平成 26 (2014) 年度 夏入試

東京大学情報理工学系研究科創造情報学専攻

創造情報学

注意事項

1. 試験開始の合図まで, この問題冊子を開かないこと.
2. この表紙の下部にある受験番号欄に受験番号を記入しなさい.
3. 4問中3問を選択して, 日本語ないし英語で解答すること.
4. 解答用紙は3枚配られる. 1問ごとに必ず1枚の解答用紙を使用すること. 解答用紙のおもて面に書ききれないときには, うら面にわたってもよい.
5. 解答用紙の指定された箇所に, 受験番号およびその用紙で解答する問題番号を忘れずに記入すること.
6. 解答用紙および問題冊子は持ち帰らないこと.

受験番号＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿

このページは空白.

このページは空白.

第1問

N 人の気象予報士がそれぞれ予測した N 個の予報データを使って、一人のシニア予報士が天気予報を行うような天気予報システムを考える。具体的には、以下のような状況を考える。

気象予報士が N 人いて、毎日1回、明日の天気の確率分布を予測する。ただし、天気は「晴れる (=1)」か「晴れない (=0)」の2値しかとらない確率変数とし、毎日の天気は日ごとに独立であるとする。

t 日目に i 番目の予報士が予測した確率分布を $P_i^{(t)}(X)$ $(X \in \{1,0\})$ とする, ただし、$0 < P_i^{(t)}(X) < 1$ $(X \in \{1,0\})$ とする。N 人の予報士を統括するシニア予報士がいて、シニア予報士が予測する予測確率分布を基に、それらの重みつき平均として予測確率分布 $\hat{P}^{(t)}(X)$ $(X \in \{1,0\})$ を計算し、t 日目にはこれを出力する。t 日目の i 番目の予報士に対する重みを $v_i^{(t)}$ $(\sum_{i=1}^{N} v_i^{(t)} = 1, v_i(t) > 0$ $(i = 1, \ldots, N))$ とする (図1参照)。つまり、t 日目において $\hat{P}^{(t)}(X)$ は $\sum_{i=1}^{N} v_i^{(t)} P_i^{(t)}(X)$ として与えられる (図1参照)。

t 日目には、シニア予報士が予測したあと、その正解である天気のデータ x_t がその日に手に入るとする。このような予測を毎日繰り返すことを考える。

以上の設定のもとで以下の設問に答えよ。

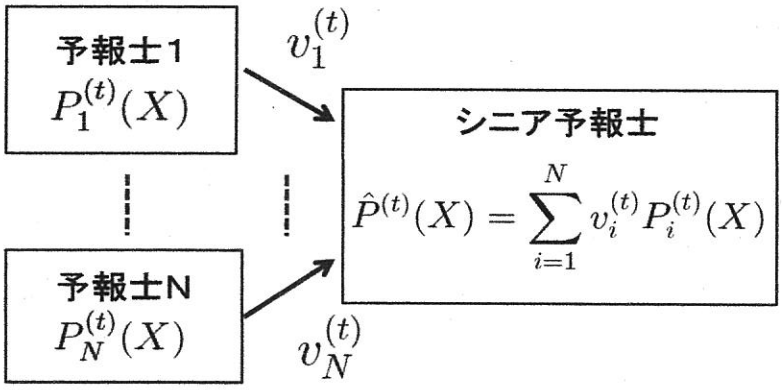

図1　シニア予報士の重みつき予測方式

(1) シニア予報士は、t 日目における i 番目の予報士に対する重み $v_i^{(t)}$ の比を、$t-1$ 日目までの過去のデータ列 $x^{t-1} = x_1 \ldots x_{t-1}$ に対する i 番目の予報士の尤度 ($w_i^{(t-1)}$ とかく) の比に等しくなるように与えるとする。つまり、シニア予報士の重み付けは、各 t について以下を満たす。

$$v_1^{(t)} : \cdots : v_N^{(t)} = w_1^{(t-1)} : \cdots : w_N^{(t-1)}.$$

ただし、x^{t-1} に対する i 番目の予報士の尤度は $w_i^{(t-1)} = \prod_{j=1}^{t-1} P_i^{(j)}(X = x_j)$ と計算される (以下、

$P_i^{(j)}(X = x_j)$ を $P_i^{(j)}(x_j)$ と記す)。ここで、$P_i^{(0)}(x_0) = 1$ $(i = 1, \ldots, N)$ とする。

このとき、各 i 番目の予報士について、$w_i^{(t)}$ と $w_i^{(t-1)}$ との関係を示し、$t+1$ 日目の重み $v_i^{(t+1)}$ を計算する式を $w_i^{(t-1)}$ と $P_i^{(t)}(x_t)$ $(i = 1, \ldots, N)$ を用いて示せ。

