Obsah

Kapitola 1 ..... **1**
Úvod 1
1.1 Cíl kurzu 1
1.2 Základní informace o systému *STATISTICA* 1
1.3 Produkty řady *STATISTICA* 2
1.3.1 Analytické moduly 2
1.3.2 Průmyslová řešení a nástroje Six Sigma 3
1.3.3 Podnikové systémy *STATISTICA* 4

Kapitola 2 ..... **7**
Data mining 7
2.1 Co je data mining? 7
2.2 Podstata data miningu 7
2.3 Širší pohled na data mining 8
2.4 Data miningový cyklus 8
2.5 Metodika DM 8
2.5.1 Průzkum oblasti a porozumění datům 9
2.5.2 Získávání dat z databází a datových skladů 9
2.5.3 Příprava dat 9
2.5.4 Typy proměnných 9
2.5.5 Proměnné vysvětlující a vysvětlované 10
2.5.6 Typické úlohy 11
2.5.7 Učení a modelování 11
2.5.8 Hlavní techniky data miningu 12
2.5.9 Ověření, nasazení, monitorování a údržba modelu 12

Kapitola 3 ..... **15**
STATISTICA a data mining 15
3.1 Nástroje řady *STATISTICA* doporučené pro DM 15
3.2 Dokumenty systému *STATISTICA* 15
3.2.1 Tabulky 15
3.2.2 Grafy 15
3.2.3 Makra 16
3.2.4 Pracovní sešity 16
3.2.5 Protokoly 16
3.2.6 Data miningové projekty 16
3.3 *STATISTICA* Data Miner: Přehled 16
3.4 Interaktivní analýzy 17
3.4.1 Příklad 18
3.5 Program *STATISTICA* Data Miner 19
3.5.1 Příklad 19
3.5.2 Rozsáhlejší projekty 23
3.5.3 Unikátní analytické uzly 24
3.5.4 Instantní projekty 25

Kapitola 4 ..... **27**
Získávání dat z databází 27
4.1 Co je to databáze 27
4.2 Jazyk SQL 28
4.3 *STATISTICA* Query 29
4.4 Ukázky SQL dotazů 31
4.5 In-Place Database Processing 34

Kapitola 5 ..... **37**

Pro potřeby kurzu společnosti StatSoft CR s.r.o. Copyright © 2008

Příprava dat ... 37
5.1 Popis a průzkum dat ... 37
5.2 Čištění dat, odlehlé hodnoty (outliers) ... 39
5.3 Chybějící hodnoty ... 41
5.4 Odvozené proměnné ... 43
5.5 Normalizace dat ... 43
5.6 Selekce příznaků ... 44
5.7 Extrakce příznaků ... 44

Kapitola 6 ... 47
Tvorba modelu ... 47
6.1 Nejbližší soused ... 47
6.1.1 Podobnost, vzdálenost ... 47
6.1.2 Nejbližší soused formálně ... 48
6.2 Lineární modely a nejmenší čtverce ... 49
6.2.1 Lineární model formálně ... 49
6.2.2 Učení lineárního modelu – metoda nejmenších čtverců ... 50
6.2.3 Lineární modely pro regresi ... 50
6.2.4 Lineární modely pro klasifikaci ... 51
6.3 Logistická regrese ... 53
6.3.1 Logistická regrese pro dichotomii ... 54
6.3.2 Učení modelu ... 54
6.4 Rozšíření báze ... 54
6.5 Klasifikační a regresní stromy ... 56
6.5.1 Proč stromy? ... 56
6.5.2 Není strom jako strom... ... 58
6.5.3 Jak roste strom? ... 58
6.5.4 Stromy typu CHAID ... 58
6.5.5 Stromy typu CART ... 59
6.5.6 Stromy typu QUEST ... 59
6.5.7 Poznámky ... 59
6.6 Shluková analýza (clustering) ... 60
6.6.1 Hierarchické shlukování ... 60
6.6.2 Metoda k-průměrů (k-means) ... 62
6.6.3 Volba optimálního počtu shluků ... 63
6.6.4 EM shlukování ... 63

Kapitola 7 ... 65
Ohodnocení modelu ... 65
7.1 Ilustrační příklady ... 65
7.2 Složitost modelu a přeučení modelu ... 67
7.2.1 Chyba na trénovacích a testovacích datech ... 67
7.2.2 Chyba a složitost modelu ... 68
7.3 Volba modelu a jeho ohodnocení ... 69
7.3.1 Trénovací, ověřovací a testovací data ... 69
7.3.2 Křížové ověření (Cross-validation) ... 70
7.4 Grafické metody hodnocení modelu ... 71
7.4.1 Graf navýšení (Lift chart) ... 71
7.4.2 ROC křivka ... 72

Kapitola 8 ... 75
Nasazení modelu ... 75
8.1 Export modelů ... 75
8.1.1 Export do SVB ... 75
8.1.2 Export do C++ ... 75
8.1.3 Export do PMML ... 76
8.2 Monitorování a údržba ... 76
8.2.1 Když model nefunguje... ... 76
8.2.2 Když model přestává fungovat... ... 77

 Pro potřeby kurzu společnosti StatSoft CR s.r.o. Copyright © 2008