機械学習演習問題

1. 最尤推定

1.1

最尤推定とは与えられたデータからそれが従う確率分布の母数を点推定する方法である。より具体的には、ランダムなサンプル $x_1,\ x_2,\ ..,\ x_n$ が得られた時、それらがある未知のパラメータ θ によって決まる確率分布 $p(x\ ;\theta)$ からサンプルされたと仮定し、観測データが得られる確率（尤度）が（あ）となるパラメータ θ を求めるというものである。

尤度関数 $L(\theta)$ はパラメータが θ と仮定した時の $x_1,\ x_2,\ ..,\ x_n$ の同時確率と定義される。つまり $L(\theta) = p(X_1 = x_1,\ X_2 = x_2,\ ..,\ X_n = x_n\ ;\theta) = p(x_1\ ;\theta)p(x_2\ ;\theta)\cdots p(x_n\ ;\theta) = \prod_{i=1}^{n} p(x_i\ ;\theta)$ である。ここで2番目の等式が成り立つためには $x_1,\ x_2,\ ..,\ x_n$ が（い）という仮定が必要である。尤度関数が最大となるパラメータ θ を求める際には、計算が簡単になるため尤度関数に対数をとった対数尤度関数が用いられる。対数尤度関数は $\log L(\theta) =$（う）と表される。そして最尤推定量 $\hat{\theta}$ はこの対数尤度関数 $\log L(\theta)$ が最大となるパラメータ θ なので、停留点を計算するか最適化手法により求める。

1.1.1 （あ）に当てはまるのはどれか。

1. 最大
2. 最小
3. 0
4. 1

1.1.2 （い）に当てはまるのはどれか。

A. 互いに独立である ($p(X_i = x_i,\ X_j = x_j) = p(X_i = x_i)p(X_j = x_j)$ と書ける)
B. 互いに独立でない
C. 同じ分布からサンプルされた
D. 異なる分布からサンプルされた

1. AかつC
2. AかつD
3. BかつC
4. BかつD

1.1.3 （う）に当てはまるのはどれか。

1. $\sum_{i=1}^{n} \log p(x_i\ ;\theta)$
2. $\prod_{i=1}^{n} \log p(x_i\ ;\theta)$
3. $\sum_{i=1}^{n} p(x_i\ ;\theta)$
4. $\prod_{i=1}^{n} p(x_i\ ;\theta)$

1.2

コインを5回投げたら「表、裏、裏、裏、表」という結果になった。

1.2.1 表が出る確率をθとすると尤度関数はどう表せるか。
1. $\theta^1(1-\theta)^4$
2. $\theta^2(1-\theta)^3$
3. $\theta^3(1-\theta)^2$
4. $\theta^4(1-\theta)^1$

1.2.2 対数尤度関数はどう表せるか。
1. $\log\theta + 4\log(1-\theta)$
2. $2\log\theta + 3\log(1-\theta)$
3. $3\log\theta + 2\log(1-\theta)$
4. $4\log\theta + \log(1-\theta)$

1.2.3 対数尤度関数をθについて偏微分した結果はどう表せるか。
1. $\frac{1}{\theta} - \frac{4}{1-\theta}$
2. $\frac{2}{\theta} - \frac{3}{1-\theta}$
3. $\frac{3}{\theta} - \frac{2}{1-\theta}$
4. $\frac{4}{\theta} - \frac{1}{1-\theta}$

1.2.4 表が出る確率θの最尤推定量はいくつか。
1. 1/5
2. 2/5
3. 3/5
4. 4/5

1.3 最尤推定について正しい説明はどれか
1. 一般的に学習するデータ数が増えれば増えるほど、よい推定となる
2. サンプル数が少なくても過学習が起きることはない
3. 最尤推定は他のどの推定よりもよい推定である
4. パラメータに依存しない確率分布からデータがサンプルされたと仮定しても良い

2. サポートベクターマシン(SVM)

2.1
サポートベクターマシンは（あ）であり、分類にも回帰にも使われるが２値分類に使われることが多い。２値分類においては、データ点xを$y(x) = w^T\phi(x) + b$の（い）によって分類する（$\phi(x)$は特徴ベクトル）。これは境界$y(x) = w^T\phi(x) + b = 0$によって各クラスのデータを分離するものと考えられる。各クラスのデータ点と境界との最短距離をマージンと呼び、SVMではマージンが（う）となるような境界$y(x) = w^T\phi(x) + b\ (= 0)$(におけるパラメータ$w,\ b$)を学習する。そして（え）にあるデータ点をサポートベクトルとよぶ。

