機械学習演習問題（線形回帰、ロジスティック回帰、確率的勾配法）

1. 線形回帰

1.1
線形回帰は（あ）の１つであり、目的変数が説明変数の係数に対して線形なモデルを学習するものである。線形回帰では訓練データに適合する回帰係数を学習する際、（い）の最小化を行う。

1.1.1 （あ）にあてはまるのはどれか。
1. 教師あり学習の回帰手法
2. 教師あり学習の分類手法
3. 教師なし学習の回帰手法
4. 教師なし学習の分類手法

1.1.2 （い）にあてはまるのはどれか。
1. 尤度関数
2. 二乗誤差
3. 0/1損失
4. 対数損失

1.2 N個の説明変数($X_1, X_2, .., X_N$)と目的変数 Y があるとすし、線形回帰を行うとする。X_1 と Y の相関係数が-0.95であったとき、正しい説明はどれか。
1. X_1 と Y の関係性は弱い
2. X_1 と Y の関係性は強い
3. X_1 の回帰係数は正、かつ絶対値が小さい
4. X_1 の回帰係数は負、かつ絶対値が大きい

1.3 リッジ回帰において正則化係数を十分に大きくすると、一般に回帰係数はどうなるか。
1. いくつかの回帰係数は完全に0となる
2. いくつかの回帰係数は0に近づくが、完全に0とはならない
3. いくつかの回帰係数は無限大になる
4. 問題に依存していずれにもなる

1.4 ラッソ回帰において正則化係数を十分に大きくすると、一般に回帰係数はどうなるか。
1. いくつかの回帰係数は完全に0となる
2. いくつかの回帰係数は0に近づくが、完全に0とはならない
3. いくつかの回帰係数は無限大になる
4. 問題に依存していずれにもなる

1.5 線形回帰について誤っている説明はどれか

1. 外れ値に敏感である
2. 説明変数間に相関があると、良い推定ができない可能性がある
3. モデルの表現力を下げた方がよい場合もある
4. 訓練データが多いほど過学習が起きやすい

1.6 モデルの複雑性（表現力）が大きくなると、バイアスとバリアンスはどうなるか。
1. バイアスとバリアンスのいずれも大きくなる
2. バイアスとバリアンスのいずれも小さくなる。
3. バイアスは大きくなるが、バリアンスは小さくなる
4. バイアスは小さくなるが、バリアンスは大きくなる

1.7 以下の訓練データが与えられた。

X	Y
-1.0	0.0
0.5	0.65
1.0	1.1
2.0	1.5

1.7.1 Y=aX+bと仮定すると、回帰係数a, bはどのようになるか。最も近いものを選べ。
1. a = 1.0, b = 0.5
2. a = 0.5, b = 0.5
3. a = 0.5, b = 1.0
4. a = 0.5, b = -0.5

1.7.2 最適な回帰係数a, bにおいて、二乗誤差 $\sum_{n=1}^{N}(y_n-(ax_n+b))^2$ の値はどうなるか。
1. 0
2. 0.01
3. 0.02
4. 0.2

1.7.3 上の結果をテストデータで検証すると、テスト時の誤差が0になった。新たに $Y=a_0+a_1X+a_2X^2$ と仮定し回帰係数を学習すると、どのようなことが起きると考えられるか。
1. 過学習が起きる
2. 未学習が起きる
3. バイアスが大きくなる
4. バリアンスが小さくなる

2. ロジスティック回帰

2.1

ロジスティック回帰は（あ）の１つであり、２値分類では、入力が一方のクラスに属する確率を出力するように学習する。ロジスティック回帰では学習の際、（い）の最大化を行う。

2.1.1 （あ）にあてはまるのはどれか。
1. 教師あり学習の回帰手法
2. 教師あり学習の分類手法
3. 教師なし学習の回帰手法
4. 教師なし学習の分類手法

2.1.2 （い）にあてはまるのはどれか。
1. 尤度関数
2. 二乗誤差
3. 0/1損失
4. いずれでもない

2.2 ラベルがy={0, 1}の1次元入力xに対して、次のロジスティック回帰モデル $p(y=1|x)=\sigma(w_0+w_1x)$ を考える。ただし $\sigma(\cdot)$ はシグモイド関数である。

