Univerzita Karlova v Praze

Matematicko-fyzikální fakulta

BAKALÁŘSKÁ PRÁCE

Lenka Goduľová

Vybrané problémy a metody při zpracování mnohorozměrných finančních dat

Katedra pravěpodobnosti a matematické statistiky

Vedoucí bakalářské práce: RNDr. Jitka Zichová, Dr.

Studijní program: Matematika

Studijní obor: Finanční matematika

Praha 2014

Na tomto mieste by som sa rada poďakovala vedúcej mojej bakalárskej práce RNDr. Jitke Zichovej, Dr. za cenné rady a ochotu, s ktorou mi venovala svoj čas a za poskytnuté materiály, mojim rodičom a súrodencom za všestrannú podporu počas celého môjho doterajšieho štúdia a Petronelle Antoniewiczovej za trpezlivosť a podporu.

Prohlašuji, že jsem tuto bakalářskou práci vypracovala samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova v Praze má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle § 60 odst. 1 autorského zákona.

V Prahe dne 22.5.2014 Lenka Goduľová

Název práce: Vybrané problémy a metody při zpracování mnohorozměrných finančních dat

Autor: Lenka Goduľová

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: RNDr. Jitka Zichová, Dr.

Abstrakt: Předložená bakalářská práce se zabývá zpracováním vícerozměrných dat. Úkolem bylo aplikovat vybrané metody na finanční data. Skládá se z teoretické části a zpracování konkrétní databáze. V prvních čtyřech kapitolách jsou shrnuté základní vztahy a pojmy týkající se náhodného vektora, náhodné veličiny, vícerozměrných dat a testu nezávislosti v kontingenční tabulce. Následující část je věnována popisu vybraných metod, kterými jsou shluková analýza a diskriminační analýza. V praktické části jsou tyto metody aplikované na databázi klientů německé banky.

Klíčová slova: náhodný vektor, mnohorozměrné rozdelení, mnohorozměrný náhodný výběr, kontingenční tabulka, shluková analýza, diskriminačná analýza

Title: Selected problems and methods in multivariate data analysis

Author: Lenka Goduľová

Department: Department of Probability and Mathematical Statistics

Supervisor: RNDr. Jitka Zichová, Dr.

Abstract: The bachelor thesis deals with processing multidimensional data. The task was to apply selected methods on financial data. The thesis is composed of the theoretical section and the analysis of a particular database. The first four chapters deal with basic relations and definitions concerning random vector and variable, multidimensional data and the independence test in a contingency table. The following section is devoted to defining the particular methods selected: cluster analysis and discriminant analysis. In the practical section these methods are applied to a database of clients of a German bank.

Keywords: random vector, multivariate distribution, multivariate random variable, contingency table, cluster analysis, discriminant analysis.

Názov práce: Vybrané problémy a metódy pri spracovaní mnohorozmerných finančných dát

Autor: Lenka Goduľová

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedúcí bakalárskej práce: RNDr. Jitka Zichová, Dr.

Abstrakt: Predložená bakalárska práca sa zaoberá spracovaním mnohorozmerných dát. Úlohou bolo aplikovať vybrané metódy na finančné dáta. Pozostáva z teoretického základu a spracovania konkrétnej databáze. V prvých štyroch kapitolách sú zhrnuté základné vzťahy a pojmy týkajúce sa náhodného vektora, náhodnej veličiny, mnohorozmerných dát a testu nezávislosti v kontingenčnej tabuľke. Následná časť je venovaná popisu vybraných metód, ktorými sú zhluková analýza a diskriminačná analýza. V praktickej časti sú tieto metódy aplikované na databázu klientov nemeckej banky.

Klúčové slová: náhodný vektor, mnohorozmerné rozdelenie, mnohorozmerný náhodný výber, kontingenčná tabuľka, zhluková analýza, diskriminačná analýza

Obsah

Úvod

V dnešnej dobe existuje mnoho metód na analýzu dát. V tejto práci si ukážeme dve metódy spracovania mnohorozmerných dát. Dáta tohoto typu používajú napríklad poisťovne, či banky používajúce rozsiahle databáze klientov. Popíšeme si zhlukovú a diskriminačnú analýzu.

Pre porozumenie zhrniene v troch prvých kapitolách teóriu, prostedníctvom ktorej budeme nazerať na skúmaný problém. Postupne sa budeme venovať vektorom a maticiam, náhodným veličinám či vektorom a v neposlednej rade aj mnohorozmerným dátam.

V štvrtej kapitole jednotlivé metódy popíšeme. Rozoberieme spôsob ich použitia a uvedieme niektoré ich vlastnosti. Posledná kapitola je venovaná aplikácii spomínaných analýz na dátach nemeckej banky pomocou softvérových produktov Mathematica, Microsoft Office Excel a NCSS. Výpočty a výstupné protokoly z jednotlivých softvérov nájdeme na priloženém CD.

Kapitola 1

Vektory a matice

V prvej kapitole definujeme základné pojmy, ktoré budeme následne využívať. Oboznámime sa s pojmami vetor a matica. Vymenujeme si ich niektoré vlastnosti. Teoretické poznatky pochádzajú z [6], [7] a [8].

1.1 Vektor a matice

1.1.1 Vektor

Telesom T nazveme aspoň dvojprvkovú množinu spĺňajúcu axiómy pre operácie násobenia a sčítania. Týmito axiómami sú myslené komutativita sčítania, asociatitívne zákony, existencia nulového, jednotkového a opačného prvku, inverzný prvok a distributívne zákony. Ak navyše platí komutativita pre násobenie, potom hovoríme o komutatívnom telese.

Vektorovým priestorom nad telesom T je neprázdna množina V, ktorej všetky prvky spĺňajú nasledujúce axiómy:

$$
\begin{aligned}
&\forall \mathbf{a}, \mathbf{b} \in V: && \mathbf{a}+\mathbf{b}=\mathbf{b}+\mathbf{a} \\
&\forall \mathbf{a}, \mathbf{b}, \mathbf{c} \in V: && (\mathbf{a}+\mathbf{b})+\mathbf{c}=\mathbf{a}+(\mathbf{b}+\mathbf{c}) \\
&\forall \mathbf{a}, \mathbf{b} \in V, \forall \alpha \in T: && \alpha(\mathbf{a}+\mathbf{b})=\alpha \mathbf{a}+\alpha \mathbf{b} \\
&\forall \mathbf{a} \in V, \forall \alpha, \beta \in T: && (\alpha+\beta)\mathbf{a}=\alpha \mathbf{a}+\beta \mathbf{a} \\
&\forall \mathbf{a} \in V, \forall \alpha, \beta \in T: && \alpha(\beta \mathbf{a})=(\alpha \beta)\mathbf{a} \\
&\forall \mathbf{a} \in V, \exists 1 \in T: && 1\mathbf{a}=\mathbf{a} \\
&\forall \mathbf{a} \in V, \exists 0 \in T, \exists \mathbf{0} \in V: && 0\mathbf{a}=\mathbf{0}.
\end{aligned}
\tag{1.1}
$$

Prvky telesa T nazývame skaláry a prvkom vektorového priestoru hovoríme vektory. Príkladom vektorového priestoru je vektorový priestor nad telesom reálnych čísel R. V texte práce ním bude k-rozmerný reálny priestor, ktorý označíme R^k. Vektor dĺžky k si môžeme predstaviť ako usporiadaný rad čísel $x_1, x_2 \ldots, x_k$. Stĺpcový vektor $\mathbf{0}$, ktorý je uvedený v (1.1), nazveme nulovým vektorom. Má všetky zložky rovné nule. Stĺpcový vektor budeme značiť $\boldsymbol{x} = (x_1, ..., x_k)^T$.

Vektory sa vyznačujú rôznymi vlastnosťami a charakteristikami. Nás bude zaujímať lineárna závislosť. Hovoríme, že k-rozmerné vektory $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_s$ sú lineárne závislé, ak existujú reálne čísla $c_1, c_2 \ldots, c_s$ (aspoň jedno nenulové) také, že

$$
c_1\boldsymbol{x}_1 + c_2\boldsymbol{x}_2 + \ldots + c_s\boldsymbol{x}_s = \mathbf{0}. \tag{1.2}
$$

V určitých prípadoch je možné lineárnu závisloť určiť na základe toho, ako jednotlivé vektory vyzerajú. V prípade, že sa medzi skúmanými vektormi nachádza nulový vektor alebo nejaký vektor je násobkom iného, sú tieto vektory lineárne závislé.

1.1.2 Matice

Maticou $\boldsymbol{A}$ typu $m \times n$ budeme rozumieť súbor prvkov:

$$\boldsymbol{A} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \dots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix}, \tag{1.3}$$

$\boldsymbol{A} = (a_{i,j})$, kde $i = 1, \dots, m$ a $j = 1, \dots, n$, $a_{i,j} \in R$.
Vektor $(a_{i1}, a_{i2}, \dots, a_{in})$ je i -tým riadkom a $(a_{1j}, a_{2j}, \dots, a_{mj})^T$ je j -tým stĺpcom matice A. Transponovaná matica A^T typu $n \times m$ je matica, vytvorená vzájomnou výmenou riadkov a stĺpcov matice A. Za platnosti $A = A^T$, hovoríme o matici symetrickej. Štvorcová matica A je špeciálnym typom matice A, kde počet riadkov je rovný počtu stĺpcov.

Matice môžeme sčítať alebo násobiť ľubovoľnou konštantou. Rovnosť matíc A a B nastáva práve vtedy keď $a_{ij} = b_{ij}, \quad i = 1, 2, \dots, m$ a $j = 1, 2, \dots, n$. Matica s prvkami $a_{ij} = 0, \quad i = 1, 2, \dots, m$ a $j = 1, 2, \dots, n$ sa nazýva nulová matica. Prvky a_{ii} tvoria (hlavnú) diagonálu štvorcovej matice A. Štvorcovú maticu A typu $n \times n$, ktorá má mimo diagonálu samé nuly, nazveme diagonálnou maticou. Ak sú ešte navyše na diagonále samé jednotky, tak hovoríme o jednotkovej matici. Štvorcovú maticu A nazveme pozitívne semidefinitnou (definitnou), ak platí $\boldsymbol{x}^T A \boldsymbol{x} \geq 0$ ($\boldsymbol{x}^T A \boldsymbol{x} > 0$) pre $\boldsymbol{x} \neq 0$. Determinant matice A budeme značiť symbolom $|A|$. Viac o determinantoch a vlastnostiach matíc môžeme nájsť v [7].

Kapitola 2

Náhodná veličina a náhodný vektor

Náhoda je niečo, čo vopred nevieme ovplyvniť. V elementárnej teórii pravdepodobnosti pracujeme s pojmom náhodný pokus, ktorého výsledkom sú náhodné javy. My budeme pracovať s elementárnymi javmi ω. Sú to všetky možné, ďalej nezjednodušiteľné výsledky náhodného pokusu. Náhodná veličina potom priradí náhodným javom číselné hodnoty. V tejto kapitole si ukážeme spôsob, prostredníctvom ktorého sa dostaneme k náhodným veličinám pomocou teórie pravdepodobnosti a zavedieme pojmy stredná hodnota a rozptyl náhodnej veličiny. Pokračovať budeme ďalším dôležitým pojmom náhodný vektor, rozoberieme si podrobnejšie rozdelenie pravdepodobnosti a distribučnú funkciu. Teoretické poznatky pochádzajú z [2], [3], [4], [8] a [11].

