機械学習演習問題解答

1. 最尤推定

1.1.1 **1**

最尤推定とは尤度が最大となるパラメータを推定する手法である。

1.1.2 **1**

各サンプルが独立に、同じ分布から取り出されるということを独立同分布という。

1.1.3 **1**

対数の性質から $\log \prod_{i=1}^{n} p(x_i ; \theta) = \sum_{i=1}^{n} \log p(x_i ; \theta)$

1.2.1 **2**

「表、裏、裏、裏、表」となるのは $\theta \times (1-\theta) \times (1-\theta) \times (1-\theta) \times \theta = \theta^2 (1-\theta)^3$

1.2.2 **2**

$\log \theta^2 (1-\theta)^3 = 2 \log \theta + 3 \log(1-\theta)$

1.2.3 **2**

$\frac{\partial}{\partial \theta} \log \theta = \frac{1}{\theta}$, $\frac{\partial}{\partial \theta} \log(1-\theta) = \frac{-1}{1-\theta}$ なので、 $2 \log \theta + 3 \log(1-\theta)$ の偏微分は $\frac{2}{\theta} - \frac{3}{1-\theta}$

1.2.4 **2**

最尤推定量とは対数尤度関数が最大となる点であり、対数尤度関数をθについて偏微分した値が0になる点なので、 $\frac{2}{\theta} - \frac{3}{1-\theta} = 0$ より $\hat{\theta} = \frac{2}{5}$

1.3 **1**

1が正しい。2は誤り。サンプル数が少なくなるほど過学習は起きやすくなる。3も誤り。最尤推定が常に最も適切な推定手法である。4は誤り。最尤推定は確率分布のパラメータを推定する手法なので、データはパラメータに依存した確率分布からサンプルされるものと仮定する。

2. サポートベクターマシン(SVM)

2.1.1 **1**

SVMは教師あり学習である。

2.1.2 **1**

データ点 x を $y(x) = w^T \phi(x) + b$ の正負で分類する。ラベルは{-1, +1}なので、 $sign(w^T \phi(x) + b)$ によって分類する。

2.1.3 **2**

SVMではマージンが最大となるように学習する。

2.1.4 **1**
マージン上にあるデータ点をサポートベクトルという。ソフトマージンSVMにおいてはマージン上または、マージン内部または、誤分類されたデータ点をサポートベクトルという。

2.2.1 **3**
サポートベクトルは境界から最も近いデータ点なので、このデータセットでは正負それぞれ、4, 3がサポートベクトルになる。

2.2.2 **1**
4, 3がサポートベクトルとなるので、分離境界はx = 3.5という式でかける。つまりx - 3.5 = 0よりw = 1, b = -3.5である。

2.2.3 **2**
(4.5, 1)はサポートベクトルにならないので、学習結果（wとbの値）は変化しない。

2.3 **3**
3が誤り。SVMではサポートベクトル（決定境界から近い点・マージン上の点）がデータ予測に利用される。つまり決定境界から離れたデータは予測に影響しない。1, 2, 4は正しい。

2.4 **2**
2が正しい。1は誤り。SVMは決定的な出力を行うため、ラベルに属する確率は出力しない。3は誤り。SVMは汎化性能の高い分類器だが、すべての場合において最も汎化性能の高い分類器とは限らない。4は誤り。サポートベクトルの数の最小値は正例・負例1つずつの2である。

2.5 **4**
4が誤り。パラメータの変更だけでは、テスト時の正解率は予測ができないので、Cが大きいほうがテスト時の正解率が高くなるとはいえない。1は正しい。$C \to \infty$のときはマージンから内側方向への誤差(ξ_i)に対して∞のペナルティを与えるので、マージン内に訓練データが入ることや誤分類を一切許容しないようになる。2は正しい。$C \to 0$のときはマージンから内側方向への誤差(ξ_i)に対してペナルティを与えないので、誤分類が多くなる。3は正しい。Cが大きいと完全に分離することに過度になり、汎化性能が小さくなるので、一般に過学習を防ぐためにはCを小さくするほうがよい

