1 Normalizácia dát

Máme vstupné dáta $X = \{x_1, x_2, \ldots, x_N\}$.

1.1 Decimálne škálovanie na interval $< -1, 1 >$

Napíšte funkciu `decscale(x)`, ktorá transformuje vektor `x` vydelením jeho zložiek číslom 10^d, kde d je najmenšie celé číslo také, že platí

$$\forall i \in \{1, \ldots, N\} : -1 \leq x_i \leq 1.$$

Napríklad

```
>> decscale([23.5 -5 0])
ans =

    0.2350   -0.0500         0

>> decscale([0.007 -0.02 0.00082])
ans =

    0.0700   -0.2000    0.0082
```

Pomocou príkazu `plot` zobrazte do jedného grafu `x` a `decscale(x)` – napr. `x` na x-ovú a `decscale(y)` na y-ovú osu.

1.2 Min-max normalizácia na interval $< A, B >$

Napíšte funkciu `mmscale(x)`, ktorá transformuje vektor `x` tak, že jeho zložky lineárne namapuje na interval $< A, B >$. Typicky sa robí min-max normalizácia na interval $< 0, 1 >$ alebo na interval $< -1, 1 >$.

Napríklad

```
>> mmscale([0.12 3 -123],-1,1)
ans =

   0.95429   1.00000  -1.00000

>> mmscale([0.12 3 -123],0,1)
ans =

  0.97714  1.00000  0.00000
```

Pomocou príkazu `plot` zobrazte do jedného graf `x` a `mmscale(x,A,B)` – napr. `x` na x-ovú a `decscale(y)` na y-ovú osu.

1.3 Normalizácia podľa smerodatnej odchýlky

Smerodatná odchýlka (standard deviation) je funkcia

$$sd(X) = \sigma_X = \sqrt{\frac{\sum (X - \bar{X})^2}{N - 1}},$$

kde $\bar{X} = \frac{1}{N} \cdot \sum_{i=1}^{N} x_i$ je stredná hodnota.

Táto normalizácia urobí takú transformáciu, aby dáta mali strednú hodnotu 0 a rozptyl 1.

Napíšte funkciu `sdscale(x)`, ktorá transformuje vektor `x` tak, že jeho zložky budú mať strednú hodnotu 0 a rozptyl 1.

Napríklad

```
>> sdscale([1 2 3])
ans =

        -1         0         1

>> sdscale([-2 2 7])
ans =

  -0.96099  -0.07392   1.03491
```

1.4 Sigmoidálna normalizácia

Sigmoida je funkcia

$$f(x) = \frac{1}{1 + e^{-\lambda x}}, \text{ kde } \lambda > 0 \text{ sa nazýva } \textit{strmosť}.$$

Sigmoida má definičný obor $(-\infty, +\infty)$ a obor hodnôt $(0, 1)$.

Jej graf pre $\lambda = 1$ si môžeme zobrazit napr. príkazmi

```
>> x=-10:0.2:10;
>> plot(x,1./(1+exp(-x)))
```

Napíšte funkciu `sigmscale(x,l)`, ktorá transformuje prvky vektora `x` podľa sigmoidy so strmosťou `l`.

Napríklad

```
>> sigmscale(-3:3,2) ans =

  0.00247  0.01799  0.11920  0.50000  0.88080  0.98201  0.99753
```

Naprogramujte aj transformáciu `sigmscale_inv(x,l)` inverznú k transformácii `sigmscale(x,l)`.

