機械学習とデータマイニング　レポート課題

28G23027 川原尚己

**課題A**

**[問１]**以下の(1)~(3)式にて省略されている計算手順を明記すること．

**[問２]**式(4)の主張が任意のとに対して成り立つことを示すこと．

**回答**

**[問１]**

を式(A1)のように定義する．

このとき，は式(A2)のように表される．

となるは式(A3)のようになる．

は下に凸な放物線であるから，となるにおいて最小値をとる．以上より，式(1)が導出された．

　次に式(2)を示す．であるから，が成り立つ．式(2)の左辺を直接変形することにより，以下の等式を得る．

は独立であるから，が成り立つ．よって，式(A4)の第二項はに等しく，式(2)が得られる．

　最後に式(3)を示す．

であり，は分布に対しての期待値で表せるから式(A5)が成り立つ．

以上より，式(3)の左辺は式(A6)のように表される．

「学習問題の具体例」の式(4)より，であるから，式(3)が得られる．

**[問２]**

「学習問題の具体例」の8ページにおいて，であり，はの下限であるから，が成り立つ．よって，Markovの不等式より式(A7)がなりたつ．

「学習問題の具体例」の７ページよりが成り立つから，式(A8)が得られる．

このとき，任意のに対し，なる整数が存在する．そのようなをとるとき，式(A8)より式(A9)が成り立つ．

以上より，式(4)がとに対して成り立つことを示せた．

**課題B**

資料中の「分布の位置推定」の話に出てくる学習アルゴリズムはどのような一致性を満たすか．「学習法の一致性」の定義に出てくるはそれぞれ何に相当し，なぜ一致性が約束できるか示すこと．

**回答**

「分布の位置推定」の話での学習アルゴリズムにおいてははそれぞれ以下の事柄に相当する．

：からの平均二乗誤差

：からの平均二乗誤差の下限

：この標本の平均

このとき，は式(B1)~(B3)のように表される．

このに対して式(A8)が成り立つから，に対する極限を考えることで一致性を満たすことが示せた．

**課題C**

式(5)の導出過程を明記せよ．

**回答**

式(C1)のようにPAC条件を考える．

，，「サル真似」における回目の更新の出力をとして式(C1)に代入すると，式(C2)を得る．

について評価を行うために，以下のような集合を考える．

がに含まれる確率は全てのを正しく分類できるよりも小さいことから，式(C4)を得る．

さらに，をとる確率がより大きくなる確率がより小さいことから，式(C5)を得る．

式(C1)及び式(C5)より，式(C6)を満たす任意のでPAC条件を満たすことがわかる．

以上より，標本複雑度はPAC条件を満たすの最小値であるから，式(5)を満たす．

**課題D**

二値分類の場合，の確率で，残りの確率でラベルが反転して となるような確率ノイズ を設計し，とととを含む上記のような等式を明記した上で，その設計の妥当性を説明せよ．

**回答**

以下のような確率ノイズを考える．

ただしは区間の一様分布に従う確率変数である．

このとき，というを考えると，99% の確率で正しいラベルが出力され，1% の確率でラベルが反転して出力されるため，題意を満たす設計であり，妥当性がある．

**課題E**

以下の3つの不等式が成り立つことを図や数式などを用いて説明せよ．

1. 実数直線上の有限区間において
2. 二次元平面における識別線において

３．における長方形において

**回答**

任意のに対して，式(E1)が成り立つ．

式(E1)の統合が成立するのは，任意の二値ラベルの組み合わせでも入力データを分類可能な分類面が存在するときである．そのため，今回の課題においては分類不可能な入力データと二値ラベルの組を少なくとも一つ例示すればよい．

１．

アイコン

自動的に生成された説明実数数直線上において，図１のような３点を分類できる有限区間は存在しないため，式(6)は成立する．

図１　実数数直線上において分類できない例

２．

グラフィカル ユーザー インターフェイス, アイコン

自動的に生成された説明二次元平面において，図２のような４点を分類できる識別線は存在しないため，式(7)は成立する．

図２　二次元平面上において分類できない例

３．

どの点も次元長平面上にない場合を考えればよい．

この点のうち，ある軸において最小値または最大値であるような点は丁度個存在する．これら個の点を頂点とする領域を考える．は，個の軸に対する最小値または最大値をすべて含んでいるため，残りの一点は必ずに含まれることになる．の頂点であるようなすべての点に同一のラベルを，残りの一点にもう一方のラベルを割り当てることを考えると，このような分類は実現できない．よって，式(8)は成立する．

**課題F**

以下の等式 (９) と (10) がなぜ成立するか，具体的に説明すること．

**回答**

式(9)の左辺は分布から得られた独立標本を用いて分類器を学習する．その後，よりランダムに選択したに対し，を用いた場合の汎化性能の期待値を表している．一方で，式(9)の右辺は学習データを固定した場合において得られたの汎化性能の期待値に対し，さらに学習データ全体において期待値をとったものである．任意の学習データは分布に従って生成されることから，学習データを固定して学習を行ったかどうかにかかわらず，学習によって得られた分類器の汎化性能の期待値は両者の間で等しくなる．

また，式(10)の右辺のはとは独立した事象であるから，式(F1)のように表せる．

学習データとしてが選ばれる確率はであるから，最終的に(F2)式が得られる．

以上より，式(9)及び式(10)が成立することが示された．