2.1.1 （あ）にあてまるのはどれか。
1. 教師あり学習

2. 教師なし学習
3. 強化学習
4. いずれでもない

2.1.2 （い）にあてまるのはどれか。
1. 正負
2. 大小
3. 0か0でないか
4. +1か-1か

2.1.3 （う）にあてまるのはどれか。
1. 最小
2. 最大
3. 0
4. 無限大

2.1.4 （え）にあてまるのはどれか。
1. マージン上
2. 境界線（面）上
3. マージンの外側
4. マージン内部と外側

2.2
以下のデータセット(xは特徴量、yはラベル)が与えられたとき、式$f(x) = sign(wx + b)$をSVMによって学習する。

x	y
1	-1
2.5	-1
3	-1
4	1
5	1

2.2.1 学習の結果サポートベクトルはいくつになるか。
1. 0
2. 1
3. 2
4. 3

2.2.2 wとbの値はどのようになるか。
1. w = 1, b = -3.5
2. w = -1, b = 3.5
3. w = 3.5, b = -1

4. w = -3.5, b = 1

2.2.3 新たに学習データセットとして(x, y) = (4.5, 1)が追加されたとき、wとbの値はどうなるか。
1. wもbも変わる
2. wもbも変わらない
3. wは変わるがbは変わらない
4. bは変わるがwは変わらない

2.3 SVMについて誤っている説明はどれか。
1. 一般にカーネル関数は暗に高次元の特徴空間へ写像することを意味するのでモデルの表現力が上がる
2. マージンの外側のデータは予測に影響を与えない
3. 識別では決定境界から最も離れたデータが予測に利用される
4. カーネルはある特徴ベクトルφ(x)の内積として定義される

2.4 SVMについて正しい説明はどれか。
1. 入力に対してラベルに属する確率を出力する
2. サポートベクトルの数は学習データ全体に対して少数である
3. すべての場合において、最も汎化性能の高い分類器である
4. サポートベクトルの数の最小値は1である

2.5 データの誤分類を許すソフトマージンSVMにおいて最小化する目的関数は、$C\sum_{i=1}^{n}\xi_i+\frac{1}{2}||w||^2$ である。ただし、ξ_i はデータ点 x_i がマージン内または誤分類されたときに正の値をとるスラック変数と呼ばれるものである。このソフトマージンSVMにおけるパラメータ C ついて誤っているのはどれか。
1. $C\to\infty$ のときは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
2. $C\to 0$ のときは誤分類が多くなる
3. 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
4. C が大きいほうがテスト時の正解率が高くなる

2.6 SVMにおいて最もよく使われるカーネルの1つにRBFカーネル($k(x,\ x')\ =\ \exp(-\frac{||x-x'||^2}{2\sigma^2})$)がある。3点 $z_1,\ z_2,\ x$ あり、z_1 と x は距離が非常に近く、z_2 と x は距離が非常に遠いと仮定する。このとき $k(z_1,\ x),\ k(z_2,\ x)$ の値はどのようになるか。
1. $k(z_1,\ x)\approx 0,\ k(z_2,\ x)\approx 1$
2. $k(z_1,\ x)\approx 1,\ k(z_2,\ x)\approx 0$
3. $k(z_1,\ x)>>1,\ k(z_2,\ x)<<0$
4. $k(z_1,\ x)<<0,\ k(z_2,\ x)>>1$

2.7 正規化線形カーネル $k(x,\ x')\ =\ \frac{x^Tx'}{||x||\ ||x'||}$ に対応する特徴ベクトルφ(x)はどれか。
1. $\phi(x)\ =\ \frac{x}{||x||}$
2. $\phi(x)\ =x$
3. $\phi(x)\ =x^Tx$

4. $\varphi(x) = \frac{x}{\overrightarrow{||x||}}$

2.8 非線形な分離に不適切なカーネルはどれか。
1. 線形カーネル
2. rbfカーネル
3. 多項式カーネル
4. シグモイドカーネル

2.9 決定的な出力をする分類器はどれか。
1. ロジスティック回帰モデル
2. SVM
3. K-Means
4. 主成分分析

3. 最近傍法, k近傍法(k-NN)

