機械学習演習問題解答（線形回帰、ロジスティック回帰、確率的勾配法）

1. 線形回帰

1.1.1 **1**
線形回帰は教師あり学習で、回帰手法の一つである。

1.1.2 **2**
線形回帰では、教師データyとモデルの出力値の二乗誤差を最小化すること（最小二乗法）で、パラメータを求める。

1.2 **2**
相関係数が-0.95ということは負の相関が強いのでX_1とYの関係性は強いと言える。そこから、X_1の回帰係数は0でないことがわかる。しかし、X_1の回帰係数の大きさについてはわからない。

1.3 **2**
リッジ回帰では$||w||^2$を小さくするように学習するので、正則化係数を十分に大きくするといくつかの回帰係数は0に近づく。しかし完全に0とはならない。

1.4 **1**
ラッソ回帰では$|w|$を小さくするように学習するので、正則化係数を十分に大きくするといくつかの回帰係数は完全に0となる。つまりスパースな回帰係数wが得られる。回帰において説明変数の選択などに利用されることもある。

1.5 **4**
4が誤り。訓練データが多いほど過学習は起きにくい。逆に訓練データが少ないと過学習が起きやすい。1, 2, 3は正しい。1については訓練データにおいて外れ値を取り除くなどの対策などがある。2は多重共線性ともよばれる。3については複雑なモデルでは表現力が上がる一方で過学習しやすくなるので、説明変数の数を減らしたり、正則化などをすることでモデルの表現力を下げることもある。

1.6 **4**
モデルの複雑性（表現力）が大きくなると、バイアスは小さくなるがバリアンスは大きくなる。バイアス・バリアンストレードオフの問題である。

1.7.1 **2**
計算によっても求められるが、プロットしてみるとa = 0.5, b = 0.5が最も誤差が小さくなることが分かる。

1.7.2 **3**

$\sum_{n=1}^{4} (y_n - (0.5x_n + 0.5))^2 = 0 + 0.01 + 0.01 + 0 = 0.02$ となる。

1.7.3 **1**
テスト時の誤差が0なので1次の線形回帰が最も汎化性能が高いと考えることができる。このデータに対して2次の線形回帰をすると、テスト時の誤差が0より大きくなることが期待される。つまり過学習が起きると考えられる。

2. ロジスティック回帰

2.1.1 **2**
ロジスティック回帰は教師あり学習で、名前から間違えやすいが分類手法の１つである。

2.1.2 **1**
ロジスティック回帰ではパラメータを最尤推定で求める。

2.2.1 **2**
シグモイド関数は $\sigma(h) = \frac{1}{1+\exp(-h)}$ と定義される。

2.2.2 **2**
$\sigma(0) = \frac{1}{1+\exp(0))} = \frac{1}{1+1} = 0.5$

2.2.3 **4**
シグモイド関数の値域が(0, 1)なので、$p(y = 1|x)$ の値域も(0, 1)である。

2.2.4 **1**
$p(y = 1|x) = \sigma(3 - x)$ であり、この値が0.5より大きいか小さいかで判断するので、決定境界は $3 - x = 0$ つまり $x = 3$

2.2.5 **3**
$x = 2.6$ は $p(y = 1|x = 2.6) = \sigma(0.4) > 0.5$ より1に分類される。$x = 3.1$ は $p(y = 1|x = 3.1) = \sigma(-0.1) < 0$ より0に分類される。

2.3.1 **4**
いずれでも完全に分類することはできない。1, 2, 3のいずれも決定境界は線形となるが、データを完全に分類するためには非線形な決定境界が必要である。

2.3.2 **3**
3が適切である。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ロジスティック回帰を適用することで完全に分類可能である。例えば、$p(y = 1|x) = \sigma(w_0 + w_1\, x_1 x_2)$ など。1, 2では決定境界は線形なままなので完全に分類することはできない。