2.6 **2**
$||x - z_1||^2 \approx 0,\ ||x - z_2||^2 \approx \infty$ なので $\exp(-\frac{||x - z_1||^2}{2\sigma^2}) \approx 1,\ \exp(-\frac{||x - z_2||^2}{2\sigma^2}) \approx 0$ つまり
$k(z_1,\ x) \approx 1,\ k(z_2,\ x) \approx 0$ である。

2.7 **1**
$k(x,\ x') = \varphi(x)^T \varphi(x)$ で定義され、 $\frac{x^T x'}{||x||\ ||x'||} = (\frac{x}{||x||})^T (\frac{x'}{||x'||})$ なので $\varphi(x) = \frac{x}{||x||}$

2.8 **1**
1が不適。線形カーネルでは線形分離しかできない。2, 3, 4は非線形な分離が可能である。

2.9 **2**
SVMは決定的な出力をする分類器である。ロジスティック回帰は確率的な出力をする分類器である。k-means、主成分分析は分類器ではない。

3. 最近傍法, k近傍法(k-NN)

3.1.1 **1**
最近傍法では距離が最小となる点を参照する

3.1.2 **1**
k-NNでは距離が近い順にk個のデータ点を参照する

3.1.3 **2**
最近傍法と1-NNは同じである。

3.2.1 **2**
x=0.8から距離が最小のデータ点はx=0.7なのでラベルにはx=0.7のラベル0が割り当てられる。同様にx=2.2から距離が最小のデータ点はx=2.0なのでラベルにはx=2.0のラベル1が割り当てられる。

3.2.2 **4**
x=0.8から距離が近い順に3点取り出すと、x=0.7, 1.0, 1.6なのでラベルにはそれらのラベル0, 1, 1の最頻値である1が割り当てられる。同様にx=2.2から距離が近い順に3点取り出すと、x=2.0, 2.5, 1.6なのでラベルにはそれらのラベル1, 0, 1の最頻値である1が割り当てられる。

3.2.3 **4**
3-NNと同様に距離が近い順に5点取り出して対応するラベルの最頻値を割り当てれば良い。

3.2.4 **1**
距離が近い順に7点だが、これは全データなのでx=0.8, 2.2のいずれに対しても、全データのラベルの最頻値である0が割り当てられる。

3.3.1 **3**
データ点(0.5, 0.5)以外のデータは正しく分類されるので、精度は⅘=0.8

3.3.2 **1**
k=1のときは全てのデータ点が誤って分類されるので、精度は0である。k=3, 5のときの精度は0.8なので、精度が最も低いのはk=1

3.4 **3**
3が正しい。1は誤り。kの値が大きいほど分類精度は良くなるとはかぎらない。2は誤り。kの値が小さいほど決定境界は線形の部分が多く角張った教会になる。3は誤り。決定境界は部分線形ではあるが、線形ではない。

3.5 **3**
3が誤り。k-NNでは陽に訓練ステップがないので、訓練ステップは陽に存在せず、かかる時間はkの値に依存しない。予測にかかる時間も訓練データ数には依存するが、kの値にはほとんど依存しない。1, 2, 4は正しい。2に関して、高次元になると各データとの距離がほとんど同じになってしまう（次元の呪い）ため、適切に機能しなくなる。

4. 主成分分析

4.1.1 **1**
主成分分析において分散が大きい成分は重要な成分である。

4.1.2 **1**
各主成分は互いに直交するように選ばれる。

4.2.1 **1**
データ点が一直線上にあり、この方向成分が第1主成分となる。また主成分はノルム（長さ）が1に正規化されるので $(1/\sqrt{2},\ 1/\sqrt{2})$

4.2.2 **2**
第2主成分は第1主成分に直交する成分なので $(1/\sqrt{2},\ -1/\sqrt{2})$

4.2.3 **3**
データ点(-1, -1)を第1主成分 $(1/\sqrt{2},\ 1/\sqrt{2})$ 上へ射影して得られる座標は(内積の性質より) $(1/\sqrt{2},\ 1/\sqrt{2})(-1,\ -1)^T = -\sqrt{2}$ となる。同様に(0, 0), (1, 1)に対しては $0,\ \sqrt{2}$ となる。