3.1
最近傍法ではあるデータ点のラベルを予測する際、訓練データの中でそのデータ点との距離が（あ）となるデータ点のラベルを割り当てるという手法である。k-NNでは訓練データの中でそのデータ点との距離が（い）データ点をk個取り出し、それらのラベルの最頻値を割り当てる手法である。最近傍法はk-NNにおいてk=（う）のときと同じである。

3.1.1 （あ）に当てはまるのはどれか。
1. 最小
2. 最大
3. 0
4. 無限大

3.1.2 （い）に当てはまるのはどれか。
1. 近い順に
2. 遠い順に
3. 0の
4. 無限大の

3.1.3 （う）に当てはまるのはどれか。
1. 0
2. 1
3. 2
4. ∞

3.2
以下のデータセット(xは特徴量、yはラベル)が与えられた。

x	y
-0.1	0

0.7	0
1.0	1
1.6	1
2.0	1
2.5	0
3.3	0

3.2.1 最近傍法で入力xからラベルyを予測する。x=0.8, 2.2が与えられた時のラベルyの組み合わせとして適切なのはどれか。
1. 0, 0
2. 0, 1
3. 1, 0
4. 1, 1

3.2.2 3-NN(k-NNにおいてk=3のとき)で入力xからラベルyを予測する。x=0.8, 2.2が与えられた時のラベルyの組み合わせとして適切なのはどれか。
1. 0, 0
2. 0, 1
3. 1, 0
4. 1, 1

3.2.3 5-NNで入力xからラベルyを予測する。x=0.8, 2.2が与えられた時のラベルyの組み合わせとして適切なのはどれか。
1. 0, 0
2. 0, 1
3. 1, 0
4. 1, 1

3.2.4 7-NNで入力xからラベルyを予測する。x=0.8, 2.2が与えられた時のラベルyの組み合わせとして適切なのはどれか。
1. 0, 0
2. 0, 1
3. 1, 0
4. 1, 1

3.3
以下のデータセット(x_1, x_2 は特徴変数、y はラベル)が与えられた。

x_1	x_2	y
0	0	0
0	1	0

1	0	0
1	1	0
0.5	0.5	1

3.3.1 3-NNにおいて、一個抜き交差検証（k分割交差検証においてk=データの個数としたもの）により得られる精度はいくつか。
1. 0
2. 0.4
3. 0.8
4. 1

3.3.2 k-NNにおいて、一個抜き交差検証により得られる精度が最も低いkはどれか。
1. 1
2. 3
3. 5
4. どれも同じ

3.4 k-NNに関して正しい説明はどれか。
1. kの値が大きいほど分類精度は良くなる
2. kの値が小さいほど決定境界はなめらかになる
3. 陽に訓練ステップを必要としない
4. 決定境界は線形である

3.5 k-NNに関して誤っている説明はどれか。
1. ノイズのみられるデータではkを大きくするのがよい
2. 高次元データより低次元データの方が向いている
3. kの値が大きいほど訓練ステップに時間がかかる
4. 最適なkの値は交差検証によって決めることができる

4. 主成分分析

4.1
主成分分析は教師なし学習の１つであり、データから重要な成分をみつける手法である。主成分分析において重要な成分とはデータの（あ）であり、この成分のことを主成分という。各主成分は（い）に選ばれる。

4.1.1 （あ）にあてはまるのはどれか。
1. 分散が大きい成分
2. 分散が小さい成分
3. 平均値
4. 中央値

4.1.2 （い）にあてはまるのはどれか。

1. 互いに直交するよう
2. ランダム
3. 長さが長い順
4. 長さが短い順

4.2
2次元空間において「(-1, -1), (0, 0), (1, 1)」という3つデータ点が与えられ、これらのデータに対して主成分分析を適用する。

4.2.1 第1主成分はどのように表されるか。
1. $(1/\sqrt{2},\ 1/\sqrt{2})$
2. $(1/\sqrt{2},\ -1/\sqrt{2})$
3. $(0,\ 1)$
4. $(1,\ 0)$

4.2.2 第2主成分はどのように表されるか。
1. $(1/\sqrt{2},\ 1/\sqrt{2})$
2. $(1/\sqrt{2},\ -1/\sqrt{2})$
3. $(0,\ 1)$
4. $(1,\ 0)$