2.1 Náhodná veličina

Trojicu $(\Omega, \mathcal{A}, P)$ nazveme pravdepodobnostný priestor. Ľubovoľná množina Ω je priestor elemetárnych javov ω, na ktorom je definovaná σ - algebra $\mathcal{A}$ ako neprázdny systém podmnožín priestoru Ω uzavretý na spočítateľné zjednotenie a doplnok. Symbol P predstavuje pravdepodobnostnú mieru. Borelovskú σ - algebru na R označíme $\mathcal{B}_1$. Je to najmenšia σ - algebra obsahujúca otvorené intervaly. Merateľné zobrazenie z $(\Omega, \mathcal{A}, P)$ do $(\psi, \mathcal{B})$ nazveme náhodnou veličinou X, kde ψ je výberový priestor a $\mathcal{B}$ je σ - algebra na ψ. Pre $\psi = R$ a $\mathcal{B} = \mathcal{B}_1$

reprezentuje každá hodnota $X(\omega)$ reálne číslo a platí $\{\omega \in \Omega$: X(ω) $\in$ B $\} \in \mathcal{A}$, $\forall$ B $\in \mathcal{B}_1$. Pravdepodobnostné rozdelenie náhodnej veličiny X je určené systémom $\{P[X \in B], B \in \mathcal{B}_1\}$. Špeciálne máme pre B = (-∞, x] funkciu

$$F_x(x) = P\,[X \leq x] = P\{\omega \in \Omega : X(\omega) \leq x\} \quad x \in R. \tag{2.1}$$

Funkcia $F_x(x)$ sa nazýva distribučná funkcia náhodnej veličiny X. Pozrieme sa na niektoré charakteristiky náhodných veličín. Stredná hodnota náhodnej veličiny X je definovaná predpisom:

$$EX = \int_\Omega X(\omega) dP(\omega). \tag{2.2}$$

Z vlastností integrálu plynie $E(a + bX) = a + bEX$ pre ľubovoľné a,$b \in R$. Obecný moment k-tého rádu je

$$\mu'_k = EX^k, \qquad k = 0, 1, \ldots \tag{2.3}$$

ak existuje integrál na pravej strane. Pre $k = 1$ máme strednú hodnotu.
Číslo

$$\mu_k = E(X - EX)^k, \qquad k = 0, 1, \ldots \tag{2.4}$$

nazveme centrálnym momentom rádu k v prípade, že integrál na pravej strane existuje. Najvýznamnejším centrálnym momentom je μ_2. Nazývame ho rozptyl. Označíme ho σ^2 alebo var$X$. Platí var$(a + bX) = b^2$var$X$ pre ľubovoľné a,$b \in R$. Odmocninou z rozptylu je smerodajná odchýlka.

Vzťah medzi dvoma náhodnými veličinami nám priblíži kovariancia a korelácia. Predpokladajme konečné a nenulové rozptyly náhodných veličín X a Y. Potom kovarianciou týchto náhodných veličín budeme rozumieť

$$cov(X, Y) = E(X - EX)(Y - EY). \tag{2.5}$$

Z kovariancie dvoch normovaných náhodných veličín dostaneme korelačný koeficient

$$\rho(X, Y) = cov(\frac{X - EX}{\sqrt{varX}}, \frac{Y - EY}{\sqrt{varY}}) = \frac{cov(X, Y)}{\sqrt{varX\,varY}}. \tag{2.6}$$

Je možné ukázať, že $\rho(X, Y) = \pm 1$ práve vtedy, keď $Y = a \pm bX$, $b > 0$. Čím je $\rho(X, Y)$ bližšie k 1 respektéve -1, tým je spomenutá lineárna závislosť náhodných veličín X a Y silnejšia.

2.2 Náhodný vektor

Nech R^k je k-rozmerný reálny priestor a $\mathcal{B}_k$ je systém jeho borelovských podmnožín. Náhodným vektorom $\boldsymbol{X}$ nazveme merateľné zobrazenie z $(\Omega, \mathcal{A}, P)$ do $(\psi, \mathcal{B})$, kde $\psi = R_k$ a $\mathcal{B} = \mathcal{B}_k$. Je to stĺpcový vektor o dĺžke k, ktorého zložky sú náhodné veličiny. Značíme $\boldsymbol{X} = (X_1, X_2, ..., X_k)^T$.

Strednou hodnotou náhodného vektora $\boldsymbol{X}$ budeme rozumieť stĺpcový vektor stredných hodnôt jednotlivých náhodných veličín, $E\boldsymbol{X} = (EX_1, EX_2, ..., EX_k)^T$. Platí $E(\boldsymbol{a} + \boldsymbol{B}\boldsymbol{X}) = \boldsymbol{a} + \boldsymbol{B}E\boldsymbol{X}$ pre ľubovoľný m -rozmerný reálny vektor $\boldsymbol{a}$ a ľubovoľnú maticu $\boldsymbol{B}$ typu $m \times k$. Variančnou maticou náhodného vektora $\boldsymbol{X}$ je

$$\boldsymbol{\Sigma} = var\mathbf{X} = \begin{pmatrix} varX_1 & cov(X_1, X_2) & \dots & cov(X_1, X_k) \\ cov(X_2, X_1) & varX_2 & \dots & cov(X_2, X_k) \\ \dots & \dots & \dots & \dots \\ cov(X_k, X_1) & cov(X_k, X_2) & \dots & varX_k \end{pmatrix}. \tag{2.7}$$

Matica $\boldsymbol{\Sigma}$ je symetrická a pozitívne semidefinitvna. Pozitívna semidefinitnosť vyplýva z vlastnosti $\mathrm{var}(\boldsymbol{a} + \boldsymbol{B}\boldsymbol{X}) = \boldsymbol{B}\mathrm{var}(\boldsymbol{X})\boldsymbol{B}^T$ a ďalej z predpokladu, že varZ náhodnej veličiny Z je nezáporné číslo (ak existuje). Teda pre ľubovoľné $\boldsymbol{c} \in R^k$ a pro $\mathrm{Z} = \boldsymbol{c}^T\boldsymbol{X}$, platí $\mathrm{varZ} = \mathrm{var}(\boldsymbol{c}^T\boldsymbol{X}) = \boldsymbol{c}^T var(\boldsymbol{X})\boldsymbol{c} \geq 0$.

Ďalej je možné zaviesť korelačnú maticu náhodného vektora $\boldsymbol{X}$ predpisom

$$\boldsymbol{\Gamma} = cor\mathbf{X} = \begin{pmatrix} \rho(X_1, X_1) & \rho(X_1, X_2) & \dots & \rho(X_1, X_k) \\ \rho(X_2, X_1) & \rho(X_2, X_2) & \dots & \rho(X_2, X_k) \\ \dots & \dots & \dots & \dots \\ \rho(X_k, X_1) & \rho(X_k, X_2) & \dots & \rho(X_k, X_k) \end{pmatrix}. \tag{2.8}$$

Je možné ukázať, že $\rho(X_i, X_i) = 1, \quad i = 1, \dots, k$ a $\mathrm{cor}\boldsymbol{X} = \boldsymbol{A}\ \mathrm{var}\boldsymbol{X}\ \boldsymbol{A}$, kde $\boldsymbol{A} = diag\{\frac{1}{\sqrt{varX_1}}, \dots, \frac{1}{\sqrt{varX_k}}\}$.

2.3 Distribučná funkcia

Na konci tejto kapitoly sa ešte raz vrátime k distribučnej funkcii, ktorú sme definovali v (2.1). V matematickej štatistike sa používajú dva základné typy distribučných funkcií:

- Predpokladajme, že F je funkcia skokov, ktorá má konečný alebo spočítateľný počet bodov nespojitosti a p_j je veľkosť skoku funkcie F v bode x_j, kde $\sum p_j = 1$. Potom postupnosť pravdepodobností

$$p_j = P[X = x_j], \qquad j = 1, 2, ... \tag{2.9}$$

  definuje diskrétne rozdelenie pravdepodobnosti náhodnej veličiny X.

- Nech F je absolutne spojitá a existuje funkcia $f(x)$ taká, že

$$F(x) = \int_{-\infty}^{x} f(t)dt, \tag{2.10}$$

  potom funkcia $f(x)$ sa nazýva hustota a definuje spojité rozdelenie náhodnej veličiny X.

Ak namiesto náhodnej veličiny X uvažujeme náhodný vektor $\boldsymbol{X} = (X_1, X_2, ..., X_k)^T$, definujeme jeho distribučnú funkciu predpisom

$$F_{\mathbf{X}}(x_1, \ldots, x_k) = P[X_1 \leq x_1, \ldots, X_k \leq x_k]. \tag{2.11}$$

- Náhodný vektor s diskrétne rozdelenými zložkami má združené rozdelenie určené pravdepodobnosťami

$$P[X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k], \tag{2.12}$$

  pre všetky možné kombinácie hodnôt $x_1, \ldots, x_k$ jeho zložiek.

- Náhodný vektor so spojito rozdelenými zložkami má hustotu $f_{\mathbf{X}}(x)$, pre ktorú platí

$$F_{\mathbf{X}}(x_1, \ldots, x_k) = \int_{-\infty}^{x_1} \ldots \int_{-\infty}^{x_k} f_{\mathbf{X}}(t_1, \ldots, t_k)dt_1, \ldots, dt_k. \tag{2.13}$$

Marginálne rozdelenie jednotlivých zložiek $X_1, X_2, ..., X_k$ a rozdelenie ľubovoľného podvektora náhodného vektora $\boldsymbol{X}$ je možné odvodiť zo združeného rozdelenia. Pre jednozložkový podvektor X_i platí

$$F_{X_i}(t) = \lim_{x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_k \to \infty} F_{\mathbf{X}}(x_1, \ldots, x_{i-1}, t, x_{i+1}, \ldots, x_k), \tag{2.14}$$

$$P[X_i = t] = \Sigma \ldots \Sigma P[X_1 = x_1, \ldots, X_{i-1} = x_{i-1}, X_i = t, X_{i+1} = x_{i+1}, \ldots, X_k = x_k], \tag{2.15}$$

$$f_{X_i}(t) = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} f_{\mathbf{X}}(x_1, \ldots, x_{i-1}, t, x_{i+1}, \ldots, x_k) dx_1, \ldots, dx_{i-1} dx_{i+1}, \ldots, dx_k. \tag{2.16}$$

Vo vzťahoch (2.15) a (2.16) sčítame, respektíve integrujeme cez všetky možné hodnoty zložiek $x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_k$.

Kapitola 3

Mnohorozmerné dáta

Predmetom tejto kapitoly je pojem dátová matica. Vysvetlíme si jej rôzne úpravy a pozrieme sa na jej členenie. V praktickej časti budeme pracovať s kontingenčnou tabuľkou, ktorú taktiež popíšeme. Najprv si však definujeme mnohorozmerné normálne rozdelenie. Teoretické poznatky pochádzajú z [4], [8] a [11].

3.1 Mnohorozmerné normálne rozdelenie

Jednou z rozoberaných štatistických metód bude diskriminačná analýza. Bližšie bude popísaná v kapitole 4.2. Pre použitie tejto analýzy budeme potrebovať mnohorozmerné normálne rozdelenie, ktoré si teraz definujeme.

Máme náhodný vektor $\boldsymbol{X} = (X_1, X_2, ..., X_k)^T$, vektor $\boldsymbol{\mu} = (\mu_1, \mu_2, ..., \mu_k)^T$ a symetrickú pozitívne semidefinitnú maticu $\boldsymbol{\Sigma}$ typu $k \times k$. Potom náhodný vektor $\boldsymbol{X}$ má k -rozmerné normálne rozdelenie s parametrami ($\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$), ak pre ľubovoľný vektor $\boldsymbol{s} \in$ R platí $\boldsymbol{s}^T\boldsymbol{X} \sim N(\boldsymbol{s}^T\boldsymbol{\mu}, \boldsymbol{s}^T\boldsymbol{\Sigma}\boldsymbol{s})$. V prípade, že matica $\boldsymbol{\Sigma}$ je regulárna, existuje hustota, ktorá má tvar

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{1}{2}k} \mid \boldsymbol{\Sigma} \mid^{\frac{1}{2}}} exp\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\}. \tag{3.1}$$

3.2 Mnohorozmerný náhodný výber

Uvažujme nezávislé a rovnako rozdelené náhodné vektory $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, kde $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ik})^T$ pre $i = 1, \ldots, n$. Nech $\mathrm{E}\boldsymbol{X}_i = \boldsymbol{\mu}$, $\mathrm{var}\boldsymbol{X}_i = \boldsymbol{\Sigma}$ a $\mathrm{cor}\boldsymbol{X}_i = \boldsymbol{\Gamma}$ pre i = 1,..., n.

