パターン認識・課題5

課題 5.1 決定株(Decision Stump)による識別

課題 5.1.1 決定株1:一つの特徴量に注目した閾値決定

決定木の枝分かれが1段しかないものを,決定株という.
(決定木を根っこのところで切り倒した切り株のイメージ.)

決定株識別器では,ある特徴量を1つ選び(コマンドライン引数で渡す),それに対して1つ閾値を決定する.
手計算の演習では,0/1の特徴量だったので閾値の決定をする必要がなかったが,実際は,どの値で2つに分割するかを決めなければならない.
各特徴量でそれぞれ閾値を決定し,最もうまく分類できる特徴量と閾値の組を求めることにより, その特徴量の閾値に対する大小でクラス分類を行う.

閾値の決定方法:ジニ不純度による決定

ジニ不純度は,1つの集合内に,異なるクラスのデータがどれだけ混在しているかを表す数値.

ジニ不純度は$\sum_{i=1}^{N} \frac{n_i}{N}(1-\frac{n_i}{N}) = 1 - \sum_{i=1}^{N}(\frac{n_i}{N})^2$で求められる.

ジニ不純度によって閾値を決定するアルゴリズムは以下の通り.

1. 選んだ特徴量が小さい順になるよう,学習データをソートする
2. ソートした順に特徴量の値を見ていく
    1. すべての特徴量を,閾値をもとに2クラスに分類する.
    2. 閾値より小さい側のジニ不純度と,閾値より大きい側のジニ不純度を計算し,和を求める.
3. ジニ不純度が最小になる閾値を決定する.
4. 識別器は,どの(何個目の)特徴量を使ったかと,求めた閾値の値とを保存する

レポートには,以下のサンプルデータに対し,各特徴量を選択した時のジニ不純度を記載すること.

```
 実行例
$ ./gini 0
特徴量:0 閾値:??? ジニ不純度:0.???
$ ./gini 1
特徴量:1 閾値:??? ジニ不純度:0.???
$ ./gini 2
特徴量:2 閾値:??? ジニ不純度:0.???
```

特徴量0	特徴量1	特徴量2	クラス
1	1	1	ω_1
2	2	2	ω_1
3	2	2	ω_1
2	3	1	ω_1
1	4	3	ω_2
2	3	2	ω_2
1	3	3	ω_2
1	1	3	ω_2

課題 5.1.2 決定株2:特徴量の選択

実際に決定株を使う場合,どの特徴量を利用するかが重要. ここでは,どの特徴量を利用して分類するかをジニ不純度をもとに判別する.

1. 各特徴量についてジニ不純度を最小にする閾値と,その時のジニ不純度を求める.
2. 最もジニ不純度が小さい特徴量と,その時の閾値をファイルに保存する.

課題 5.1.1で求めた各ジニ不純度の最小の特徴量が選択されたことを確認する.

```
 実行例
$ ./stump_train train.dat
特徴量:0 閾値:??? ジニ不純度:0.???
特徴量:1 閾値:??? ジニ不純度:0.???
特徴量:2 閾値:??? ジニ不純度:0.???
選択: 特徴量:? 閾値:??? ジニ不純度:0.???
（ファイルに特徴量番号と閾値を保存）
```

課題 5.1.3 決定株による認識プログラムの実行

評価用データを決定株によって認識する.

1. 決定株のパラメータを読み込む.
2. 決定株が保持している特徴量の値を取得する.
3. その値が,決定株が保持している特徴量よりも大きいか小さいかによって(1 or 2)を出力する.

```
 実行例
$ ./stump_test test1.dat
認識結果: ?
$ ./stump_test test2.dat
認識結果: ?
```

ただし,以下の例それぞれを,test1.dat, test2.datとする.

特徴量**0**	特徴量**1**	特徴量**2**	クラス
1	2	1	ω_1
3	3	2	ω_2

ヒント

決定株のデータ構造例

```
typedef struct Stump_{
    int feat_index; /* 特徴量の番号 */
    float threshold; /* 閾値 */
} Stump;
```

ジニ不純度計算の例

```
#include <stdio.h>
#include <stdlib.h>

#define N 8 /* サンプル数 */
#define P 3 /* 次元数（特徴量の数） */

typedef struct Stump_{
        int feat_index; /* 特徴量の番号 */
        float threshold; /* 閾値 */
} Stump;


float values[N];
int indices[N];

int main(){
        int i;
        float min_gini = 1.0f;
        Stump st;
        float features[N][P] = /* 値を入れて初期化 */
        int labels[N] = /* 値を入れて初期化 */

        /* 使う特徴量を決める */
        st.feat_index = /* 特徴量の次元*/;

        /* 特徴量の値にしたがって並べ替える */
        /* values配列に値をコピーする */
        /* indices配列を準備する */
        for(i=0; i<N; i++){
                values[i] = features[i][st.feat_index];
                indices[i] = i;
        }
        /* values配列の値でindices配列をソートする */

        for(/* 特徴量の小さいものから順に閾値とする */){
                float g1, g2;
                float sum_gini;

                g1 = /* 閾値以下のサンプルについてジニ不純度を計算 */
                g2 = /* 閾値以上のサンプルについてジニ不純度を計算 */

                /* ジニ不純度の和を計算 */
                sum_gini = g1 + g2;

                if(/* sum_giniが最小なら */){
                        /* sum_giniを更新，その時の閾値features[indieces[i]][st.feat_index]を保
                }
        }

        printf("閾値: %f\n", st.threshold);
}
```

課題 5.1 AdaBoostによる識別

課題 5.1.1 AdaBoost学習プログラムの作成

課題 5.1.2 AdaBoostによる認識プログラムの実行

提出期限

2018/7/30

レポート提出の注意は, こちら (../report.html)

戻る