INFORMACE K PRŮBĚHU ZKOUŠKY Z DVZ

Zkouška z předmětu DVZ bude mít jak část písemnou, tak ústní (k útní bude student pozván v případě nerozhodného výsledku části písemné). Otázky ke zkoušce vycházejí ze sylabu přednášky a z témat probraných na přednáškách a cvičeních:

1. Dobývání znalostí - popis a metodika procesu CRISP, základní pojmy: trénovací a testovací množina, učení s učitelem, učení bez učitele a rozdíl mezi nimi.
2. Nástroje pro modelování dat a jejich využití I: algoritmy pro konstrukci klasifikátorů (např. ID3, princip perceptronu, neuronové sítě a volba jejich parametrů) a jejich aplikace.
3. Nástroje pro modelování dat a jejich využití II: algoritmy pro shlukování (k-meas, EM, rozdíl mezi hierarchickým a nehierarchickým shlukováním, volba optimálního počtu shluků), konstrukce asociačních pravidel (s využitím Apriori algoritmu).
4. Metody vizualizace dat a jejich použití (box graf, histogram, scatter plot matrix, paralelní souřadnice, RadViz). Identifikace odlehlých a nesprávných hodnot.
5. Porozumění datům a jejich příprava: postupy pro diskretizaci, normalizaci a doplnění chybějících hodnot, agregace dat
6. Volba relevantních atributů: proč je tento problém důležitý? Selekce a extrakce příznaků.
7. Práce s časovými řadami: definice vzdálenosti dvou časových řad (a složitost příslušného výpočtu) a příklady použití, principy nástrojů pro redukci dimenze (prostřednictvím transformací).
8. Nástroje pro modelování dat a jejich využití III: postupy pro kombinaci více modelů - bagging, boosting, AdaBoost.
9. Vyhodnocení a využití modelů: křížová validace, bootstrapping, ROC křivka, křivka učení.
10. Zpracování přirozeného jazyka jako vstupu: "text mining" a jeho základní úlohy, metody reprezentace dokumentů jejich výhody či nevýhody (pojmy jako tokenizace, stem, stop words), definice vzdálenosti dokumentů (např. TF-IDF vector score).

Na vypracování písemné části bude mít student 60 minut. Mezi otázkami bude vždy nejméně jedna praktická úloha, jejíž zadání bude zvoleno tak, aby ji student mohl zvládnout bez použití počítače.