- Odhadom vektora stredných hodnôt $\boldsymbol{\mu}$ je výberový priemer

$$\overline{\boldsymbol{X}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i \tag{3.2}$$

so složkami

$$\overline{X}_j = \frac{1}{n}\sum_{i=1}^{n} X_{ij}, \quad j = 1, \ldots, k. \tag{3.3}$$

- Odhadom variančnej matice $\boldsymbol{\Sigma}$ je výberová variančná matica

$$\boldsymbol{S} = \frac{1}{n-1}\sum_{i=1}^{n} (\boldsymbol{X}_i - \overline{\boldsymbol{X}})(\boldsymbol{X}_i - \overline{\boldsymbol{X}})^T \tag{3.4}$$

s prvkami

$$s_{rj} = \frac{1}{n-1}\sum_{i=1}^{n} (X_{ir} - \overline{X}_r)(X_{ij} - \overline{X}_j), \quad j, r = 1, \ldots, k. \tag{3.5}$$

- Odhadom korelačnej matice $\boldsymbol{\Gamma}$ je výberová korelačná matica

$$\boldsymbol{C} = \hat{\boldsymbol{A}}\boldsymbol{S}\hat{\boldsymbol{A}}, \tag{3.6}$$

kde $\hat{\boldsymbol{A}} = diag\{\frac{1}{\sqrt{s_{11}}}, \ldots, \frac{1}{\sqrt{s_{kk}}}\}$ s prvkami

$$rij = \frac{s_{ij}}{\sqrt{s_{ii}s_{jj}}}, \qquad i, j = 1, \ldots, k. \tag{3.7}$$

3.3 Dátová matica

Predpokladajme maticu typu $n \times k$, kde riadky určujú nejaké študované objekty a stĺpce zisťované znaky na týchto objektoch. Označíme ju

$$\boldsymbol{D} = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1k} \\ x_{21} & x_{22} & \ldots & x_{2k} \\ \vdots & \vdots & \ldots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{nk} \end{pmatrix}, \tag{3.8}$$

kde prvok x_{ij} predstavuje hodnotu j -tého znaku na i -tom objekte, $i = 1, 2, \ldots, n$ a $j = 1, 2, \ldots, k$. Hodnoty x_{ij} môžeme taktiež nazvať pozorovaniami či meraniami. Riadky matice $\boldsymbol{D}$ sú k -rozmerné vektory $\boldsymbol{x}_i$. Nazývame ich viacrozmernými pozorovaniami. Pozorovania sú konkrétne realizácie náhodných veličín X_{ij}, respektíve náhodných vektorov $\boldsymbol{X}_i$. Riadky matice $\boldsymbol{D}$ sú teda realizáciou k -rozmerného náhodného výberu $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$.

Dátová matica sa rozpadá horizontálne pri niektorých štatistických metódach, kde pozorujeme situácie rozčlenenia objektov do niekoľkých podmnožín. Jednotlivé objekty v danej množine majú podobné vlastnosti. Takéto rozčlenenie je niekedy dané a cieľom je analýza jednotlivých skupín (analýza rozptylu, môžeme nájsť v [4]) alebo je cieľom samotné rozčlenenie (zhluková analýza). Zhluková analýza bude popísaná v kapitole 4.1. Ak sú sledované znaky rôznej povahy alebo merané v rôznych jednotkách, doporučuje sa aplikovať zhlukovú analýzu na štandardizované dáta

$$x^*_{ij} = \frac{x_{ij} - \bar{x}_j}{\sqrt{s_{jj}}} \quad i = 1, 2, \ldots, n \quad , j = 1, 2, \ldots, k \tag{3.9}$$

kde $\bar{x}_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij}$ je priemerná hodnota j -tého znaku cez všetky namerané realizácie a $\sqrt{s_{jj}}$ je výberová smerodatná odchýlka, teda prvok diagonály výberovej variančnej matice (3.4). Štandardizované dáta sú realizáciou náhodného výberu s nulovou strednou hodnotou a jednotkovým rozptylom.

3.4 Objekty a typy premenných

Záznamy o objektoch tvoria riadky dátovej matice. Pri použití rôznych štatistických metód predpokladáme primerane rozsiahly súbor objektov. V odbornej literatúre sa objekt nazýva aj ako prvok, jednotka respektíve štatistická jednotka a znak ako premenná. Na každom z n objektov pozorujeme realizáciu k-rozmerného náhodného vektora. Poznávanie závislosti medzi jednotlivými objektami prostredníctvom meraných znakov je typickým cieľom štatistickej analýzy. Klasifikácia premenných zo štatistického hľadiska je podrobnejšie popísaná napríklad v knihe [5].

Znaky je možné klasifikovať z rôznych hľadísk. Základné delenie je:

- *Kvalitatívne*, kde nie je možné jednoznačné vyjadrenie číslom
- *Kvantitatívne*, ktoré sa dajú vyjadriť jednoznačne číselne

Ďalšou charakteristikou premenných je počet hodnôt, ktoré môžu nadobúdať. Rozlišujeme premenné :

- *Spojité*
- *Nespojité*

Spojité nadobúdajú nespočetne mnoho hodnôt. Najčastejšie vznikajú vážením a meraním objektov. Majú teda číselné vyjadrenie. V prípade vzniku nespojitých resp. diskrétnych premenných dochádza k nadobúdaniu spočítateľne alebo konečne mnoho hodnôt. Tieto premenné delíme na nominálne (ak nemá zmysel porovnávať ich hodnoty, respektíve zoradiť hodnoty podľa veľkosti), ordinálne (ak má zmysel ich hodnoty porovnávať). Poslednou skupinou nespojitých premenných sú alternatívne (inak binárne alebo dichotomické), ktoré nadobúdajú len dve hodnoty.

Varianty znakov sú hodnoty, ktoré môžu jednotlivé diskrétne znaky nadobúdať. V prípade jedného študovaného znaku ($k = 1$) s malým počtom hodnôt vytvoríme tabuľku, ktorá bude obsahovať všetky varianty a početnosť ich výskytu v súbore n objektov. Početnosť l-tej varianty značíme n_l pre $l = 1, 2, \ldots, r$ a pre celý rozsah súboru n je

$$n = \sum_{l=1}^{r} n_{l.}. \tag{3.10}$$

Hore popísaná tabuľka sa nazýva tabuľka rozdelenia početností a proces jej vytvárania je triedenie prvého stupňa resp. jednorozmerné triedenie. Pre $k = 2$ vznikne kontingenčná tabuľka s dvomi výstupmi. Kontingenčná tabuľka pre znaky Y, Z vyzerá následovne:

Y	Z	$\sum$
	$1 \ldots c$	
1	$n_{11} \ldots n_{1c}$	$n_{1.}$
$\ldots$	$\ldots$	$\ldots$
r	$n_{r1} \ldots n_{rc}$	$n_{r.}$
$\sum$	$n_{.1} \ldots n_{.c}$	n

Tabuľka 3.1: Kontingenčná tabuľka

Udáva prehľad variánt prvého a druhého znaku. Riadky zodpovedajú možným hodnotám prvého znaku a sĺpce zobrazujú hodnoty druhého znaku. V bunkách tabuľky nájdeme pozorované početnosti kombinácií hodnôt znakov Y a Z. Na krajoch sa nachádazajú pozorované marginálne početnosti definovanej ako riadkový a stĺpcový súčet. Celkový rozsah súboru je

$$n = \sum_{i=1}^{r} n_{i.} = \sum_{j=1}^{c} n_{.j} = \sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij}. \tag{3.11}$$

Tento proces vytvárania nazývame triedenie druhého druhu alebo dvojrozmerné triedenie. Nezávislosť náhodných veličín Y a Z nastáva práve vtedy, keď

$$p_{ij} = p_{i.}p_{.j} \quad pre \quad i = 1, 2, \ldots, r, \quad j = 1, 2, \ldots, c \tag{3.12}$$

Hypotézu H_0 nezávislosti Y a Z je môžné testovať pomocou kontingenčnej tabuľky. Za platnosti H_0 má veličina

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - \frac{n_{i.}n_{.j}}{n})^2}{\frac{n_{i.}n_{.j}}{n}} \tag{3.13}$$

asymptoticky rozdelenie chí-kvadrát s (r - 1)(c - 1) stupňami voľnosti. Ak prekročí testová štatistika χ^2 (1 - α)-kvantil rozdelenia chí-kvadrát s (r - 1)(c - 1) stupňami voľnosti, zamietame H_0 na hladine α. Aproximácia pomocou chí-kvadrát rozdelenia v praxi dáva rozumné výsledky, ak pre teoretické početnosti platí

$$\frac{n_{i.}n_{.j}}{n} > 5 \quad pre \quad i = 1, 2, \ldots, r \quad , j = 1, 2, \ldots, c. \tag{3.14}$$

Odvodenie testu nezávislosti alebo informácie o p -rozmernej kontingenčnej tabuľke nájdeme v [4], [9].

Kapitola 4

Mnohorozmerné štatistické metódy

Táto kapitola sa zaoberá dvoma prístupmi k spracovaniu mnohorozmerných dát. Prvým z nich bude zhluková analýza, ktorá triedi objekty do skupín s podobnými vlastnosťami z hľadiska sledovaných znakov. Problém zaradenia nového objektu do už vybraných skupín rieši diskriminačná analýza, ktorá je druhou metódou rozoberanou v nasledujúcom texte. Teoretické poznatky pochádzajú z [1], [4] a [8].

4.1 Zhluková analýza

Pri zhlukovej analýze skúmame n objektov. Sú charakterizované k znakmi či už diskrétnej alebo spojitej povahy. Cieľom je nájsť skupiny s podobnými objektami. Budeme predpokladať štandardizované dáta (3.9), ktoré mám vytvoria dátovú maticu $\boldsymbol{D}$ zavedenú v (3.8).

Zhluky podobných objektov budeme hľadať pomocou vzdialeností. Uvažujeme i -tý a j -tý objekt, ktorým prislúchajú jednotlivé riadky $\boldsymbol{x}_i$ a $\boldsymbol{x}_j$ matice $\boldsymbol{D}$. Dva základné postupy výpočtu vzdialenosti dvoch objektov sú :

- **Euklidovská vzdialenosť**

$$d_E(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sqrt{\sum_{r=1}^{k} (x_{ir} - x_{jr})^2}. \tag{4.1}$$

- **Manhattanská vzdialenosť**

$$d_M(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sum_{r=1}^{k} \mid x_{ir} - x_{jr} \mid . \tag{4.2}$$

V závislosti na type dát sa rozhodujeme, ktorú vzdialenosť použijeme. Ak sú merané znaky nezávislé a normálne rozdelené s rovnakými rozptylmi, tak použijeme euklidovskú vzdialenosť. Toto je doporučené i pre prípad vecne podobných znakov, ktoré sú rovnako dôležité pri klasifikácii objektov. Pri existencii odľahlých pozorovaní, v prípade porušenia podmienky normality rozdelenia vektora znakov je vhodnejšou voľbou Manhattanská vzdialenosť. Tú je možné zvoliť i v prípade diskrétnych celočíselných znakov.

Následujúce metódy popisujú vzájomnú vzdialenosť zhlukov S_1 a S_2 :

- **Metóda najbližšieho suseda**

$$D_{min}(S_1, S_2) = min_{\boldsymbol{x}_i \in S_1, \boldsymbol{x}_j \in S_2} d(\boldsymbol{x}_i, \boldsymbol{x}_j), \tag{4.3}$$

- **Metóda najvzdialnejšieho suseda**

$$D_{max}(S_1, S_2) = max_{\boldsymbol{x}_i \in S_1, \boldsymbol{x}_j \in S_2} d(\boldsymbol{x}_i, \boldsymbol{x}_j), \tag{4.4}$$

- **Metóda priemernej vzdialenosti**

$$D_{priem}(S_1, S_2) = d(\overline{\boldsymbol{x}}_1, \overline{\boldsymbol{x}}_2), \tag{4.5}$$

kde d je miera vzdialenosti objektov a
$S_1 = \{\boldsymbol{x}_{i_1}, \boldsymbol{x}_{i_2}, \ldots, \boldsymbol{x}_{i_r}\}$,
$S_2 = \{\boldsymbol{x}_{i_{r+1}}, \boldsymbol{x}_{i_{r+2}}, \ldots, \boldsymbol{x}_{i_m}\} \quad pre \quad 1 \leq r < m \leq n$,
$\overline{\boldsymbol{x}}_1 = \frac{1}{r} \sum_{j=1}^{r} \boldsymbol{x}_{i_j}$,
$\overline{\boldsymbol{x}}_2 = \frac{1}{m-r} \sum_{j=r+1}^{m} \boldsymbol{x}_{i_j}$.
Zhlukovacie algoritmy delíme na

- **Hierarchické**
- **Nehierarchické**

4.1.1 Hierarchická aglomeratívna metóda

Najprv popíšeme hierarchickým aglomeratívnym algoritmom. Pre vytváranie zhlukov je potrebné mať k dispozícii maticu vzdialeností $\mathbf{\Delta}$, ktorej prvky sú $d_{ij} = D(S_i, S_j)$.

