{(1)}$ である。このニューラルネットワークでは第二層において非線形活性化関数 $f : \mathbb{R} \to \mathbb{R}$ を適用する。よって第二層の第 j ノードの出力は $z_j^{(2)} = f(u_j^{(2)})$ である。第三層の第 k ノードへの入力は $u_k^{(3)} = \sum_i w_{kj}^{(3)} z_j^{(2)}$ である。第三層の第 k ノードは入力をそのまま出力 $z_k^{(3)} = u_k^{(3)} = y_k$ する。第三層のの第 k ノードの出力は入力 $\boldsymbol{x}$ の関数であることを意識して, $\boldsymbol{y}_k(\boldsymbol{x})$ などと書く。

訓練サンプル $(\boldsymbol{x}, \boldsymbol{d})$ に対する二乗誤差は以下の通りである。

$$E(\boldsymbol{W}) = \|\boldsymbol{y}(\boldsymbol{x}) - \boldsymbol{d}\|_2^2 \tag{12}$$

ここでは, このニューラルネットワークの確率的勾配降下法を行うために必要な勾配 $\frac{\partial E}{\partial w_{ji}^{(2)}}$ および $\frac{\partial E}{\partial w_{kj}^{(3)}}$ を逆誤差伝播法によって求める。

1. $\frac{\partial E}{\partial w_{kj}^{(3)}}$ を求めよ
2. $\frac{\partial E}{\partial w_{ji}^{(2)}}$ を求めよ

9.2　一般的な NN の逆誤差伝播法の導出

入力ベクトルを $\boldsymbol{x} \in \mathbb{R}^D$, 目標値を $\boldsymbol{d} \in \mathbb{R}^K$ の回帰を考える。訓練サンプル $(\boldsymbol{x}, \boldsymbol{d})$ とする (もちろん訓練サンプルは多数必要であるが、確率的勾配降下法を適用することを考え、ここでは訓練サンプルを一つだけ考える)。

この回帰のための L 層 (入力層は $\ell = 1$, 中間層は $\ell = 2, \ldots, L-1$, 出力層は $\ell = L$) からなるニューラルネットワークを考える。第 ℓ 層の第 i ノードへの入力を $u_i^{(\ell)}$, 出力を $z_i^{(\ell)}$ と書く。また第 ℓ 層の第 i ノードと第 $\ell+1$ 層の第 j ノードの間の重みパラメータを $w_{ji}^{(\ell+1)}$ と書く。全ての層の重みパラメータの集合を便宜的に $\boldsymbol{W}$ を書く。

第 ℓ 層の第 j ノードへの入力は $u_j^{(\ell)} = \sum_i w_{ji}^{(\ell)} z_i^{(\ell-1)}$ である。第 ℓ 層の第 j ノードの出力は $z_j^{\ell} = f(u_j^{(\ell)})$ である。ここで f は活性化関数である。活性化関数は層ごとに異なる場合もあるが、簡単のために活性化関数には層を表すインデクスは

つけていない。第 ℓ 層の第 j ノードと第 $\ell+1$ 層の第 k ノードをつなぎリンクの重みは $w_{kj}^{(ell+1)}$ である。第 L 層の第 k ノードの出力は $z_k^{(L)} = f(u_k^{(L)}) = y_k$ である。第 L 層の出力は入力 $\boldsymbol{x}$ の関数であることを意識して, その第 k 要素を $\boldsymbol{y}_k(\boldsymbol{x})$ などと書く。

訓練サンプルに対する二乗誤差は以下の通りである。

$$E(\boldsymbol{W}) = \|\boldsymbol{y}(\boldsymbol{x}) - \boldsymbol{d}\|_2^2 \tag{13}$$

ここでは, このニューラルネットワークの確率的勾配降下法を行うために必要な勾配 $\frac{\partial E}{\partial w_{ji}^{(L)}}$ および $\frac{\partial E}{\partial w_{kj}^{(\ell)}}$ を求める。

1. $\frac{\partial E}{\partial w_{ji}^{(L)}}$ を求めよ
2. $\frac{\partial E}{\partial u_j^{(\ell)}} \equiv \delta_j^{(\ell)}$ とする。$\frac{\partial E}{\partial w_{ji}^{(\ell)}}$ を $\delta_k^{(\ell+1)}, k = 1, 2, \ldots,$ の式で求めよ