(2) シニア予報士が t に関して T 日間予測を繰り返していくとき、設問 (1) の重みの与え方の下で予測確率分布を日ごとに出力し、重みを更新するアルゴリズムを示し、その計算量の N と T に関するオーダを求めよ。ただし、各予報士に対する初期の重みは以下のように与える。

$$v_1^{(0)} = \ldots = v_N^{(0)} = 1/N.$$

(3) T 日間、上記予測を行ったとき、実際に観測されたデータ列 $x^T = x_1 \ldots x_T$ に対するシニア予報士の累積予測損失を

$$Loss(x^T) = \sum_{t=1}^{T} \left(-\log \hat{P}^{(t)}(x_t) \right)$$

と定めるとき、この値を $P_i^{(t)}(x_t)$ $(i = 1, \ldots, N,\ t = 1, \ldots, T)$ と N を用いて表せ。ただし、対数は自然対数とする。

(4) 設問 (3) で定めた、シニア予報士の T 日間の累積予測損失は、各予報士の中で達成される最小の T 日間の累積予測損失よりも高々 $\log N$ しか大きくならないことを証明せよ。ここに、各 i 番目の予報士の T 日間の累積予測損失は $\sum_{t=1}^{T}(-\log P_i^{(t)}(x_t))$ として定義されるとする $(i = 1, \ldots, N)$。

第 2 問

ページングによる仮想記憶に関する以下の問いに解答せよ。

システムはプロセッサ、主記憶、二次記憶装置（磁気ディスク装置など）により構成される。プロセッサは論理アドレスを用いてメモリアクセスを発行する。アクセス対象が主記憶上にあるばあい、論理アドレスはアドレス変換テーブルにより主記憶の物理アドレスに変換される。アクセス対象が主記憶上にない場合には、主記憶上のある 1 ページを選択し、二次記憶装置にあるアクセス対象を含むページと交換し、新しい物理アドレスをアドレス変換テーブルに追加する。

まず、用語を説明する。

- ページミス率：全メモリアクセス中アクセスするページが主記憶上に存在しない率。
- ページ置換アルゴリズム：主記憶に新たなページを割り付けるために置換するページを選択するアルゴリズム。
- LRU（Least Recently Used）：ページ置換アルゴリズムの一つ。もっとも遠い過去にアクセスされたページを置換対象として選択するもの。
- FIFO(First In First Out)：ページ置換アルゴリズムの一つ。もっとも遠い過去に主記憶上に割り付けられたページを置換対象として選択するもの。

(1) 主記憶のアクセス時間が Tm, 二次記憶のアクセス時間が Ts とする。ページのサイズは 4KB であるとする。データ転送自身には時間がかからないと仮定する時、主記憶より十分に大きいベクトルデータを順次アクセスする場合のページミス率と平均アクセス時間を求めよ。また、Tm=100ns, Ts= 1 ms の時の具体的数値を求めよ。

(2) 通常のプログラムでは、同じメモリデータを繰り返しアクセスする。その場合、ページ置換アルゴリズムは性能に大きくかかわる。代表的なページ置換アルゴリズムとして、LRU（Least Recently Used）と FIFO (First In First Out)が知られている。FIFO より LRU の置換アルゴリズムを用いると実行速度がより向上するプログラム構造の概略を示せ。

(3) ページ置換アルゴリズムとして LRU を厳密に実現することは困難である。そのため、近似的なアルゴリズムが用いられる。LRU を近似するページ置換アルゴリズムを説明せよ。

第３問

画像処理とパターン認識技術により、手書き文字が「C」または「I」のいずれであるかを識別する問題を考える。図１に、ある入力画像Ｘを示す。iとjをそれぞれ画素（ピクセル）のx座標、y座標を示す整数値とし、ある画素(i,j)における画素値を$f(i,j)$とする。ただし、左上隅の画素を座標原点$(0,0)$とする。

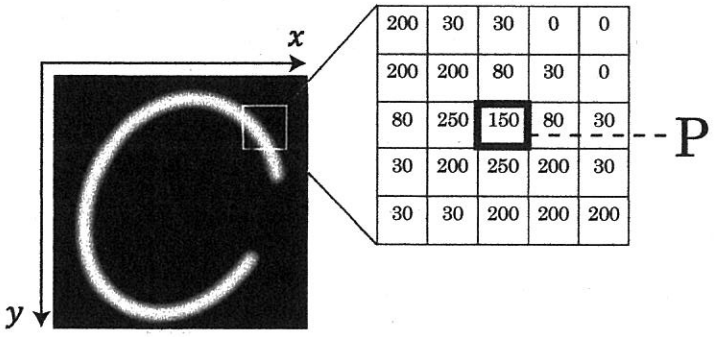

図 1： 入力画像Ｘ（図中の数字は各画素の画素値を示す）

以下に示す手順により、入力画像の性質を定量化する特徴量の抽出を行う。まず、入力画像に3×3の大きさの線形フィルタを適用する。線形フィルタの係数を表す配列を$h(m,n)$（m,nは整数値）とすると、画素(i,j)におけるフィルタの出力値 $g(i,j)$ は次式のようになる。

$$g(i,j)=\sum_{n=-1}^{1}\sum_{m=-1}^{1} f(i+m,j+n)h(m,n).$$

次の各問に答えよ。

(1) 図２に示す線形フィルタＬを画像Ｘへ適用する。図１中の画素Ｐにおけるフィルタの出力値を求めよ。

$h(-1,-1)=0$	$h(0,-1)=1$	$h(1,-1)=0$
$h(-1,0)=1$	$h(0,0)=-4$	$h(1,0)=1$
$h(-1,1)=0$	$h(0,1)=1$	$h(1,1)=0$