1.5 Normalizácia mnohorozmerných dát

Dáta pre dobývanie znalostí sú obvykle tabuľky s mnohými stĺpcami. Keď sa robí normalizácia takýchto dát, tak sa normalizuje celá tabuľka po stĺpcoch. Tu je treba dať pozor na to, že každý stĺpec má iný rozsah hodnôt a bude obecne transformovaný inou funkciou. Vyššie uvedené transformácie boli definované pre riadkové vektory. Naprogramujte nové verzie dekatického škálovania, min-max škálovania, normalizácie podľa smerodatnej odchýlky a sigmoidálnej transformácie s novými názvami `decscalem(x,c)`, `mmscalem(x,A,B,c)`, `sdscalem(x,c)`,

`sigmscalem(x,l,c)`, ktoré majú ako prvý parameter (dvojrozmernú) maticu `x`. Posledný parameter `c` je riadkový vektor rovnakej dĺžky ako riadky matice `x` a obsahuje iba hodnoty 0 a 1. Príslušná normalizácia sa vykoná na stĺpci i iba vtedy, keď `c[i]` je nenulové.

Napríklad

```
>> yy=5*rand(5,3)
yy =

  0.41501  0.33185  1.89638
  2.05899  0.09640  0.80592
  2.66147  4.35444  3.80238
  0.85257  4.59147  2.21583
  2.31878  4.61248  0.05122

>> decscalem(yy,[1 0 1])
ans =

  0.04150  0.33185  0.18964
  0.20590  0.09640  0.08059
  0.26615  4.35444  0.38024
  0.08526  4.59147  0.22158
  0.23188  4.61248  0.00512

>> mmscalem(yy,-1,1,[1 0 1])
ans =

  -1.00000   0.33185  -0.01622
   0.46362   0.09640  -0.59762
   1.00000   4.35444   1.00000
  -0.61045   4.59147   0.15410
   0.69491   4.61248  -1.00000
```

2 Generovanie testovacích dát

Pre programovanie a testovanie rôznych metód dobývania znalostí je užitočné vediet generovať dáta so zadanými vlastnosťami a tiež vedieť generovať "náhodné" dáta, resp. náhodne vyberať dáta z danej množiny.

Na generovanie náhodných dát je treba poznať rozloženie pravdepodobnosti (probability distribution) pre dáta. Matica veľkosti $m \times n$ dát s rovnomerným rozložením (uniform distribution) z intervalu A, B sa vygeneruje príkazom

```
>> A=3; B=10;
>> m=3; n=4;
>> (B-A)*rand(m,n)+A
ans =

  3.31198  6.28767  7.37211  8.75329
  5.75145  9.82846  8.29843  3.21501
  7.42435  4.56690  5.99315  6.79303
```

Funkcia `randn` generuje náhodné hodnoty s normálnym rozložením so strednou hodnotou 0 a rozptylom 1 (tj. smerodatnou odchýlkou $\sqrt{1} = 1$). Matica veľkosti $m \times n$ dát s normálnym rozložením (normal distribution) so strednou hodnotou S a rozptylom R (tj. smerodatnou odchýlkou $\sqrt{\mathtt{R}}$) sa vygeneruje príkazom

```
>> S=2; R=3;
>> m=3; n=4;
>> sqrt(R)*randn(m,n)+S
ans =

   1.76953   1.02155   2.90496   3.54724
   1.81023   9.41794  -1.39530   3.38295
   4.53440  -2.64642   3.07368   0.76086
```

2.1 Generovanie vektora s dvomi zhlukmi

Napíšte funkciu `randv2n(n1,S1,R1,n2,S2,R2)`, ktorá vygeneruje vektor (riadkový) obsahujúci `n1` hodnôt s normálnym rodelením so strednou hodnotou `S1` a rozptylom `R1` a ďalej `n2` hodnôt s normálnym rodelením so strednou hodnotou `S2` a rozptylom `R2`.

Napríklad

```
>> randv2n(3,-10,1,4,10,1)
ans =

  1.0e+01  *

   0.86684   0.97688  -0.94214  -0.85999   1.09027   1.09225  -1.19626
```

Urobte histogram vygenerovaných dát.