Aglomeratívny algoritmus má nasledujúci priebeh:

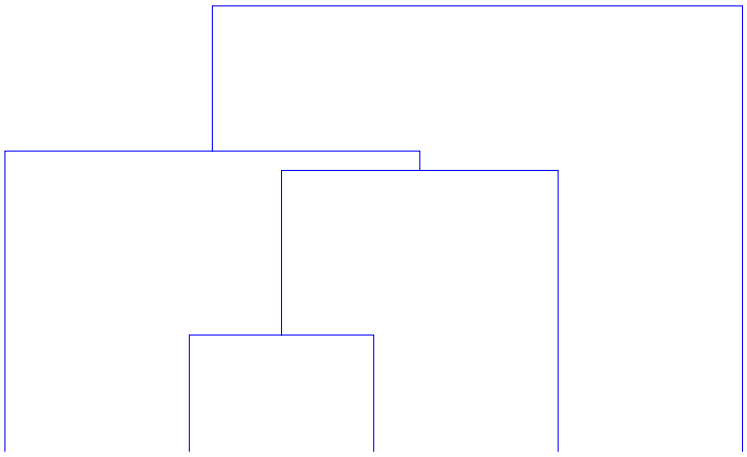
Počiatok: Každý objekt tvorí samostatný zhluk.

Krok 1 : Určíme maticu vzdialenstí objektov $\mathbf{\Delta^{(1)}}$, nájdeme ${d_{rs}}^{(1)} = min({d_{ij}}^{(1)}; i, j = 1, 2, \ldots, n)$ a spojíme objekty r, s do skupiny.

Krok t : Pre $1 < t < n$ analogicky určíme maticu vzdialeností objektov a zhlukov $\mathbf{\Delta^{(t)}}$, nájdeme ${d_{rs}}^{(t)} = min({d_{ij}}^{(t)}; i, j = 1, 2, \ldots, n - t + 1)$ a spojíme zhluky r, s do novej skupiny.

Koniec nastáva, ak sa všetky objekty nachádzajú v jednom zhluku. Algoritmus je možné ukončiť skôr.

Priebeh algoritmu je zachytený grafom nazvaným dendrogram.



Obr. 4.1: Jednoduchý príklad dendrogramu

Na obrázku 4.1 máme zobrazenú jednoduchú simuláciu hierarchického algoritmu vytvorenú pomocou softvéru Mathematica. Na začiatku pozorujeme 5 objektov (5 bodov na spodnej časti obrázku), ktoré sú postupom času na základe euklidovskej vzdialenosti spájané do zhlukov. Očíslujeme ich od 1 do 5 zľava doprava. Prvým spojením, respetíve najbližie sú k sebe body 2 a 3. Najvzdialenejším bodom a posledným pridaným je 5. Pri pohľade na dendrogram môžeme rozhodnúť o skoršom ukončení algoritmu. Vzdialenosť určujú vertikálne čiary.

Vidíme, že piaty objekt je ďaleko od ostaných. Do zhluku ho nezaradíme a algoritmus ukončíme po pridaní objektu 1 do zhluku. Tímto spôsobom odlúčime netypický objekt 5 od skupiny ostatných podobných objektov. Zhlukovanie je však možné ukončiť aj pred pridaním objektu 4 do zhluku (2,3), lebo vzdialenosť objektu 4 od zhluku (2,3) je výrazne väčšia ako vzdialenosť objektov 2 a 3. Okamih ukončenia algoritmu je záležitosťou subjektívnej voľby analytika. Týmto dôjde k rozdeleniu dendrogramu na dve časti a to na realizované spojenia a neuskutočnené spojenia. Rozdelenie je určené priamkou vedenou v úrovni, ktorej súradnicu napríklad v programe NCSS nazývame cluster cutoff.

4.1.2 Nehierarchická metóda

Pre nehierarchické algoritmy máme k dispozícii predom určený počet skupín, do ktorých budeme objekty triediť. Každá skupina má svojho zástupcu. Nazýva sa centroid. Môže ním napríklad byť aritmetický priemer (3.2) vektorov reprezentujúcich jednotlivé objekty v skupine.

Jednou z používaných metód v rámci tohto prístupu je metóda K - means clustering, ktorá s vopred určeným počtom zhlukov vytvára optimálny rozklad objektov. Je založená na minimalizácii súčtov štvorcov $K^{(q)}$ cez všetky možné usporiadania objektov do skupín. Ak predpokladáme usporiadanie do h zhlukov, potom je

$$K^{(q)} = \sum_{r=1}^{h} \sum_{i:x_i \in P_r{}^{(q)}} \sum_{j=1}^{k} (x_{ij} - c^{(q)}{}_{rj})^2, \tag{4.6}$$

kde q je číslo usporiadania objektov do zhlukov,
$P^{(q)}{}_r$ znamená r -tý zhluk pri q-tom usporiadaní objektov,
x_{ij} predstavuje hodnotu j-tého znaku v i -tom objekte,
$c^{(q)}{}_{rj}$ určuje hodnotu j-tého znaku v centroide r-tého zhluku $P^{(q)}{}_r$.

4.2 Diskriminačná analýza

Diskriminačná analýza sa zaoberá zaradením nového objektu do už existujúcich skupín, ktoré boli vytvorené napríklad na základe podobných hodnôt znakov zhlukovou analýzou. Znova predpokladáme dátovú maticu definovanú

predpisom (3.8) a $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ik})^T$, $i = 1, 2, \ldots, n$ je i -tý objekt s nameranými k znakmi. Objekty sú potom zaradené do tried $T_1, T_2, \ldots, T_p, p < n$ a tým je vytvorená testovacia databáza. Pre nový objekt $\boldsymbol{x}_i$, $i > n$ vytvoríme rozhodovacie pravidlo, diskriminačný skór, na základe ktorého tento nový objekt zaradíme do niektorej triedy. Pre skupiny $T_1, T_2, \ldots, T_p$ definujeme diskriminačné skóry $D_1^{(i)}, D_2^{(i)}, \ldots, D_p^{(i)}$ ako funkcie znakov nameraných na i -tom objekte, ktoré minimalizujú strednú hodnotu straty pri zaradení objektu do nesprávnej triedy. Stratu označíme z_{rs}, ak objekt prislúcha r -tej skupine a zaradíme ho do s -tej skupiny. Prepokladajme spojité znaky. Za podmienky, že objekt patrí do r -tej skupiny, je stredná hodnota straty

$$L_r = z_{r1} \int_{A_1} f_r(\boldsymbol{x}) d\boldsymbol{x} + \ldots + z_{rp} \int_{A_p} f_r(\boldsymbol{x}) d\boldsymbol{x} = \sum_{s=1}^{p} z_{rs} \int_{A_s} f_r(\boldsymbol{x}) d\boldsymbol{x}, \tag{4.7}$$

kde f_r je hustota vektora znakov $\boldsymbol{x}$ v r -tej triede a A_d pre $d = 1, 2, \ldots, p$ je rozklad k -rozmerného reálneho priestoru R^k do disjunktných borelovských množín. Predpokladajme, že $\pi_r > 0$ je pravdepodobnosť, že objekt patrí do r -tej triedy. Potom nepodmienená stredná hodnota straty je

$$L = \pi_1 L_1 + \ldots + \pi_p L_p. \tag{4.8}$$

Ak definujeme

$$q_s(\boldsymbol{x}) = \sum_{r=1}^{p} \pi_r z_{rs} f_r(\boldsymbol{x}), \quad s = 1, \ldots, p, \tag{4.9}$$

potom (4.8) je možné upraviť pomocou (4.7) a (4.9) na

$$\begin{aligned}
L &= \pi_1 L_1 + \ldots + \pi_p L_p \\
&= \pi_1 \sum_{s=1}^{p} z_{1s} \int_{A_s} f_1(\boldsymbol{x}) d\boldsymbol{x} + \ldots + \pi_p \sum_{s=1}^{p} z_{ps} \int_{A_s} f_p(\boldsymbol{x}) d\boldsymbol{x} \\
&= \sum_{r=1}^{p} \pi_r \sum_{s=1}^{p} z_{rs} \int_{A_s} f_r(\boldsymbol{x}) d\boldsymbol{x} \\
&= \sum_{s=1}^{p} \sum_{r=1}^{p} \int_{A_s} \pi_r z_{rs} f_r(\boldsymbol{x}) d\boldsymbol{x} \\
&= \sum_{s=1}^{p} \int_{A_s} \sum_{r=1}^{p} \pi_s z_{rs} f_r(\boldsymbol{x}) d\boldsymbol{x} \\
&= \sum_{s=1}^{p} \int_{A_s} q_s(\boldsymbol{x}) d\boldsymbol{x}.
\end{aligned} \tag{4.10}$$

Hľadáme optimálny rozklad priestoru R^k, aby bolo L minimálne. Nech ${A_1}^*, \ldots, {A_p}^*$ je rozklad priestoru R^k, pre ktorý platí

$$\boldsymbol{x} \in {A_t}^* \Rightarrow q_t(\boldsymbol{x}) \leq q_s(\boldsymbol{x}), s = 1, \ldots, p. \tag{4.11}$$

Ak označíme

$$L^* = \sum_{s=1}^{p} \int_{A^* s} q_s(\boldsymbol{x}) d\boldsymbol{x}, \tag{4.12}$$

potom platí

$$L \geq L^*. \tag{4.13}$$

Teraz sa bližie pozrieme na túto nerovnosť:

$$\begin{aligned} L = \sum_{s=1}^{p} \int_{As} q_s(\boldsymbol{x}) d\boldsymbol{x} = \sum_{s=1}^{p} \sum_{t=1}^{p} \int_{As \cap A^* t} q_s(\boldsymbol{x}) d\boldsymbol{x} = \sum_{t=1}^{p} \sum_{s=1}^{p} \int_{A^* t \cap As} q_s(\boldsymbol{x}) d\boldsymbol{x} \geq \\ \geq \sum_{t=1}^{p} \sum_{s=1}^{p} \int_{A^* t \cap As} q_t(\boldsymbol{x}) d\boldsymbol{x} = \sum_{t=1}^{p} \int_{A^* t} q_t(\boldsymbol{x}) d\boldsymbol{x} = L^*. \end{aligned} \tag{4.14}$$

Rovnosti v (4.14) sú zrejmé a nerovnosť vyplýva z (4.11).
Špeciálne volíme $z_{ss} = 0$ a $z_{rs} = 1$ pre $r \neq s$. Ak označíme $\mathrm{c}(\boldsymbol{x}) = \sum_{r=1}^{p} \pi_r f_r(\boldsymbol{x})$, dostávame $q_s(\boldsymbol{x}) = c(\boldsymbol{x}) - \pi_s f_s(\boldsymbol{x})$. Minimalizáciou podielu chybne zaradených objektov dostaneme minimálne L. Pri danom $\boldsymbol{x}$ a t platí $q_t(\boldsymbol{x}) \leq q_s(\boldsymbol{x}) \Leftrightarrow \pi_t f_t(\boldsymbol{x}) \geq \pi_s f_s(\boldsymbol{x})$ pre $s = 1, \ldots, p$. Ak nájdeme taký vektor, pre ktorý platí $\pi_t f_t(\boldsymbol{x}) > \pi_s f_s(\boldsymbol{x}), \quad s \neq t$, potom tento objekt reprezetnovaný vektorom $\boldsymbol{x}$ zaradíme do skupiny t. Ak nastáva rovnosť, môžeme vybrať akýkoľvek z maximalizujúcich indexov.