2.2.1 シグモイド関数 $\sigma(h)$ の具体的な式はどのように表されるか。
1. $\sigma(h)=\frac{1}{1+\exp(h)}$
2. $\sigma(h)=\frac{1}{1+\exp(-h)}$
3. $\sigma(h)=\frac{1}{1-\exp(h)}$
4. $\sigma(h)=\frac{1}{1-\exp(-h)}$

2.2.1 $\sigma(0)$ の値はどれか。
1. 0
2. 0.5
3. 1
4. ∞

2.2.3 $p(y=1|x)$ がとる値の範囲はどれか。
1. $(0,\ \infty)$
2. $(-\infty,\ 0)$
3. $(-\infty,\ \infty)$
4. $(0,\ 1)$

2.2.4 学習の結果 $w_0=3,\ w_1=-1$ と求まったとする。$p(y=1|x)$ が0.5より大きいなら1に分類し、0.5より小さいなら0に分類する。このとき決定境界の式はどのように表されるか。
1. x = 3
2. x = -3
3. x = 2.5
4. x = -2.5

2.2.5 上の問題と同じ状況において、x=2.6, 3,1はそれぞれどちらに分類されるか。

1. 0, 0
2. 0, 1
3. 1, 0
4. 1, 1

2.3 以下のデータ(x_1, x_2 は説明変数、y は目的変数)が与えられた。

x_1	x_2	y
0	0	1
0	1	0
1	0	0
1	1	1

2.3.1 ロジスティック回帰で分類することを考える。正しい説明はどれか。
1. $p(y=1|x) = \sigma(w_0 + w_1 x_1)$ では完全に分類できる
2. $p(y=1|x) = \sigma(w_0 + w_2 x_2)$ では完全に分類できる
3. $p(y=1|x) = \sigma(w_0 + w_1 x_1 + w_2 x_2)$ では完全に分類できる
4. いずれでも完全に分類することはできない

2.3.2 上のデータをロジスティック回帰で分類するための工夫として適切なのはどれか。
1. 使用する説明変数を減らす
2. 学習の際に正則化を行う
3. 高次の特徴空間やへ写像したり、新たな特徴量にしてから適用する
4. いずれも適切でない

2.4 ロジスティック回帰に関して誤っている説明はどれか。
1. 表現力の高いモデルに対して正則化すると、決定境界は複雑になる。
2. 説明変数間に相関があると、良い推定ができない可能性がある
3. 一般にモデルに入れる説明変数の数が多いほど表現力が上がる
4. ロジスティック回帰においても、ラッソ正則化において正則化係数を十分に大きくするといくつかの係数は完全に0となる

3. 確率的勾配法

3.1 確率的勾配降下法とは最適化手法の一つであり、バッチ学習である最急降下法をオンライン学習に改良したものである。つまり、最適化するパラメータをθ、最小化する目的関数を$L(\theta) = \sum_{i=1}^{N} L_i(\theta)$ (i は学習サンプルのインデックス)とすると一回の更新が最急降下法では（あ）であるが、確率的勾配降下法では（い）である。

（あ）と（い）の組み合わせとして正しいのはどれか。ただしη を学習率とする。
1. （あ） $\theta \leftarrow \theta + \eta \nabla_\theta L(\theta)$ 、（い） $\theta \leftarrow \theta + \eta \nabla_\theta L_i(\theta)$
2. （あ） $\theta \leftarrow \theta + \eta \nabla_\theta L_i(\theta)$ 、（い） $\theta \leftarrow \theta + \eta \nabla_\theta L(\theta)$
3. （あ） $\theta \leftarrow \theta - \eta \nabla_\theta L(\theta)$ 、（い） $\theta \leftarrow \theta - \eta \nabla_\theta L_i(\theta)$

4. （あ）$\theta \leftarrow \theta - \eta \nabla_\theta L_i(\theta)$、（い）$\theta \leftarrow \theta - \eta \nabla_\theta L(\theta)$

3.2 確率的勾配降下法に関して誤っている説明はどれか。
1. 訓練データサイズが大きい場合はバッチ最急降下法より確率的勾配降下法を用いるのがよい
2. 確率的勾配降下法を適用するためには目的関数のパラメータに関する勾配が計算できる必要がある
3. 訓練データの順番を変えると結果も変わる可能性がある
4. 結果は学習率には依存しない

