Vytěžování dat - Rozhodovací stromy

Jan Hrdlička

November 29, 2011

Prořezávání stromů

- Cílem této úlohy je zjistit, jak se projevuje prořezávání stromu na úspěšnosti klasifikace na testovací množině.

Prořezávání stromů

- Cílem této úlohy je zjistit, jak se projevuje prořezávání stromu na úspěšnosti klasifikace na testovací množině.
- Rozhodovací stromy už znáte z přednášky.

Dataset

- Dataset je uměle generovaný na začátku souboru zadani10.m, který si stáhnete ze stránek cvičení. Úkolem bude tento soubor doplnit.

Dataset

- Dataset je uměle generovaný na začátku souboru zadani10.m, který si stáhnete ze stránek cvičení. Úkolem bude tento soubor doplnit.
- Dataset se skládá z "naměřených" hodnot (proměnná obs) a z tříd do kterých se snažíme klasifikovat (proměnná class).

Dataset

- Dataset je uměle generovaný na začátku souboru zadani10.m, který si stáhnete ze stránek cvičení. Úkolem bude tento soubor doplnit.
- Dataset se skládá z "naměřených" hodnot (proměnná obs) a z tříd do kterých se snažíme klasifikovat (proměnná class).
- Třídy jsou 2: "green" a "blue".

Přehled úlohy

- V m-file se nejprve vygeneruje soubor dat. Poté se v cyklu opakovaně pouští náhodné rozřazení na trénovací a testovací množinu. Vaším úkolem je uvnitř cyklu...

Přehled úlohy

- V m-file se nejprve vygeneruje soubor dat. Poté se v cyklu opakovaně pouští náhodné rozřazení na trénovací a testovací množinu. Vaším úkolem je uvnitř cyklu...
- Vytvořit rozhodovací strom z trénovací množiny

Přehled úlohy

- V m-file se nejprve vygeneruje soubor dat. Poté se v cyklu opakovaně pouští náhodné rozřazení na trénovací a testovací množinu. Vaším úkolem je uvnitř cyklu...
- Vytvořit rozhodovací strom z trénovací množiny
- Zjistit chybu na testovací množině pro různé úrovně prořezání.

Přehled úlohy

- V m-file se nejprve vygeneruje soubor dat. Poté se v cyklu opakovaně pouští náhodné rozřazení na trénovací a testovací množinu. Vaším úkolem je uvnitř cyklu...
- Vytvořit rozhodovací strom z trénovací množiny
- Zjistit chybu na testovací množině pro různé úrovně prořezání.
- Zarovnat chyby po prořezání podle hloubky zbylého stromu

Přehled úlohy

- V m-file se nejprve vygeneruje soubor dat. Poté se v cyklu opakovaně pouští náhodné rozřazení na trénovací a testovací množinu. Vaším úkolem je uvnitř cyklu...
- Vytvořit rozhodovací strom z trénovací množiny
- Zjistit chybu na testovací množině pro různé úrovně prořezání.
- Zarovnat chyby po prořezání podle hloubky zbylého stromu
- Tabulku chyb vykreslit pomocí grafu boxpolot

Vytvoření stromu

- K vytvoření stromu použijte funkci classregtree.

Vytvoření stromu

- K vytvoření stromu použijte funkci classregtree.
- DULEŽITÉ: nastavte parametry "prune" na "off" a "splitmin" na 2

Vytvoření stromu

- K vytvoření stromu použijte funkci classregtree.
- DULEŽITÉ: nastavte parametry "prune" na "off" a "splitmin" na 2
- K vytvoření stromu použijte trénovací množinu

Zjištění chyby (misclassification cost)

- Ke zjištění chyby použijte funkci test (je jich v matlabu vice, tahle je metoda u třídy @classregtree)

Zjištění chyby (misclassification cost)

- Ke zjištění chyby použijte funkci test (je jich v matlabu vice, tahle je metoda u třídy @classregtree)
- Pusťte ji na testovací množině.

Zjištění chyby (misclassification cost)

- Ke zjištění chyby použijte funkci test (je jich v matlabu vice, tahle je metoda u třídy @classregtree)
- Pusťte ji na testovací množině.
- Funkce vráti vektor chyb pro různé prořezání stromu. Na prvním místě je chyba pro strom bez prořezání, na posledním místě chyba pro strom s hloubkou 1.

Zarovnání chyb

- Chyby srovnejte do tabulky costs podle hloubky stromu po prořezání, ne podle úrovně prořezání

Zarovnání chyb

- Chyby srovnejte do tabulky costs podle hloubky stromu po prořezání, ne podle úrovně prořezání
- Příklad: během dvou cyklů naleznete 2 různě hluboké stromy, hloubky 4 a hloubky 6. Metoda test vám vrátí vektor s délkou 4 resp. 6. Vašim cílem je, aby vždy poslední indexy byly v tabulce pod sebou. Poslední indexy totiž reprezentují chybu pro strom hloubky 1, předposlední 2 atd..

Faktické náležitosti protokolu

Váš protokol by měl obsahovat:

- Boxplot daných úrovní rozhodovacích stromů

Faktické náležitosti protokolu

Váš protokol by měl obsahovat:

- Boxplot daných úrovní rozhodovacích stromů
- Vysvětlení boxplotu