Teraz dosadíme za $f_s(\boldsymbol{x})$ hustotu k -rozmerného normálneho rozdelenia (3.1) so strednou hodnotou $\boldsymbol{\mu_s}$ a regulárnou variančnou maticou $\boldsymbol{\Sigma_s}$ pre $s = 1, \ldots, p$. Nerovnosť $\pi_t f_t(\boldsymbol{x}) > \pi_s f_s(\boldsymbol{x}), s \neq t$ je ekvivalentná s nerovnosťou

$$ln\pi_t + lnf_t > ln\pi_s + lnf_s, \quad s \neq t. \tag{4.15}$$

Ak označíme

$$D_s = -\frac{1}{2} ln|\boldsymbol{\Sigma_s}| - \frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu_s})^T {\boldsymbol{\Sigma_s}}^{-1} (\boldsymbol{x} - \boldsymbol{\mu_s}) + ln\pi_s, \tag{4.16}$$

tak nerovnosť (4.15) platí práve vtedy keď $D_t > D_s$ pre $s \neq t$. Nový objekt x_{n+1} zaradíme do triedy T_t ak

$$D_t^{(n+1)} = max(D_s^{(n+1)}, s = 1, 2, \ldots, p) \tag{4.17}$$

Za odhad $\widehat{\pi_s}$ pravdepodobnosti π_s je možné vziať relatívnu početnosť $\frac{n_s}{n}$, kde n_s je počet objektov testovacej databáze v triede T_s a $n_1 + n_2 + \ldots + n_p = n$. Ďalšou možnou voľbou je $\widehat{\pi_s} = \frac{1}{p}$ pre $s = 1, 2, \ldots, p$. Priemer meraných znakov v triede T_s je

$$\overline{\boldsymbol{x}}_s = \frac{1}{n_s} \sum_{k:\boldsymbol{x}_k \in T_s} \boldsymbol{x}_k. \tag{4.18}$$

Pre variančnú maticu znakov v s -tej triede je odhadom výberová variančná matica

$$\boldsymbol{S}_s = \frac{1}{n_s - 1} \sum_{k:\boldsymbol{x}_k \in T_s} (\boldsymbol{x}_k - \overline{\boldsymbol{x}}_s)(\boldsymbol{x}_k - \overline{\boldsymbol{x}}_s)^T. \tag{4.19}$$

Predpokladajme normálne rozdelenie vektorov znakov, potom môžeme hovoriť o dvoch typoch diskriminačných skórov:

- **Kvadratický diskriminačný skór** počítaný pre objekt $\boldsymbol{x}_i$ a triedu T_s, ktorý vznikne dosadením odhadov (4.18) a (4.19) do (4.16)

$$D_s^{(i)} = -\frac{1}{2} ln|\boldsymbol{S}_s| - \frac{1}{2}(\boldsymbol{x}_i - \overline{\boldsymbol{x}}_s)^T \boldsymbol{S}_s^{-1}(\boldsymbol{x}_i - \overline{\boldsymbol{x}}_s) + ln\widehat{\pi_s}. \tag{4.20}$$

Kvadratický diskriminačný skór určuje obecný prístup. Ak sa triedy odlišujú len strednou hodnotou a majú rovnakú variančnú maticu, používa sa

- **Lineárny diskriminačný skór**

$$D_s^{(i)} = \overline{\boldsymbol{x}}_s^T \boldsymbol{S}^{-1} \boldsymbol{x}_i - \frac{1}{2} \overline{\boldsymbol{x}}_s^T \boldsymbol{S}^{-1} \overline{\boldsymbol{x}}_s + ln\widehat{\pi_s}, \tag{4.21}$$

pre s - tu triedu a i -tý objekt, kde
$\boldsymbol{S} = \frac{1}{n-p} \sum_{j=1}^{p} (n_j - 1)\boldsymbol{S}_j$.

V knihe [4] sa lineárna diskriminačná analýza nedoporučuje pre prípad viac ako dvoch tried objektov.

Kapitola 5

Spracované dáta

V praktickej časti budeme demonštrovať použitie popísaných metód na voľne dostupných dátach istej nemeckej banky [10], pričom konkrétne budeme pracovať s databázou žiadateľov o úver v počte 1000. Celkovo táto databáza tvorí dátovú maticu typu (3.8). Objektami sú klienti banky s jednotlivými priradenými znakmi, ktoré sú diskrétneho alebo spojitého charakteru. V nasledujúcich podkapitolách budú jednotlivé znaky popísané.

5.1 Kontingenčná tabuľka

Najprv sa budeme zaoberať testom nezávislosti v kontingenčných tabuľkách (tabuľka 3.1) pre tri diskrétne znaky. Vybranými znakmi sú bilancia na účte, informácia o pôvode a vlastníctvo. Pod jednotlivými znakmi sa nachádzajú kategórie, do ktorých sú klienti rozdelení:

- *Bilancia na účte*
  - Žiadna bilancia alebo dlh
  - Kladná bilancia
  - Žiaden účet
- *Cudzinec*
  - Áno
  - Nie

- *Vlastníctvo*
  - Dom alebo pozemok
  - Životná poistka
  - Auto
  - Iné

Nasleduje prehľad kontingenčných tabuliek pozorovaných početností n_{ij}, teoretických početností $\frac{n_{i.}n_{.j}}{n}$ a hodnôt jednotlivých sčítancov veličín χ^2 zo vzorca (3.13).

	Vlastníctvo				
Cudzinec	**Dom alebo pozemok**	**Životná poistka**	**Auto**	**Nič**	Celkom
Áno	19	14	2	2	**37**
Nie	263	218	330	152	**963**
Celkom	**282**	**232**	**332**	**154**	

Tabuľka 5.1: Kontingenčná tabuľka pozorovaných početností znakov Cudzinec a Vlastníctvo

	Vlastníctvo				
Cudzinec	**Dom alebo pozemok**	**Životná poistka**	**Auto**	**Iné**	Celkom
Áno	10,43	8,58	12,28	5,70	37
Nie	271,57	223,42	319,72	148,30	963
Celkom	282	232	332	154	

Tabuľka 5.2: Teoretické početnosti znakov Cudzinec a Vlastníctvo

	Vlastníctvo				
Cudzinec	**Dom alebo pozemok**	**Životná poistka**	**Auto**	**Nič**	Celkom
Áno	7,03	3,42	8,61	2,40	**21,46**
Nie	0,27	0,13	0,33	0,09	**0,82**
Celkom	**7,30**	**3,55**	**8,94**	**2,49**	**22,28**

Tabuľka 5.3: Hodnoty sčítancov χ^2 rozdelenia znakov Cudzinec a Vlastníctvo

	Bilancia na účte			
Vlastníctvo	**Žiaden dlh alebo zostatok**	**Kladný zostatok**	**Žiaden účet**	Celkom
Dom alebo pozemok	70	75	137	**282**
Životná poistka	80	55	97	**232**
Auto	74	92	166	**332**
Nič	50	47	57	**154**
Celkom	**274**	**269**	**457**	

Tabuľka 5.4: Kontingenčná tabuľka pozorovaných početností znakov Bilancia na účte a Vlastníctvo

	Bilancia na účte			
Vlastníctvo	**Žiaden dlh alebo zostatok**	**Kladný zostatok**	**Žiaden účet**	Celkom
Dom alebo pozemok	77,27	75,86	128,87	**282**
Životná poistka	63,57	62,41	106,02	**232**
Auto	90,97	89,31	151,72	**332**
Nič	42,20	41,43	70,38	**154**
Celkom	**274**	**269**	**457**	

Tabuľka 5.5: Teoretické početnosti znakov Bilancia na účte a Vlastníctvo

	Bilancia na účte			
Vlastníctvo	**Žiaden dlh alebo zostatok**	**Kladný zostatok**	**Žiaden účet**	Celkom
Dom alebo pozemok	0,68	0,01	0,51	**1,21**
Životná poistka	4,25	0,88	0,77	**5,89**
Auto	3,16	0,08	1,34	**4,59**
Nič	1,44	0,75	2,54	**4,74**
Celkom	**9,54**	**1,72**	**5,17**	**16,43**

Tabuľka 5.6: Hodnoty sčítancov χ^2 rozdelenia znakov Bilancia na účte a Vlastníctvo

	Bilancia na účte			
Cudzinec	**Žiaden dlh alebo zostatok**	**Kladný zostatok**	**Žiaden účet**	Celkom
Áno	15	6	16	**37**
Nie	259	263	441	**963**
Celkom	**274**	**269**	**457**	

Tabuľka 5.7: Kontingenčná tabuľka pozorovaných početností znakov Bilancia na účte a Cudzinec

	Bilancia na účte			
Cudzinec	**Žiaden dlh alebo zostatok**	**Kladný zostatok**	**Žiaden účet**	Celkom
Áno	10,14	9,95	16,91	**37**
Nie	263,86	259,05	440,09	**963**
Celkom	**274**	**269**	**457**	

Tabuľka 5.8: Teoretické početnosti znakov Bilancia na účte a Cudzinec

	Bilancia na účte			
Cudzinec	**Žiaden dlh alebo zostatok**	**Kladný zostatok**	**Žiaden účet**	Celkom
Áno	2,33	1,57	0,05	**3,95**
Nie	0,09	0,06	0,00	**0,15**
Celkom	**2,42**	**1,63**	**0,05**	**4,10**

Tabuľka 5.9: Hodnoty sčítancov χ^2 rozdelenia znakov Bilancia na účte a Cudzinec

V tabuľkách 5.3, 5.6 a 5.9 nájdeme sčítance χ^2 rozdelenia daných znakov a v pravom dolnom rohu hodnotu testovej štatistiky χ^2 zo vzorca (3.13). Hladina testu je 0.05. Pri dvojici znakov Cudzinec a Vlastníctvo porovnávame hodnotu $\chi^2 = 22{,}28$ s 0,95-kvantilom χ^2 rozdelenia s troma stupňami voľnosti, ktorá je $\chi^2 = 7{,}82$. Testová štatistika prekračuje kvantil, teda zamietame nezávislosť znakov. Rovnako je to pri druhej dvojici znakov Vlastníctvo a Bilancia na účte. Testovú štatistiku $\chi^2 = 16{,}43$ porovnávame s 0,95-kvantilom χ^2 rozdelenia o šesť stupňov voľnosti 12,59. V poslednom prípade k prekročeniu 0,95-kvantilu nedôjde. Testová štatistika $\chi^2 = 4{,}1$ je menšia ako 0,95-kvantil χ^2 rozdelenia s dvoma stupňami voľnosti 5,99. Takže na základe našich dát nemôžeme zamietnuť nezávislosť znakov Cudzinec a Bilancia na účte.

Závislosť spojitých znakov bude vyšetrená na základe korelačnej matice (3.6). Spojité znaky, ktoré budeme posudzovať, sú :

- *Vek*
- *Výška úveru*
- *Doba splatnosti úveru*

Doba splatnosti úveru je v intervale od 4 mesiacov do 72 mesiacov. Výška úveru sa pohybuje v rozmedzí 250 - 18 424 nemeckých mariek a vek klientov je od 19 do 75 rokov.

Korelačná matica spojitých znakov je

$$\begin{pmatrix} 1. & -0.04 & 0.03 \\ -0.04 & 1. & 0.62 \\ 0.03 & 0.62 & 1. \end{pmatrix}, \tag{5.1}$$

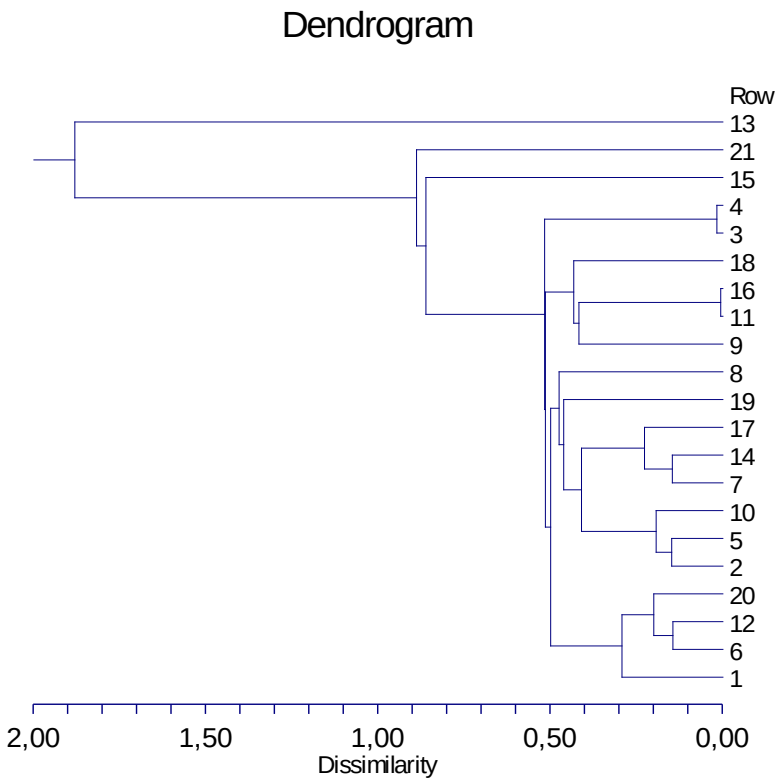
kde stĺpce, respektíve riadky reprezentujú spojité znaky v poradí Vek, Výška úveru a Doba splatnosti úveru. Najsilnejšia lineárna závislosť je medzi znakom Výška úveru a Doba splatnosti úveru, čo odráža úverovú poliku bánk. Korelácia medzi znakmi Vek a Doba splatnosti úveru a aj Vek a Výška úveru je veľmi slabá.