3.3 パラメータの更新の際に、学習速度を早めるために慣性項を追加するモメンタムという手法がある。モメンタムなしのパラメータの更新を $\theta_{t+1} \leftarrow \theta_t - \eta \nabla_{\theta_t} L(\theta_t)$ とすると、モメンタムありのときのパラメータの更新式は以下のどれか。ただし、$\alpha (> 0)$ を慣性項のパラメータとする。
1. $\theta_{t+1} \leftarrow \theta_t - \eta \nabla_{\theta_t} L(\theta_t) + \alpha(\theta_t - \theta_{t-1})$
2. $\theta_{t+1} \leftarrow \theta_t - \eta \nabla_{\theta_t} L(\theta_t) - \alpha(\theta_t - \theta_{t-1})$
3. $\theta_{t+1} \leftarrow \theta_t - \eta \nabla_{\theta_t} L(\theta_t) + \alpha(\theta_t)$
4. $\theta_{t+1} \leftarrow \theta_t - \eta \nabla_{\theta_t} L(\theta_t) - \alpha(\theta_t)$

3.4 以下のデータが与えられた。

X	Y
-1.0	0.0
0	0.5
2.0	1.5

$Y = aX + b$ と仮定し、そのパラメータa, bを求める。このとき最小化する目的関数を $L(a,b) = \frac{1}{2} \sum_{n=1}^{N} (y_n - (ax_n + b))^2$ とする。

3.4.1 目的関数のa, bに関する偏微分 $\frac{\partial L}{\partial a}, \frac{\partial L}{\partial b}$ はどのように表せるか。
1. $\frac{\partial L}{\partial a} = \sum_{n=1}^{N} (y_n - (ax_n + b))x_n$, $\frac{\partial L}{\partial b} = \sum_{n=1}^{N} (y_n - (ax_n + b))$
2. $\frac{\partial L}{\partial a} = -\sum_{n=1}^{N} (y_n - (ax_n + b))x_n$, $\frac{\partial L}{\partial b} = -\sum_{n=1}^{N} (y_n - (ax_n + b))$
3. $\frac{\partial L}{\partial a} = \sum_{n=1}^{N} (y_n - (ax_n + b))^2$, $\frac{\partial L}{\partial b} = \sum_{n=1}^{N} (y_n - (ax_n + b))^2$
4. $\frac{\partial L}{\partial a} = -\sum_{n=1}^{N} (y_n - (ax_n + b))$, $\frac{\partial L}{\partial b} = -\sum_{n=1}^{N} (y_n - (ax_n + b))$

3.4.2 バッチ最急降下法で最小化すると一回目のパラメータの更新量はa, bそれぞれいくらになるか。ただし a, b の初期値をそれぞれ $1.0,\ 0.0$、学習率 $\eta = 0.1$ とする。
1. 2.0, -1.0
2. 0.2, -0.1

3. -2.0, 1.0
4. -0.2, 0.1

3.4.3 確率的勾配降下法で最小化すると一回目のパラメータの更新量はa, bそれぞれいくらになるか。ただしa, bの初期値をそれぞれ$1.0,\ 0.0$、学習率$\eta = 0.1$、一回目の訓練サンプルをx=-1.0とする。
1. 1.0, -1.0
2. 0.1, -0.1
3. -1.0, 1.0
4. -0.1, 0.1

3.4.4 引き続き、二回目のパラメータの更新量はa, bそれぞれいくらになるか。ただし二回目の訓練サンプルをx=0.0とする。
1. 0.0, 0.4
2. 0.0, 0.04
3. 0.1, -0.4
4. 0.1, -0.04

3.4.5 モメンタムありの確率的勾配降下法でパラメータを更新すると、問題3.4.3の後二回目のパラメータの更新量はa, bそれぞれいくらになるか。ただし慣性項のパラメータ$\alpha = 0.9$, 二回目の訓練サンプルをx=0.0とする。
1. 0.0, 0.04
2. -0.09, 0.09
3. -0.09, 0.05
4. -0.09, 0.13

3.4.6 収束するまでパラメータの更新を行うと、パラメータa, bはどのような値になるか。
1. 1, 1
2. 0.5, 0.5
3. 1, 0.5
4. 1, -0.5

3.5 勾配降下法の工夫として学習率を最初は大きくし、学習が進むにつれて小さくするような工夫がある。学習率の初期値を$\eta_0 = 0.1$とし、学習ステップtに応じて学習率を$\eta = \eta_0 \frac{1}{t+1}$と変えるとすると、$t = 100$において学習率はどのようになるか。最も近いものを選べ。
1. 0.1
2. 0.01
3. 0.001
4. 0.0001