4.2.3 第1主成分によって張られる１次元空間へデータ点「(-1, -1), (0, 0), (1, 1)」を射影すると、それぞれの点の座標は１次元空間においてどのように表されるか。
1. 「$(-\sqrt{2}),\ (0),\ (-\sqrt{2})$」
2. 「$(\sqrt{2}),\ (0),\ (-\sqrt{2})$」
3. 「$(-\sqrt{2}),\ (0),\ (\sqrt{2})$」
4. 「$(\sqrt{2}),\ (0),\ (\sqrt{2})$」

4.2.4 上の問題で１次元空間に射影されたデータを元の２次元空間に再構成すると、再構成誤差（元のデータと再構成したデータとの誤差）は何%か。
1. 0%
2. 10%
3. 30%
4. 40%

4.3
2次元空間において「(-2, -2), (-1, 1), (0, 0), (1, -1), (2, 2)」という5つのデータ点が与えられ、これらのデータに対して主成分分析を適用する。

4.3.1 第1主成分と第2主成分の組み合わせとして適当なのはどれか。
1. $(1/\sqrt{2},\ 1/\sqrt{2})$, $(1/\sqrt{2},\ -1/\sqrt{2})$
2. $(\sqrt{2},\ \sqrt{2})$, $(1/\sqrt{2},\ -1/\sqrt{2})$
3. $(1/\sqrt{2},\ -1/\sqrt{2})$, $(-1/\sqrt{2},\ 1/\sqrt{2})$
4. $(1/\sqrt{2},\ 1/\sqrt{2})$, $(-1/\sqrt{2},\ -1/\sqrt{2})$

4.3.2 第1主成分によって張られる1次元空間へデータ点「(-2, -2), (-1, 1), (0, 0), (1, -1), (2, 2)」を射影すると、それぞれの点の座標は１次元空間においてどのように表されるか。
1. 「$(-2\sqrt{2}),\ (\sqrt{2}),\ (0),\ (\sqrt{2}),\ (2\sqrt{2})$」
2. 「$(\sqrt{2}),\ (-1),\ (0),\ (1),\ (-\sqrt{2})$」
3. 「$(-2\sqrt{2}),\ (0),\ (0),\ (0),\ (2\sqrt{2})$」
4. 「$(2\sqrt{2}),\ (0),\ (0),\ (0),\ (2\sqrt{2})$」

4.3.3 第1主成分と第2主成分によって張られる2次元空間へデータ点「(-2, -2), (-1, 1), (0, 0), (1, -1), (2, 2)」を射影すると、それぞれの点の座標は2次元空間においてどのように表されるか。
1. 「$(-2\sqrt{2},\ 0),\ (\sqrt{2},\ -\sqrt{2}),\ (0,\ 0),\ (\sqrt{2},\ \sqrt{2}),\ (\sqrt{2},\ 0)$」
2. 「$(\sqrt{2},\ \sqrt{2}),\ (-1,\ 1),\ (0,\ 0),\ (1,\ -1),\ (-\sqrt{2},\ -\sqrt{2})$」
3. 「$(-2\sqrt{2},\ 0),\ (0,\ -\sqrt{2}),\ (0,\ 0),\ (0,\ \sqrt{2}),\ (2\sqrt{2},\ 0)$」
4. 「$(\sqrt{2},\ \sqrt{2}),\ (0,\ 0),\ (0,\ 0),\ (0,\ 0),\ (\sqrt{2},\ \sqrt{2})$」

4.3.4 上の問題で１次元空間に射影されたデータと２次元空間に射影されたデータを元の２次元空間に再構成すると、再構成誤差はどちらが大きくなるか。
1. 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
2. 第1主成分と第2主成分によって張られる２次元空間に射影されたデータ
3. どちらも同じ
4. 実験しないとわからない

4.4 主成分分析の使用目的について誤っているのはどれか。
1. データの特徴を抽出する
2. 高次元データを低次元にする
3. 高次元データを可視化する
4. データを非線形変換する

4.5 主成分の数について述べた以下の説明の中で誤っているのはどれか。
1. 主成分の数の最大値はデータの特徴量の数である
2. 主成分の数は目的によらず、多いほうが良い
3. 一般に可視化の際は、主成分の数を3以下にすることが多い
4. 主成分の数が少ないほど、データを射影した際に情報が落ちる