4.2.4 **1**
データ点は一直線上にあるので、第1主成分に射影しても情報は損失しないので、再構成によって完全にもとの情報を復元できる。つまり再構成誤差は0%である。

4.3.1 **1**
(-2, -2), (0, 0), (2, 2)という一直線に並んだ3点とそれに直交する(-1, 1), (0, 0), (1, -1)という一直線に並んだ3点があり、前者の方が分散が大きいので第1主成分は、ノルム（長さが）1に正規化されることを考慮して $(1/\sqrt{2},\ 1/\sqrt{2})$ となる。そして第2主成分は第1主成分に直交する成分なので $(1/\sqrt{2},\ -1/\sqrt{2})$ となる。

4.3.2 **3**
データ点(-2, -2)を第1主成分 $(1/\sqrt{2},\ 1/\sqrt{2})$ 上へ射影して得られる座標は $(1/\sqrt{2},\ 1/\sqrt{2})(-2,\ -2)^T = -2\sqrt{2}$ となる。同様に(-1, 1), (0, 0), (1, -1), (2, 2)に対しては $0,\ 0,\ 0,\ 2\sqrt{2}$ となる。

4.3.3 **3**
データ点(-2, -2)を第2主成分 $(1/\sqrt{2},\ -1/\sqrt{2})$ 上へ射影して得られる座標は $(1/\sqrt{2},\ -1/\sqrt{2})(-2,\ -2)^T = 0$ となる。つまりデータ点(-2, -2)を第1主成分と第2主成分によっ

て張られる2次元空間へ射影して得られる座標は問題4.3.2の結果も利用して、$(-2\sqrt{2},\ 0)$と表せる。同様に(-1, 1), (0, 0), (1, -1), (2, 2)に対しては第2主成分$(1/\sqrt{2},\ -1/\sqrt{2})$上へ射影して得られる座標は$-\sqrt{2},\ 0,\ \sqrt{2},\ 0$なので第1主成分と第2主成分によって張られる2次元空間へ射影して得られる座標はそれぞれ$(0,\ -\sqrt{2}),\ (0,\ 0),\ (0,\ \sqrt{2}),\ (2\sqrt{2},\ 0)$となる。

4.3.4 **1**
第1主成分によって張られる１次元空間に射影すると「$(-2\sqrt{2}),\ (0),\ (0),\ (0),\ (2\sqrt{2})$」となるが、(-1, 1), (1, -1)はどちらも(0, 0)と同じ(0)に射影されており、情報が損失していることがわかる。しかし、第1主成分と第2主成分によって張られる２次元空間に射影すると$(-2\sqrt{2},\ 0),\ (0,\ -\sqrt{2}),\ (0,\ 0),\ (0,\ \sqrt{2}),\ (2\sqrt{2},\ 0)$となり、いずれも区別でき情報を損失していないことがわかる。よって再構成誤差は第1主成分によって張られる１次元空間に射影されたデータのほうが大きくなる。

4.4 **4**
4が誤り。データを変数間の相関がないように線形変換することはできるが、非線形変換を行うことはできない。1, 2, 4は正しい。

4.5 **2**
2が誤り。一般的に主成分分析の目的は次元削減なので、主成分の数は少ないほうがよい。1, 3, 4は正しい。3において主成分の数を3以下にしたほうが良いのは、2次元や3次元が人間に理解できる（見て分かる）次元だからである。

5. k平均クラスタリング(k-means)

5.1 **3**
k-meansは教師なしクラスタリング手法である。

5.2 **3**
k-meansアルゴリズムは以下のとおりである。
1) 各クラスタ中心の初期値を設定する
2) 各データ点に、最も距離が近いクラスタを割り当てる
3) 各クラスタの平均ベクトル（中心）を計算する
4) 収束するまで2, 3の処理を繰り返す
つまりc → b → a → d