2.2 Generovanie zhlukov v 2D

Naprogramujte funkciu `randn2d(p)`, ktorá má ako parameter maticu `p` tvaru $n \times 5$. Táto funkcia vygeneruje maticu rozmerov $\sum_{i=1}^{n} \mathtt{p}(i) \times 2$ obsahujúcu náhodné body v 2D, pričom `p(`i`,1)` hodnôt je zo zhluku s normálnym rodelením so strednou hodnotou `p(`i`,2)` v prvej súradnici a strednou hodnotou `p(`i`,3)` v druhej súradnici, rozptylom `p(`i`,4)` v prvej súradnici a rozptylom `p(`i`,5)` v druhej súradnici (pre $i = 1, \ldots, n$).

Napríklad:

```
>> rm=randn2d([6,-10,-10,1,1; 8,10,10,2,2])
rm =

  1.0e+01  *

  -0.81834  -1.08722
   1.24685   0.79264
  -0.92659  -0.99046
  -0.96725  -1.02791
```

```
   1.00210   1.14543
   1.05893   0.97251
   0.99114   1.02167
  -1.05398  -0.93580
   0.75075   1.27756
  -1.03174  -1.05069
   0.90385   0.86900
   0.95572   0.85197
   0.99392   0.95680
  -0.94696  -1.15904
```

Zobrazte vygenerované zhluky do 2D grafu.

2.3 Generovanie vzorky z dát

Navrhnite funkciu `selectk(x,k)`, ktorá z matice `x` náhodne vyberie `k` riadkov.

Napríklad:

```
>> rm=rand(8,2)
rm =

  0.02559  0.73449
  0.45501  0.37769
  0.50644  0.09478
  0.31154  0.12610
  0.72651  0.93408
  0.06366  0.83468
  0.91405  0.95055
  0.39877  0.00754

>> srm=selectk(rm,4)
srm =

  0.50644  0.09478
  0.06366  0.83468
  0.45501  0.37769
  0.31154  0.12610
```

Zobazte pôvodné a vybrané dáta do grafu rozdielnymi farbami alebo značkami.

2.4 Náhodné rozdelenie dát do skupín

Naprogramujte funkciu `rearrange(M)`, ktorá náhodne rozdelí riadky matice `M` do skupín veľkosti k (posledná skupina môže mať počet prvkov menší než k). Výstupom je matica s náhodne permutovanými riadkami, kde i-ta skupina sú riadky od $(i-1)k+1$ do $\max(ik, n)$, kde n je počet riadkov matice.

Napríklad:

```
>> a=rand(6,3)
a =
```

```
  0.18244  0.48563  0.15539
  0.55104  0.74847  0.55122
  0.95554  0.64857  0.56828
  0.12789  0.43557  0.76303
  0.80228  0.25556  0.74825
  0.07529  0.16154  0.42594

>> rearrange(a)
ans =

  0.07529  0.16154  0.42594
  0.55104  0.74847  0.55122
  0.95554  0.64857  0.56828
  0.18244  0.48563  0.15539
  0.80228  0.25556  0.74825
  0.12789  0.43557  0.76303
```

3 Analýza atribútu

Využite zobrazenie hodnôt vhodným zobrazením a jednoduché štatistické atribúty ako je priemer a rozptyl na analýzu hodnôt atribútu, ktorý je vo vektore, ktorý získate volaním funkcie `dataa` zo súboru http://ksvi.mff.cuni.cz/~mraz/datamining/dataa.m. Pomohla by zhluková analýza?

4 Kontingenčné tabuľky, χ^2 a Fisherov test

Pri cvičných zoskokoch boli parašutisti rozdelení do dvoch skupín A a B, ktoré skákali na dvoch rôznych miestach. V jednej skupine bolo mnoho zranených. Celkové počty sú uvedené v nasledujúcej tabuľke:

	zranených	zdravých	Celkom
Skupina A	2	38	40
Skupina B	5	10	15
Celkom	7	48	55

Table 1: Kontingenčná tabuľka

Spočítajte χ^2-test a Fisherov test na hladine významnosti $\alpha = 0.05$. Odpovedzte na otázku, či rozdiely medzi skupinami sú náhodné, alebo nie.