5. k平均クラスタリング(k-means)

5.1 k-meansは教師なし（あ）手法の１つである。（あ）にあてまるのはどれか。
1. 分類
2. 回帰
3. クラスタリング
4. 次元削減

5.2 k-meansのアルゴリズムにおいて以下の(a)~(d)の要素がある。これらを正しく並べたものはどれか。
(a) 各クラスタの平均ベクトル（中心）を計算する

(b) 各データ点に、最も距離が近いクラスタを割り当てる
(c) 各クラスタ中心の初期値を設定する
(d) 収束するまで2, 3の処理を繰り返す
1. a → b → c → d
2. c → a → b → d
3. c → b → a → d
4. a → c → b → d

5.3
1次元の入力「-2.7, -1.3, 0.7, 3.5, 5.1」に対してk-meansを適用する。

5.3.1 k=2とし、クラスタA, Bを考え、それぞれのクラスタ中心の初期値を-3.0, 0とする。 1回目のステップで各入力はどちらのクラスタに割り当てられるか。
1. 「A, B, B, B, B」
2. 「A, A, B, B, B」
3. 「A, A, A, B, B」
4. 「A, A, A, A, B」

5.3.2 各クラスタの中心はどう更新されるか。
1. -2.7, 0
2. -2.7, 2.0
3. -2.0, 3.1
4. -1.1, 4.3

5.4.3 ２回目のステップで各入力はどちらのクラスタに割り当てられるか。
1. 「A, B, B, B, B」
2. 「A, A, B, B, B」
3. 「A, A, A, B, B」
4. 「A, A, A, A, B」

5.3.4 各クラスタの中心はどう更新されるか。
1. -2.7, 0
2. -2.7, 2.0
3. -2.0, 3.1
4. -1.1, 4.3

5.3.5 最終的に各入力はどちらのクラスタに割り当てられるか。
1. 「A, B, B, B, B」
2. 「A, A, B, B, B」
3. 「A, A, A, B, B」
4. 「A, A, A, A, B」

5.3.6 同様にk=2だが、クラスタA, Bに対してクラスタ中心の初期値を-3.0, 5.0とする。 1回目のステップで各入力はどちらのクラスタに割り当てられるか。
1. 「A, B, B, B, B」
2. 「A, A, B, B, B」

3. 「A, A, A, B, B」
4. 「A, A, A, A, B」

5.3.7 各クラスタの中心はどう更新されるか。
1. -2.7, 0
2. -2.7, 2.0
3. -2.0, 3.1
4. -1.1, 4.3

5.3.8 最終的に各入力はどちらのクラスタに割り当てられるか。
1. 「A, B, B, B, B」
2. 「A, A, B, B, B」
3. 「A, A, A, B, B」
4. 「A, A, A, A, B」

5.3.9 k=3とし、クラスタA, B, Cを考え、それぞれのクラスタ中心の初期値を-3.0, 0, 5.0とする。 1 回目のステップで各入力はどのクラスタに割り当てられるか。
1. 「A, B, B, C, C」
2. 「A, A, B, B, C」
3. 「A, A, B, C, C」
4. 「A, A, A, B, C」

5.3.10 各クラスタの中心はどう更新されるか。
1. -2.7, -0.3, 4.3
2. -2.0, 2.1, 5.1
3. -2.0, 0.7, 4.3
4. -1.1, 3.5, 5.1

5.3.11 最終的に各入力はどちらのクラスタに割り当てられるか。
1. 「A, B, B, C, C」
2. 「A, A, B, B, C」
3. 「A, A, B, C, C」
4. 「A, A, A, B, C」

5.4 k-meansアルゴリズムにおいて最適な解を得るために用いられる工夫として適当なのはどれか。
(a) クラスタ中心の初期値を変えて実行する
(b) 繰り返し回数を調整する
(c) クラスターの数を調整する
(d) クラスターの数をできるだけ多くする
1. aとc
2. aとd
3. bとd
4. aとbとc

5.5 k-meansとk-NN両方に当てはまる特徴はどれか。

1. 教師あり学習である
2. クラスタリング手法である
3. パラメータkを学習する
4. いずれのアルゴリズムも距離計算を行う