5.3.1 **1**
各データ点とクラスタ中心-3.0, 0の距離を計算し、近い方のクラスタを割り当てると、「A, B, B, B, B」となる。

5.3.2 **2**
クラスタAは「-2.7」の1点のみなので中心は-2.7、クラスタBは「-1.3, 0.7, 3.5, 5.1」の4点なので中心は$\frac{-1.3+0.7+3.5+5.1}{4} = 2.0$と更新される。

5.3.3 **2**

各データ点とクラスタ中心-2.7, 2.0の距離を計算し、近い方のクラスタを割り当てると、「A, A, B, B, B」となる。

5.3.4 **3**
クラスタAは「-2.7, -1.3」の2点なので中心は $\frac{-2.7-1.3}{2} = -2.0$ 、クラスタBは「0.7, 3.5, 5.1」の3点なので中心は $\frac{0.7+3.5+5.1}{3} = 3.1$ と更新される。

5.3.5 **2**
3回目のステップにおいて、各データ点とクラスタ中心-2.7, -1.3の距離を計算し、近い方のクラスタを割り当てると、「A, A, B, B, B」となり、2回目のステップのときと変化がないので、収束したと判断する。つまり最終的なクラスタリング結果は「A, A, B, B, B」である。

5.3.6 **3**
各データ点とクラスタ中心-3.0, 5.0の距離を計算し、近い方のクラスタを割り当てると、「A, A, A, B, B」となる。

5.3.7 **4**
クラスタAは「-2.7, -1.3, 0.7」の2点なので中心は $\frac{-2.7-1.3+0.7}{3} = -1.1$ 、クラスタBは「3.5, 5.1」の2点なので中心は $\frac{3.5+5.1}{2} = 4.3$ と更新される。

5.3.8 **3**
2回目のステップにおいて、各データ点とクラスタ中心-1.1, 4.3の距離を計算し、近い方のクラスタを割り当てると、「A, A, A, B, B」となり、1回目のステップのときと変化がないので、収束したと判断する。つまり最終的なクラスタリング結果は「A, A, A, B, B」である。

5.3.9 **1**
各データ点とクラスタ中心-3.0, 0, 5.0の距離を計算し、近い方のクラスタを割り当てると、「A, B, B, C, C」となる。

5.3.10 **1**
クラスタAは「-2.7」の1点のみなので中心は-2.7、クラスタBは「-1.3, 0.7」の2点なので中心は $\frac{-1.3+0.7}{2} = -0.3$ 、クラスタCは「3.5, 5.1」の2点なので中心は $\frac{3.5+5.1}{2} = 4.3$ と更新される。

5.3.11 **1**
2回目のステップにおいて、各データ点とクラスタ中心-2.7, -0.3, 4.3の距離を計算し、近い方のクラスタを割り当てると、「A, B, B, C, C」となり、1回目のステップのときと変化がないので、収束したと判断する。つまり最終的なクラスタリング結果は「A, B, B, C, C」である。

5.4 **4**
(a), (b), (c)はいずれも最適な解を得るための工夫である。(a)についてはクラスタリング結果が初期値に依存するため用いられる工夫である。(b)はステップ数を増やしてアルゴリズムが完全に収束することを期待して用いられる工夫である。(c)はクラスターの数を変化させることで得られる結果が異なるので、適切なkを探すための工夫である。(d)は誤りでクラス

ターの数を多くすることとクラスタリング結果には関係がない。むしろ理解でき意味のある結果を得るためにはクラスターの数は多くしすぎないほうがよい。

5.5 **4**
そもそもk-meansとk-NNは全くことなるアルゴリズムだが、名前から間違えやすいので注意が必要である。4が正しい。1は誤り。k-meansは教師なし学習でありk-NNは教師あり学習である。2は誤り。k-meansはクラスタリング手法でありk-NNは分類（識別）や回帰手法である。3は誤り。k-meansとk-NNいずれにおいてもkはユーザーが設定するパラメータであり、学習するパラメータではない。