図 2：線形フィルタＬ

次に、フィルタ適用後の画像を6×6ピクセルの解像度へ縮小し、二値化する。すなわち、変換後の画像は $f(i,j)=\{0,1\}\,(0\le i,j\le 5)$ となる。このような二値画像の例Y1、Y2を図３に示す。
これらから、画像モーメントと呼ばれる特徴量を抽出する。$(p+q)$次の画像モーメントは $M_{pq}=\sum_{i,j} i^p j^q f(i,j)$ と定義される。例えば、M_{00}は二値画像中の図形の［ (A) ］、$(M_{10}/M_{00},\ M_{01}/M_{00})$ は図形の［ (B) ］を意味していることが分かる。

(2) 空白(A)、(B)に入れるべき適切な内容をそれぞれ記述せよ。また、Y1、Y2のそれぞれから画像モーメントM_{00}、M_{10}を計算せよ。

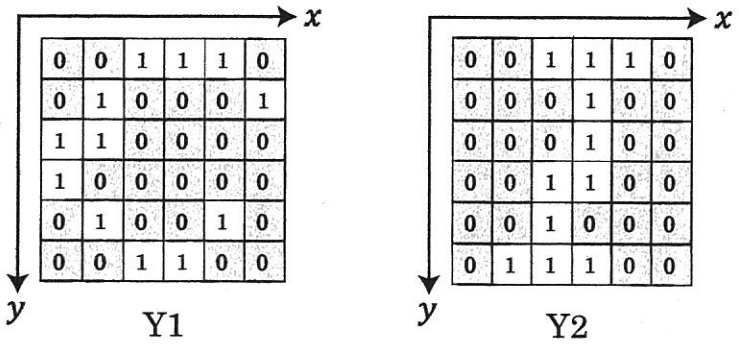

図 3： 二値画像 Y1（左）Y2（右）

以上のように、与えられた画像から抽出される 2 つの特徴量のペア(M_{00}, M_{10})を入力パターンとし、これが二つのクラス「C」または「I」のいずれであるかを識別する。表 1 に、4 つの「C」と 4 つの「I」の学習用サンプルの特徴量値を示す。また、これらを 2 次元座標系へプロットしたものを図 4 に示す。

表 1：学習用サンプル

	C1	C2	C3	C4	I1	I2	I3	I4
M_{00}	10	12	12	14	8	10	10	12
M_{10}	24	24	26	26	27	26	28	27

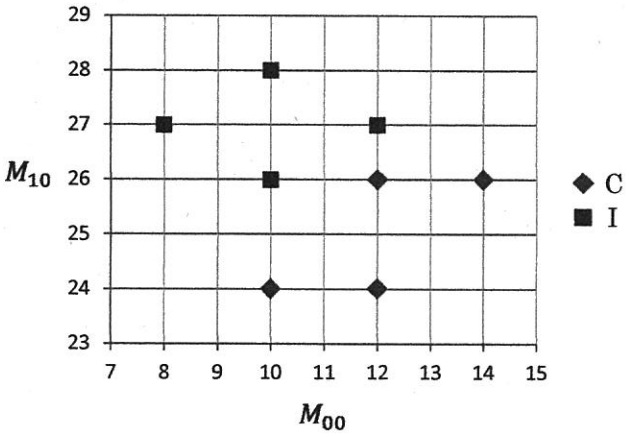

図 4： 学習用サンプル（プロット）

(3) 入力パターンをk最近傍法で識別することを考える。k最近傍法とは、入力パターンと最も距離が近い上位k個の学習用サンプルの中で、最も多い数のサンプルが所属するクラスへ識別するアルゴリズムである。今、未知の入力画像のパターンが得られ、その特徴量値は$(M_{00}, M_{10}) = (13, 27)$となった。$k$最近傍法を用い、この入力パターンが $k = 1$、$k = 3$ それぞれの場合について「C」または「I」のいずれであるかを識別し、その理由を述べよ。ただし、サンプル間の距離はユークリッド距離により定義する。

(4) 別の識別アルゴリズムとして、クラスごとの学習用サンプルの平均値までのユークリッド距離が小さい方へ入力パターンを識別する方法を考える。この時、識別境界の方程式およびこれを用いた識別方法を与えよ。また、これを用いて入力パターン$(M_{00}, M_{10}) = (13, 27)$ を識別せよ。

(5) 前問(3)、(4)の識別アルゴリズムを比較し、それぞれの長所について述べよ。

第 4 問

以下に示す情報システムに関する 8 項目から 4 項目を選択し、各項目を 4〜8 行程度で説明せよ。 必要に応じて例や図を用いてよい。

(1) tf-idf
(2) ZMP (Zero Moment Point)
(3) 分散ハッシュ
(4) 最短経路問題
(5) ベイジアンネットワーク
(6) キャリー・ルック・アヘッド
(7) クロージャ(閉包)
(8) 有限オートマトン

このページは空白.

このページは空白.