5.2 Zhluková analýza

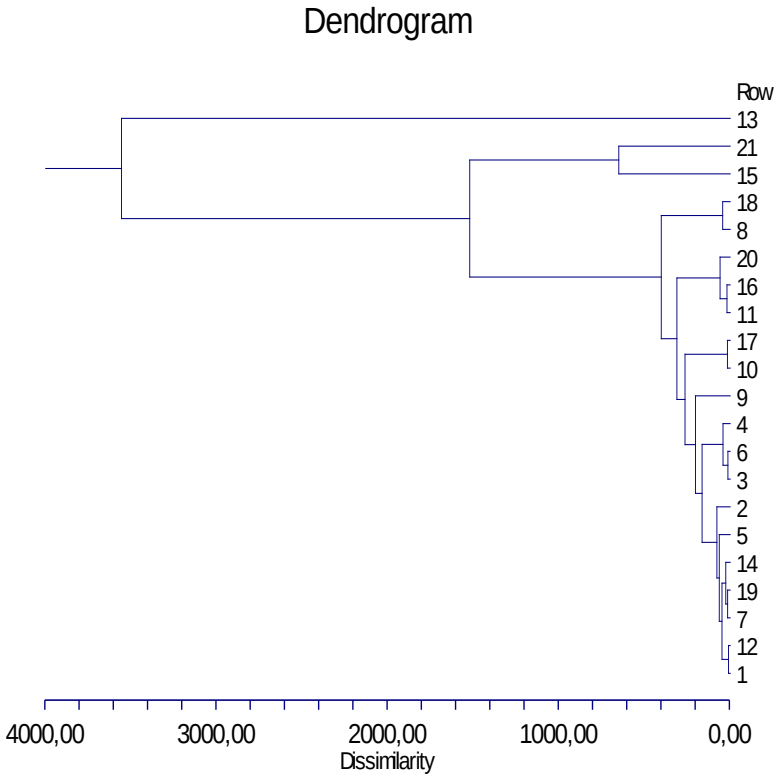
V tejto časti budeme aplikovať zhlukovú analýzu popisovanú v kapitole 4.1 na tri spojité znaky. Budeme používať rovnaké spojité znaky ako v kapitole 5.1. Vzhľadom k lepšej názornosti bude dátová matica obsahovať obmedzený počet klientov. V tejto kapitole aplikujeme na dáta hierarchickú aglomeratívnu metódu a nehierarchickú zhlukovú metódu (K-means clustering). Použitými softvérami budú Mathematica, NCSS a MS Excel.

5.2.1 Hierarchická aglomeratívna metóda

Sledovanými znakmi sú Vek, Doba splatnosti úveru a Výška úveru. Budeme sa zaoberať klientami, ktorí sú viac ako 60-roční a menej ako 64-roční. Počet klientov sme znížili na 21. Celkovo máme dátovú maticu s tromi znakmi a dvadsaťjeden objektami. Zhluková analýza bude aplikovaná na štadardizovanú a neštandardizovanú dátovú maticu s euklidovskou respektíve manhattanskou vzdialenosťou. Ako metódu na určenie vzájomnej vzdialenosti zhlukov sme zvolili metódu najbližšieho suseda. Vytvorením dendrogramov v softvére NCSS bolo zistené, že voľba vzdialenosti objektov nemá vplyv na tvar dendrogramu. Uvedieme dendrogramy s euklidovskou vzdialenosťou na štandardizované a neštandardizované dáta:



Obr. 5.1: Dendrogram štandardizovaných dát



Obr. 5.2: Dendrogram neštandardizovaných dát

Čísla v pravej časti dendrogramov predstavujú riadky v dátovej matici.

Z dendrogramu usúdime, že algoritmus je vhodné ukončiť pri štandardizovaných dátach napríklad voľbou cluster cutoff = 0.6 a pri neštandardizovaných dátach je možné použiť cluster cutoff = 1000. Ak vztýčime v bodoch cluster cutoff na vodorovnej ose zvislú priamku, rozdelíme dendrogram tak, že spojenie naľavo od tejto kolmice nebude realizované. Tento fakt neskôr podložíme číselnými hodnotami. Konkrétne hodnoty datovej matice sú:

Klient	Doba splatnosti úveru	Výška úveru	Vek
1	18	1239	61
2	12	1655	63
3	24	2032	60
4	24	1940	60
5	15	1520	63
6	12	2012	61
7	10	1364	64
8	9	3832	64
9	24	2384	64
10	10	781	63
11	24	2924	63
12	12	1255	61
13	**60**	**13756**	**63**
14	13	1409	64
15	**20**	**6468**	**60**
16	24	2957	63
17	6	753	64
18	24	3757	62
19	6	1338	62
20	12	3059	61
21	**30**	**7596**	**63**

Tabuľka 5.10: Použitá dátová matica

Aplikácia algoritmu na štandardizované dáta je naznačená v tabuľke 5.11.

Krok	Vzdialenosť	Objekty
1	0,0063	**11,16**
2	0,0176	**3,4**
3	0,1450	**6,12**
4	0,1467	**7,14**
...	...	...
17	0,5174	1,6,12,20,2,5,10,7,14,17,19,8,9,11,16,18, **3,4**
18	0,8620	1,6,12,20,2,5,10,7,14,17,19,8,9,11,16,18,3,4, **15**
19	0,8893	1,6,12,20,2,5,10,7,14,17,19,8,9,11,16,18,3,4,15, **21**
20	1,8810	1,6,12,20,2,5,10,7,14,17,19,8,9,11,16,18,3,4,15, 21, **13**

Tabuľka 5.11: Vzdialenosti zhlukov pre štandardizované dáta

Najbližšie sú k sebe objekty 11 a 16 so vzdialenosťou 0,0063. Je to prvé spojenie dvoch objektov do skupiny. Pri pohľade na konkrétne znaky vidíme, že títo klienti majú rovnakú dobu splatnosti úveru 24 a vek 63. Malý rozdiel pozorujeme vo výške úveru, kde v prvom prípade je to 2 924 nemeckých mariek a v druhom 2 957 nemeckých mariek. Druhá a tretia v poradí najmenšia vzdialenosť sa týka objektov 3 a 4 respektíve 6 a 12. U týchto klientov je zhoda v znaku Vek a Doba splatnosti úveru, no majú vyšší rozdiel vo výške úveru. Zmena nastáva v štvrtej spojenej dvojici v poradí, kde sa u klientov 7 a 14 okrem výšky úveru líši aj doba splatnosti. Celkovo však považujeme objekty týchto spojení za podobné.

Pri náhľade na tabuľku 5.11 vidíme, že posledné spojenie je so vzdialenosťou 1,8810. Došlo k pripojeniu klienta 13. Pozorujeme značné vychýlenie v hodnotách znakov Doba splatnosti úveru 60 mesiacov a Výška úveru 13 756 nemeckých mariek oproti všetkým ostatným klientom v tabuľke v tabuľke 5.10. Predchádzajúce dve spojenia majú podobnú vzdialenosť 0,8893 a 0,8620. Tu boli pripojení klienti 21 a 15. U klienta 21 stále pozorujeme značný rozdiel v hodnotách znakov Doba splatnosti a Výška úveru, no pri klientovi 15 sa rozdiel v znaku Doba splatnosti vytratí. Avšak rozdiel v znaku Výška úveru je stále značný. Následne by sme už videli zmenšujúcu sa vzdialenosť, ktorá najviac zavisí na výške úveru.

Algoritmus ukončíme po sedemnástom kroku. Vzdialenosť spojovaných zhlukov je vtedy 0,5175. Ak by sme pridali ešte osemnásty krok, tak by sa vzdialenosť zmenila na 0,8620 čo predstavuje veľký skok. Prvých sedemnásť krokov teda tvorí zhluk a posledné tri objekty ostanú nezaradené.

V prípade neštandardizovaných dát bude tabuľka vzdialeností vyzerať takto:

Krok	Vzdialenosť	Objekty
1	9,87	**1,12**
2	13,48	**3,6**
3	15,23	**7,19**
4	16,34	**10,17**
...	...	...
17	403,05	1,12,7,19,14,5,2,3,6,4,9,10,17,11,16,20, **8,18**
18	651,28	**15,21**
19	1521,91	1,12,7,19,14,5,2,3,6,4,9,10,17,11,16,20,8,18, **15,21**
20	3556,52	1,12,7,19,14,5,2,3,6,4,9,10,17,11,16,20,8,18,15,21, **13**

Tabuľka 5.12: Vzdialenosti zhlukov pre neštandardizované dáta

Na rozlišenie odľahlých objektov nemala vplyv štandardizácia dát. V oboch prípadoch by sme po sedemnástich krokoch algoritmu nezaradili rovnakých troch klientov a to 13, 21 a 15. Vyznačujú sa netypicky veľkou výškou úveru.

5.2.2 Nehierarchická K-means metóda

Túto metódu budeme znova aplikovať na rovnaké tri spojité znaky. Budú nás zaujímať klienti, ktorí majú viac ako 60 rokov. Ďalej je potrebné predom vybrať počet zhlukov. Hierarchická metóda nám rozdelila dáta na dva zhluky, preto budeme predpokladať dva zhluky. Ukážeme si, čo sa zmení, ak zvýšime počet zhlukov na štyri a päť. Dátová matica je väčšia než sme použili v hierarchickej analýze. Obsahuje 50 objektov a 3 znaky. V programe NCSS bola využitá funkcia K-means Clustering popísaná v kapitole 4.1.2.

Softvér rozdelil 50 objektov na 47 a 3 (obrázok 5.3), ak prepokladáme dva zhluky. Rozptyl v zhluku pozostávajúcich z troch objektov je v znaku Doba splatnosti úveru 0.

Znak	1. zhluk	2. zhluk
Doba splatnosti úveru	8,46	0,00
Výška úveru	2 583,41	4 321,99
Vek	4,36	1,73

Tabuľka 5.13: Smerodajné odchýlky v zhlukoch pre dané znaky

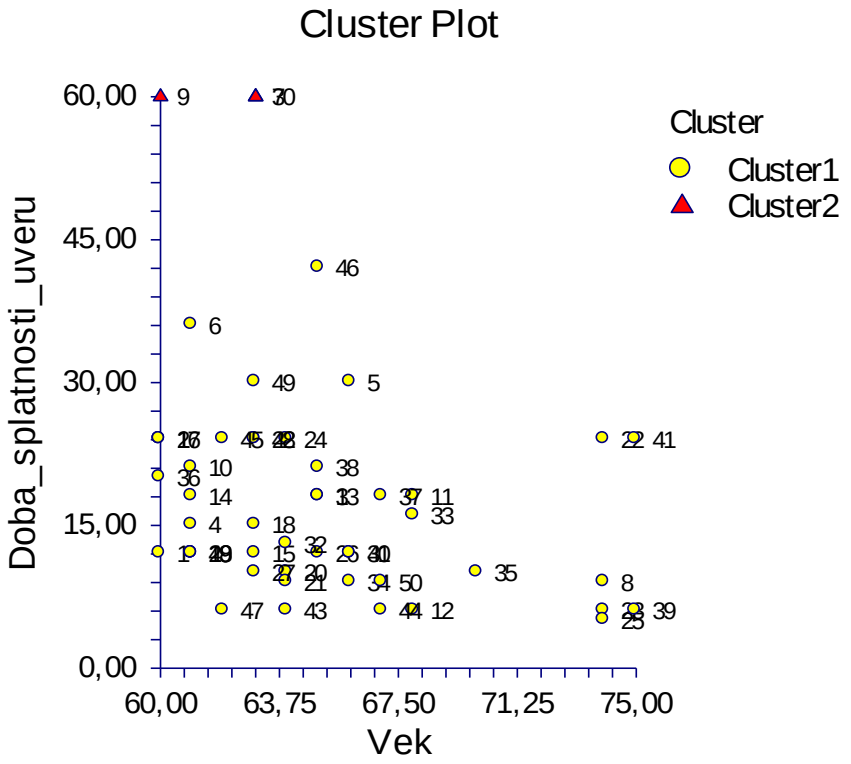
Je zjavné, že tieto tri objekty majú rovnakú Dobu splatnosti úveru. V priemeroch v tabuľke 5.14 je viditeľný veľký rozdiel medzi zhlukmi v Dobe splatnosti

úveru a Výške úveru.

