Ćwiczenie 1

Klasyfikacja danych za pomocą drzew decyzyjnych

Część teoretyczna

Wykład 3: Drzewa decyzyjne.

Zadania pomocnicze

Zapoznaj się z funkcją ClassificationTree.fit w helpie Matlaba.
Zapoznaj się z opisem zbioru danych Ionosphere – Google: "UCI Machine Learning Repository: Ionosphere Data Set".

Zadania do wykonania

Zaprojektuj drzewo klasyfikacyjne do klasyfikacji zbioru danych Ionosphere.

1. Załaduj zbiór danych Ionosphere:

   ```
   load ionosphere
   ```

   Podejrzyj zmienne.

2. Wprowadź instrukcje (modyfikacja danych):

   ```
   rand('state',#*r_k);
   X=X+rand(351,34)*0.1-0.05;
   ```

   gdzie za # wstaw numer swojej sekcji a za `r_k` aktualny rok kalendarzowy.

3. Utwórz drzewo decyzyjne:

   ```
   ctree = ClassificationTree.fit(X,Y)
   ```

   Zapoznaj się z obiektem `ctree`.

4. Zapoznaj się z graficzną i regułową reprezentacją drzewa:

   ```
   view(ctree,'mode','graph');
   view(ctree);
   ```

5. Wyznacz klasy dla każdego przykładu trenującego:

   ```
   Ynew = predict(ctree,X)
   ```

6. Wyznacz macierz przekłamań używając funkcji `plotconfusion` (wcześniej konwertuj `Y` i `Ynew` na wektory numeryczne:

   ```
   Y1(cell2mat(Y)=='b')=0;
   Y1(cell2mat(Y)=='g')=1;
   Ynew1(cell2mat(Ynew)=='b')=0;
   Ynew1(cell2mat(Ynew)=='g')=1;)
   ```

7. Oceń działanie modelu na nowych danych w procedurze kroswalidacji. Użyj funkcji `crossval` i `kfoldLoss`.

8. Ustal optymalna strukturę drzewa wykorzystując procedurę kroswalidacji. W pętli generuj drzewa dla parametru "minimalna liczba przykładów w liściu" (MinLeaf) zmienianego od 2 do 100. Każde takie drzewo oceniaj w procedurze kroswalidacji:

```
tree = ClassificationTree.fit(X,Y,'crossval','on',...
        'minleaf',m(i)); %m – minimalna liczba przykładów w liściu
blad(i) = kfoldLoss(tree); %błąd kroswalidacji w i-tej iteracji
```

   Sporządź wykres błędu w zależności od parametru m.
   Wyznacz optymalna wartość m (największa wartość m, przy której błąd utrzymuje się na niskim poziomie).

9. Pokaż optymalne drzewo w postaci graficznej:

```
OptimalTree = ClassificationTree.fit(X,Y,'minleaf',m_opt);
view(OptimalTree,'mode','graph');
```

10. Porównaj błąd osiągany przez drzewo optymalne `OptimalTree` z błędem osiąganym przez drzewo `ctree` wygenerowane przy domyślnych ustawieniach parametrów:

```
errOpt1 = resubLoss(OptimalTree) %błąd na zbiorze uczącym dla OptimalTree
err1 = resubLoss(ctree) %błąd na zbiorze uczącym dla ctree
errOpt2 = kfoldLoss(crossval(OptimalTree)) %błąd kroswalidacji dla
OptimalTree
err2 = kfoldLoss(crossval(ctree)) %błąd kroswalidacji dla ctree
```

11. Znajdź optymalny poziom przycięcia drzewa w procedurze kroswalidacji:

```
[~,~,~,bestlevel] = cvLoss(ctree,'subtrees','all','treesize','min')
```

   `bestlevel` informuje o ile poziomów trzeba przyciąć drzewo.
   W okienku graficznym możesz podejrzeć przycięte drzewo ustawiając poziom przycięcia (pruning level) na `bestlevel`.

12. Przytnij drzewo `ctree` w okienku graficznym ustawiając poziom przycięcia (pruning level) na `bestlevel`.

13. Przytnij drzewo `ctree`:

```
PruneTree = prune(ctree,'Level',bestlevel);
view(PruneTree,'mode','graph');
```

14. Dla przyciętego drzewa wyznacz błąd na zbiorze uczącym i błąd kroswalidacji. Porównaj te błędy z błędami osiąganymi przez `OptimalTree` i `ctree`.

Co powinno znaleźć się w sprawozdaniu

Sprawozdania powinny być sporządzone wg wzoru zamieszczonego na stronie.

A) Cel ćwiczenia.
B) Treść zadania.
C) Opis drzew decyzyjnych, opis zbioru danych.
D) Metodyka rozwiązania – poszczególne instrukcje Matlaba z komentarzem (zachowaj numerację zadań).

E) Zestawienie wyników (wykresy, tabele z komentarzem).
F) Wnioski końcowe.

Zadania dodatkowe dla ambitnych

1. Zaprojektuj klasyfikator zbudowany na drzewie decyzyjnym dla innego zbioru danych z UCI Machine Learning Repository (w uzgodnieniu z prowadzącym).
2. Zaprojektuj model aproksymacyjny zbudowany na drzewie regresyjnym (w uzgodnieniu z prowadzącym).
3. Wykonaj podobne ćwiczenie w innym programie, np. R, Statistica, C#, ... (w uzgodnieniu z prowadzącym).

Przykładowe zagadnienia i pytania zaliczeniowe

1. Narysuj model drzewa klasyfikacyjnego lub regresyjnego.
2. Opisz metodę uczenia drzewa decyzyjnego.
3. Opisz działanie drzewa decyzyjnego w procesie klasyfikacji nowego przykładu.
4. Kryterium stopu i ustalenie etykiety.
5. Rodzaje testów.
6. Kryterium wyboru testów.
7. Ustalenie zbioru testów kandydujących.
8. Przycinanie drzewa.

Do przygotowania na następne zajęcia

1. Zapoznać się z instrukcją do kolejnego ćwiczenia.
2. Zapoznać się z częścią teoretyczną do kolejnego ćwiczenia.
3. Wykonać zadania pomocnicze do kolejnego ćwiczenia.