2.4 **1**
1が誤り。表現力の高いモデルに対して正則化すると決定境界は滑らかになる。2, 3, 4は正しい。いずれも線形回帰にも共通する特徴である。

3. 確率的勾配法

3.1 **3**
最急降下法ではすべてのサンプルに対する目的関数を計算してからパラメータを更新し、確率的勾配降下法ではランダムな1つのサンプルに対する目的関数を計算してからパラメータを更新する。つまり、目的関数を最小化する場合はそれぞれ$\theta \leftarrow \theta - \eta \nabla_{\theta} L(\theta)$,
$\theta \leftarrow \theta - \eta \nabla_{\theta} L_i(\theta)$でパラメータを更新する。

3.2 **4**
4が誤り。学習率は重要な要素である。学習率が小さすぎると学習に時間がかかったり、非凸関数では局所最適解に陥りやすい。一方、学習率が大きすぎるとパラメータが発散したりする。学習率は適切に決める必要があり、そのための工夫をしたアルゴリズムが多く開発されている。1, 2, 3は正しい。1について、大規模なデータでバッチ最急降下法を行うとすべてのサンプルに対する目的関数を計算するのに時間がかかったり、メモリにのらなかったりする可能性がある。3について、確率的勾配降下法ではパラメータの更新がランダムに選ばれたサンプルに依存するので、局所最適解が複数存在する場合は最終的な結果が変わる可能性がある。

3.3 **1**
慣性項とは直前のパラメータの更新量に比例する項なので$\alpha(\theta_t - \theta_{t-1})$で与えられる。つまりモメンタムありの確率的勾配降下法のパラメータの更新式は、
$\theta_{t+1} \leftarrow \theta_t - \eta \nabla_{\theta_t} L(\theta_t) + \alpha(\theta_t - \theta_{t-1})$である。

3.4.1 **2**
$L(a, b) = \frac{1}{2} \sum_{n=1}^{N} (y_n - (ax_n + b))^2$のとき、$\frac{\partial L}{\partial a} =- \sum_{n=1}^{N} (y_n - (ax_n + b))x_n$, $\frac{\partial L}{\partial b} =- \sum_{n=1}^{N} (y_n - (ax_n + b))$である。

3.4.2 **4**
バッチ最急降下法での一回目のパラメータの更新量はそれぞれ
$-\eta \frac{\partial L}{\partial a} = 0.1 \sum_{n=1}^{3} (y_n - x_n)x_n = 0.1(-1.0 + 0.0 - 1.0) =- 0.2$
$-\eta \frac{\partial L}{\partial b} = 0.1 \sum_{n=1}^{3} (y_n - x_n) = 0.1(1.0 + 0.5 - 0.5) = 0.1$

3.4.3 **4**
確率的勾配降下法での一回目のパラメータの更新量はそれぞれ
$-\eta \frac{\partial L}{\partial a} = 0.1(y_i - x_i)x_i = 0.1 \times (-1.0) =- 0.1$
$-\eta \frac{\partial L}{\partial b} = 0.1(y_i - x_i) = 0.1 \times 1.0 = 0.1$

3.4.4 **2**
一回目の更新後、$a = 0.9$, $b = 0.1$となる。その後二回目のパラメータの更新量はそれぞれ
$-\eta \frac{\partial L}{\partial a} = 0.1(y_i - (0.9x_i + 0.1))x_i = 0.1 \times 0.0 = 0.0$
$-\eta \frac{\partial L}{\partial b} = 0.1(y_i - (0.9x_i + 0.1)) = 0.1 \times 0.4 = 0.04$

3.4.5 **4**

モメンタムありのときの二回目のパラメータの更新量はそれぞれ

$-\eta\frac{\partial L}{\partial a} + \alpha(\theta_t - \theta_{t-1}) = 0.0 + 0.9 \times (-0.1) = -0.09$

$-\eta\frac{\partial L}{\partial b} + \alpha(\theta_t - \theta_{t-1}) = 0.04 + 0.9 \times 0.1 = 0.13$

3.4.6 **2**

データは3点とも直線y=0.5x+0.5上にあるので、パラメータa, bはそれぞれ0.5, 0.5に収束する。

3.5 **3**

$\eta = \eta_0\frac{1}{t+1} = 0.1\frac{1}{101} \approx 0.001$