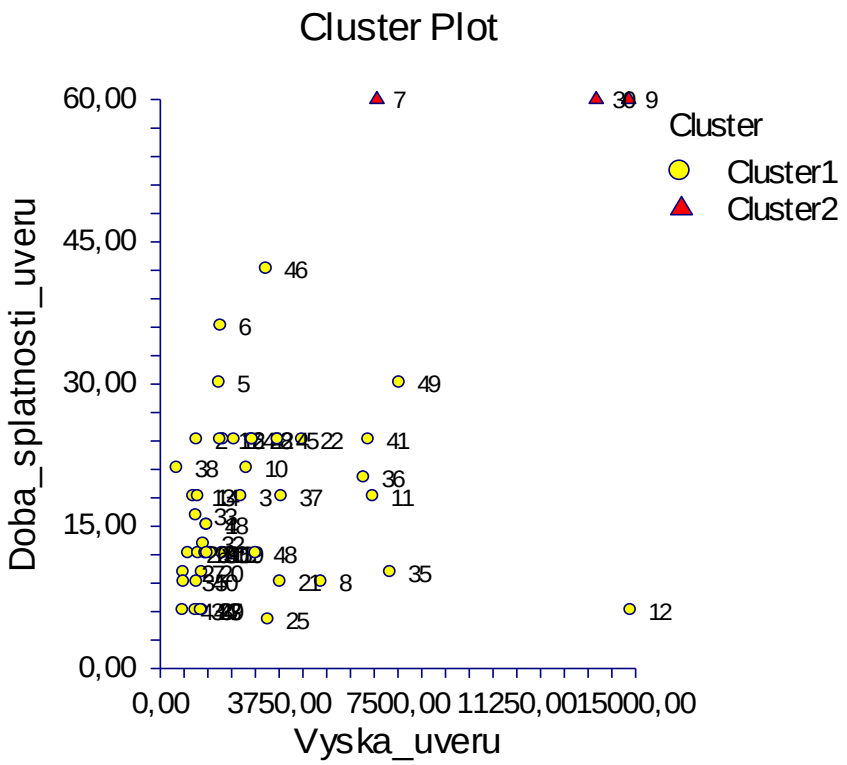
Znak	1. zhluk	2. zhluk
Doba splatnosti úveru	16,34	60,00
Výška úveru	2 830,21	11 791,33
Vek	65,09	62,00

Tabuľka 5.14: Priemery v zhlukoch pre dané znaky

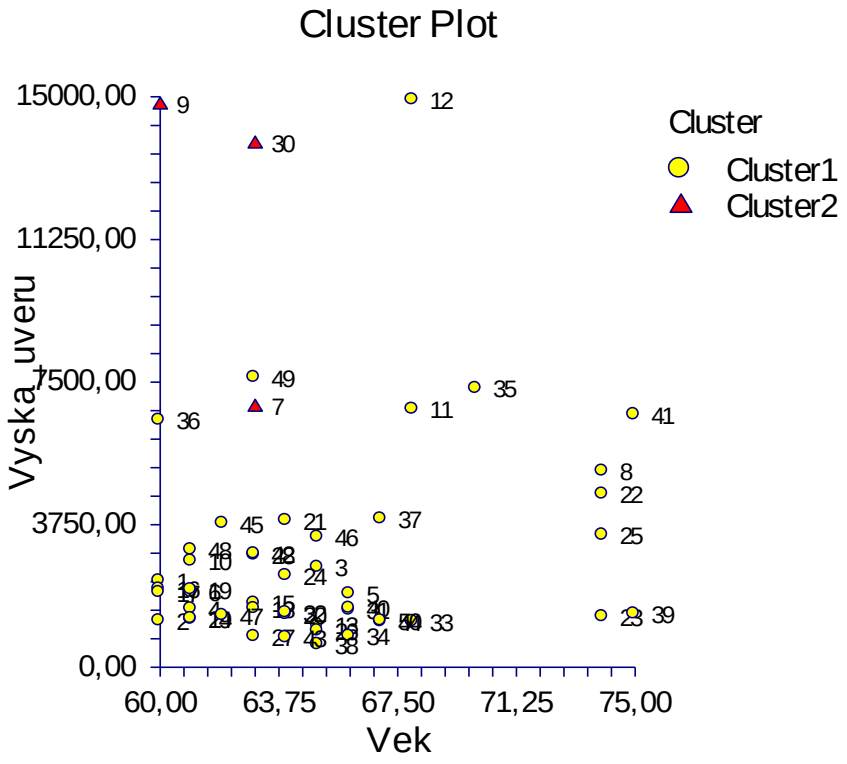
K interpretacii vytvorených zhlukov môžu pomôcť dvojrozmerné projekcie v nasledujúcich obrázkoch:



Obr. 5.3: Zobrazenie zhlukov pre znaky Doba splatnosti úveru a Vek



Obr. 5.4: Zobrazenie zhlukov znaky Doba splatnosti úveru a Výška úveru



Obr. 5.5: Zobrazenie zhlukov pre znaky Výška úveru a Vek

Na predchádzajúcich grafoch sme mohli vidieť jednotlivé zhluky farebne rozlíšené. Aj nehierarchická metóda nám odčlenila tri objekty, ktoré sa značne odlišujú od ostatných. Sú to klienti 7, 9 a 30. Všetci traja majú netypicky dlhú dobu splatnosti úveru 60 mesiacov, nízky vek a klienti 9 a 30 majú vysokú výšku úveru. Jeden z týchto objektov sa nám už objavil v hierarchickej aglomeratívnej metóde, v tabuľke 5.15 má poradie 30 (v tabuľke 5.11 má číslo 13) a ďalšie dva sú novo pridané do dátovej matice pre nehierarchické zhlukovanie. V nižšie uvedenej tabuľke sú zvýraznené. Prehľad vybraných objektov:

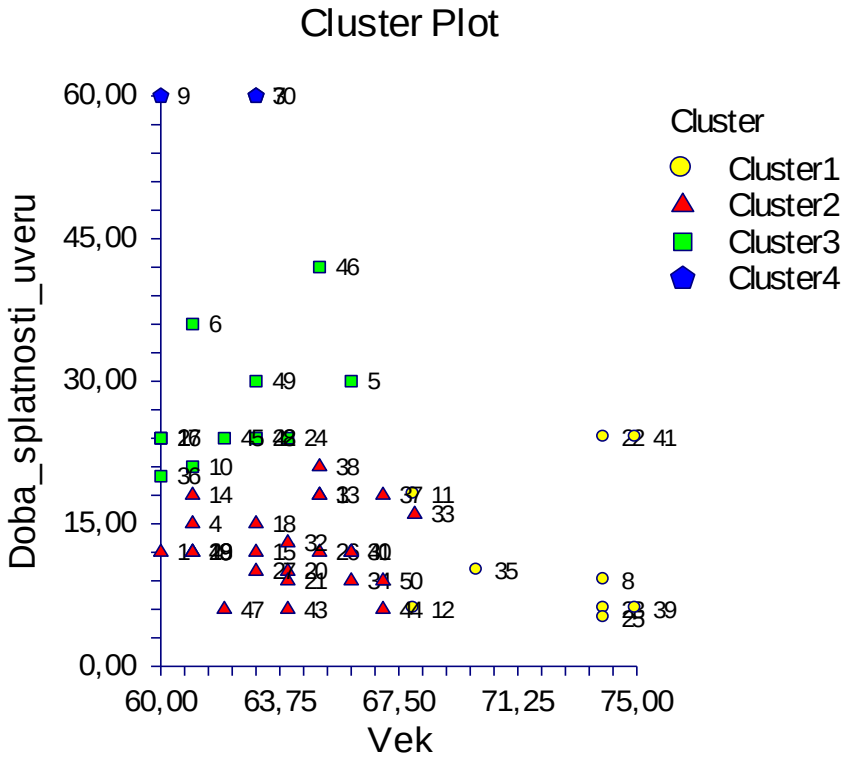
Klient	Doba splatnosti úveru	Výška úveru	Vek
4	15	1512	61
5	30	1908	66
6	36	1953	61
7	**60**	**6836**	**63**
9	**60**	**14782**	**60**
12	6	14896	68
24	24	2384	64
30	**60**	**13756**	**63**
33	16	1175	68
35	10	7308	70
38	21	571	65
39	6	1374	75
41	24	6615	75
43	6	753	64
44	6	1169	67
46	42	3394	65
47	6	1338	62
49	30	7596	63

Tabuľka 5.15: Hodnoty znakov vybraných objektov

Tučne zvýraznené sú objekty, ktoré sa nachádzajú v zhluku 2. Na obrázkoch 5.3, 5.4 a 5.5 sú vyznačené červeným trojuholníkom.

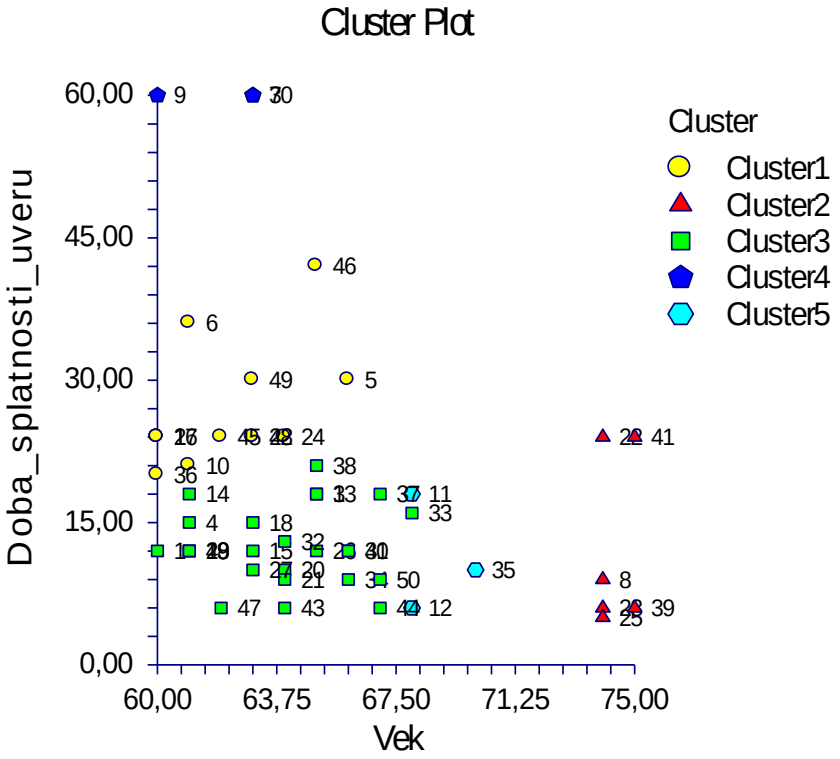
Teraz predpokladajme štyri zhluky. Vyberieme jednu dvojicu znakov, na ktorej demoštrujeme, čo sa stane. Vo zvyšných prípadoch je situácia podobná. Voľbou vyššieho počtu zhlukov docielime väčšiu podobnosť znakov v rámci jedného zhluku. Na obrázku 5.6 môžeme vidieť postupné delenie skupiny žltej farby z obrázka 5.3. Pozorujeme rozdelenie do zhlukov, kde je možné charakterizovať každú skupinu. Modrou farbou je, ako sme už spomínali hore, charakteristická dlhá doba splatnosti a nízky vek v zhluku. Naopak pre žltý zhluk pozorujeme vysoký vek, no

krátku dobu splatnosti. Červenou farbou je zobrazený nízky vek a krátka doba splatnosti úveru a zelenou nízky vek a stredne dlhá doba splatnosti.



Obr. 5.6: Zobrazenie zhlukov pre znaky Doba splatnosti úroku a Vek

Ak zvýšime počet zhlukov na päť, potom pozorujeme na obrázku 5.7 rovnaké rozdelenie objektov do zhlukov až na žltý zhluk na obrázku 5.6. Tento zhluk je rozdelený na dva vzhľadom k veku klientov. Sú označený červenou a svetlo modrou farbou.



Obr. 5.7: Zobrazenie zhlukov pre znaky Doba splatnosti úroku a Vek

5.3 Diskriminačná analýza

V poslednej kapitole pridáme podľa teoretického popisu v kapitole 4.2 ďalšieho klienta. Zaradíme ho do jednej zo skupín, ktorá vznikla použitou zhlukovou nehierarchickou metódou. Uvedieme diskriminačnú analýzu pre dve a štyri počiatočné skupiny.

Najprv predpokladajme dve počiatočné skupiny T_1(47 objektov) a T_2 (3 objekty), ktoré vytvoria testovaciu databázu. Pre správne zaradenie ďalšieho objektu je potrebné napočítať diskriminačné skóry. Keďže predpokladáme dve počiatočné skupiny, budeme počítať lineárne diskriminačné skóry (4.21) :

$$D_s^{(i)} = \overline{\boldsymbol{x}}_s^T \boldsymbol{S}^{-1} \boldsymbol{x}_i - \frac{1}{2} \overline{\boldsymbol{x}}_s^T \boldsymbol{S}^{-1} \overline{\boldsymbol{x}}_s + ln\widehat{\pi}_s \quad s = 1, 2 \quad i = 1, \ldots, 51, \tag{5.2}$$

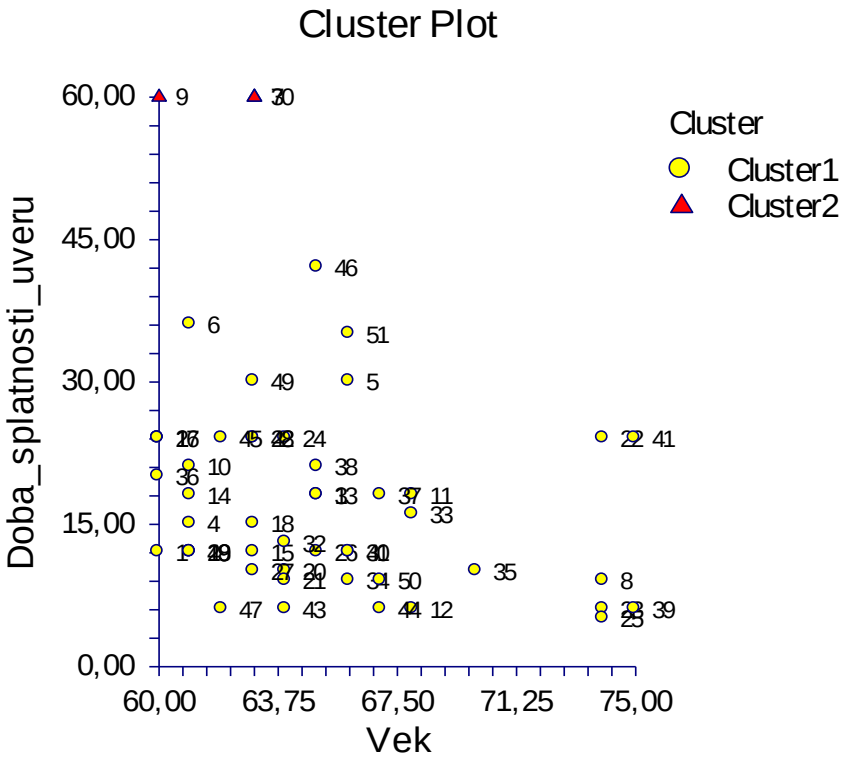
kde priemery $\overline{\boldsymbol{x}}_s$ sú uvedené v tabuľke 5.14 a matica $\boldsymbol{S}$ vyzerá nasledovne:

$$\begin{pmatrix} 68.51 & 1499.03 & -9.92 \\ 1499.03 & 7.17424 \times 10^6 & 2665.19 \\ -9.92 & 2665.19 & 18.37 \end{pmatrix}. \tag{5.3}$$

Znaky sú v poradí Vek, Výška úveru a Doba splatnosti úveru. Za odhad π_s bolo dané $\frac{3}{50}$ v prvej skupine a $\frac{47}{50}$ v druhej. Najprv aplikujeme skóry (5.2) na testovaciu databázu. Napočítaním diskriminačných skórov pomocou softvéru Mathematica a použitím pravidla (4.17) pre 50 objektov bolo zistené, že objekty boli rozdelené do rovnakých skupín ako v nehierarchickej analýze. Grafické zobrazenie skupín je viditeľné na obrázku 5.3. Preukázalo sa správne fungovanie lineárnych diskriminačných skórov.

Pridaným objektom bude klient, ktorý má 66 rokov s výškou úveru 766 mariek na 35 mesiacov. Diskriminačné skóry pre skupiny T1 a T2 sú 165,407 a 153,379. Zaradíme teda nový objekt podľa pravidla (4.17) do prvej skupiny.

Teraz porovnáme zaradenie nového objektu podľa diskriminačnej analýzy s nehierarchickou analýzou. Pre nehierarchickú analýzu budeme predpokladať dáta ako v predchádzajúcej kapitole s vyššie popísaným pridaným objektom 51. Celkovo pracujeme s maticou, ktorá má 51 objektov a 3 znaky. Pri konštrukcii budeme postupovať presne ako v predchádzajúcej kapitole. Pre kontrolu zaradenia už uvediem len graf.



Obr. 5.8: K- means metóda s novým prvkom 51

Predpokladajme štyri skupiny T1,T2,T3 a T4 tak, ako sú vyznačené na obrázku 5.6. Softvér rozdelil 50 objektov na skupiny s počtom 9, 25, 13 a 3 objekty.
Ak máme viac ako dve počiatočné skupiny, budeme počítať kvadratické diskriminačné skóry podľa (4.20):

$$D_s^{(i)} = -\tfrac{1}{2} ln|\boldsymbol{S}_s| - \tfrac{1}{2}(\boldsymbol{x}_i - \overline{\boldsymbol{x}_s})^T \boldsymbol{S}_s^{-1}(\boldsymbol{x}_i - \overline{\boldsymbol{x}_s}) + ln\widehat{\pi}_s, \quad s = 1,2,3 \quad i = 1,\ldots,51.$$

Priemery $\overline{\boldsymbol{x}}_s$ nájdeme v nasledujúcej tabuľke:

Znak	1. zhluk	2. zhluk	3. zhluk	4. zhluk
Doba splatnosti úveru	12,00	12,52	26,69	60,00
Výška úveru	5 706,22	1 615,40	3 175,31	11 791,33
Vek	72,44	63,96	62,15	62,00

Tabuľka 5.16: Priemery štyroch zhlukov

Matice $\boldsymbol{S}_i, i = 1,2,3,4$ vyzerajú nasledovne:

$\boldsymbol{S}_1$:

$$\begin{pmatrix} 61.75 & 1838.13 & 1.75 \\ 1838.13 & 1.67467 \times 10^7 & -8912.24 \\ 1.75 & -8912.24 & 8.52778 \end{pmatrix} \tag{5.4}$$

$\boldsymbol{S}_2$:

$$\begin{pmatrix} 16.51 & 453.533 & 0.271667 \\ 453.533 & 776335. & -209.192 \\ 0.271667 & -209.192 & 5.29 \end{pmatrix} \tag{5.5}$$

$\boldsymbol{S}_3$:

$$\begin{pmatrix} 39.2308 & -313.064 & 6.46795 \\ -313.064 & 3.45944 \times 10^6 & 142.532 \\ 6.46795 & 142.532 & 4.14103 \end{pmatrix} \tag{5.6}$$

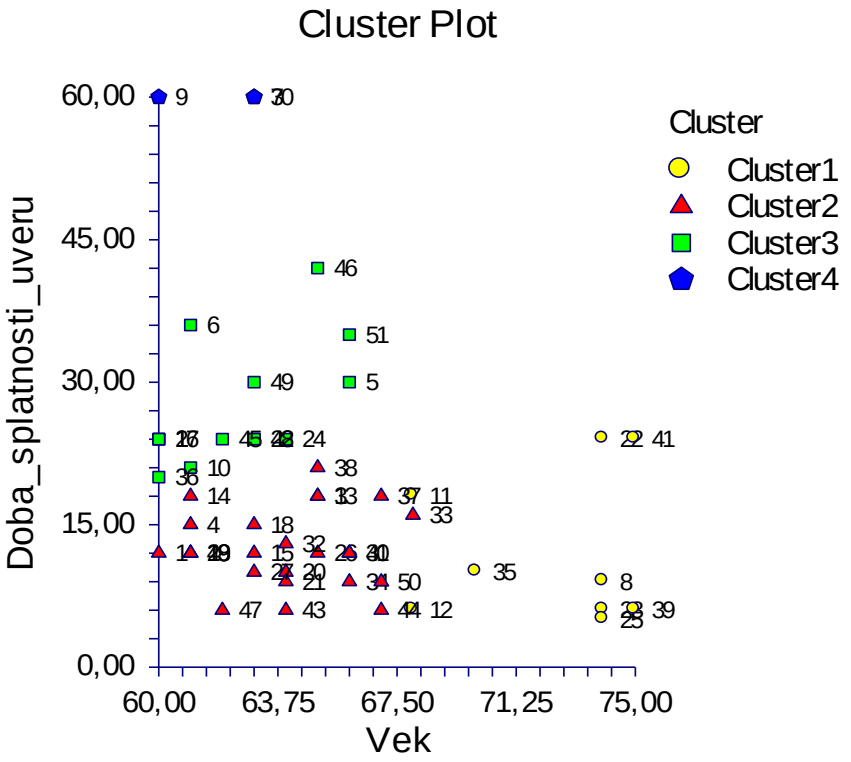
$\boldsymbol{S}_4$:

$$\begin{pmatrix} 0. & 0. & 0. \\ 0. & 1.86797 \times 10^7 & -4486. \\ 0. & -4486. & 3. \end{pmatrix} \tag{5.7}$$

Vo štvrtom zhluku sa nachádzajú len 3 pozorovania, ktoré sú netypické vzhľadom ku dobe splatnosi úveru a výške úveru. Znak Vek má v tomto zhluku nulový rozptyl. Preto zúžime testovaciu databázu na 47 objektov a budeme pracovať so skupinami T1, T2 a T3. Za odhady π_s vezmeme $\frac{p}{47}$, kde p je počet objektov

v skupine. Napočítaním kvadratických diskriminačných skórov pre 47 objektov bolo zistené, že objekty boli rozdelené do rovnakých skupín ako v nehierarchickej analýze.

Pridaným objektom bude opäť klient, ktorý má 66 rokov s výškou úveru 766 mariek na 35 mesiacov. Diskriminačné skóry pre skupiny T1, T2 a T3 sú -31.88, -26.5 a -13.99. Nový objekt zaradíme podľa (4.17) do skupiny T3. Aplikáciou nehierarchickej zhlukovej analýzy na dátovú maticu, ktorá má 51 objektov a 3 znaky sa potvrdilo zaradenie nového klienta do zelenej skupiny T3, ako je vidieť na obrázku 5.9



Obr. 5.9: K- means metóda s novým prvkom 51

Literatúra

[1] Ajvazjan S., Bežajevová Z., Staroverov O.: *Metody vícerozměrné analýzy*. Praha: SNTL, 1981.

[2] Anděl J.: *Základy matematické statistiky*. Praha: Matfyzpress, 2007.

[3] Anděl J.: *Statistické metody*. Praha: Matfyzpress, 2003.

[4] Anděl J.: *Matematická statistika*. Bratislava: SNTL/Alfa, 1978.

[5] Anderberg M. R.: *Cluster Analysis for Applications*. New York: Academic Press, 1978.

[6] Bečvář J.: *Lineární algebra*. Praha: Matfyzpress, 2005.

[7] Bican L.: *Lineární algebra a geometrie*. Praha: Academia, 2000.

[8] Hebák P., Hustopecký J.: *Vícerozměrné statistické metódy s aplikacemi*. Praha: SNTL/Alfa, 1987.

[9] Řehák J., Řeháková B.: *Analýza kategorizovaných dat v sociologii*. Praha: Academia, 1985.

[10] $http : //www.stat.uni - muenchen.de/service/datenarchiv/kredit/kredit_e.html$

[11] $http : //www.karlin.mff.cuni.cz/ kulich/vyuka/statfpm/doc/statfpmvetnik.pdf$

Zoznam obrázkov

Zoznam tabuliek

Zoznam príloh

Príloha č. 1: CD obsahujúce

- DATA.xls - zdrojove dáta
- LenkaGodulova-BP.pdf - elektronická verzia práce
- výstupné protkoly z NCSS
- Súbor Microsoft Office Excel s výpočtami
- Súbor Mathematica s výpočtami