Podstawy `R`: typy złożobe m.in. macierze.

Wszystkie zadania rozwiązujemy bez używania pętli.

Zadanie 3.1 [MG] Niech `t` będzie wektorem o n elementach będących liczbami całkowitymi ze zbioru $\{1, \ldots, k\}$. Napisz funkcję, która dokona kodowania elementów t_i (*one-hot-encode*). Funkcja powinna zwracać macierz zero-jedynkową R wymiaru $n \times k$ taką, że $r_{i,j} = 1$ wtedy i tylko wtedy gdy $t_i = j$. Taka reprezentacja jest przydatna np. w problemie klasyfikacji wieloetykietowej przy użyciu k klasyfikatorów.

Zadanie 3.2 [MG] Dokonaj przekształcenia *softmax* każdego wiersza macierzy $\mathbf{X} \in \mathbb{R}^{n \times k}$, tzn. przekształcenia postaci:

$$x_{i,j} \mapsto \frac{\exp(x_{i,j})}{\sum_{l=1}^{k} \exp(x_{i,l})}.$$

Następnie dokonaj odkodowania każdego wiersza (*one-hot decode*), tj. dla każdego wiersza należy znaleźć numer kolumny o wartości najbardziej zbliżonej do 1. Zwróć wektor n-elementowy.

Zadanie 3.3 [MG] Niech dana będzie macierz $\mathbf{X} \in \mathbb{R}^{n \times d}$. Wyznacz przedział wielowymiarowy ograniczający wartości n punktów reprezentowanych jako $\mathbf{X}$ (*bounding hyperrectangle*). Dokładniej, wyznacz i zwróć macierz $\mathbf{B} \in \mathbb{R}^{2 \times d}$ taką, że $b_{1,j} = \min_i x_{i,j}$ oraz $b_{2,j} = \max_i x_{i,j}$.

Zadanie 3.4 [MG] Niech macierz $\mathbf{X}$ wymiaru $n \times d$ reprezentuje n punktów w $\mathbb{R}^d$. Napisz funkcję, która wyznaczy i zwróci odległości między punktami z $\mathbf{X}$ oraz (danym jako drugi argument funkcji) punktem $\mathbf{y} \in \mathbb{R}^d$. Funkcja powinna zwrócić wektor $\mathbf{d} \in \mathbb{R}^n$ taki, że $d_i = \|\mathbf{x}_{i,\cdot} - \mathbf{y}\|_2$.

Zadanie 3.5 [MG] Dana jest macierz $P \geq 0$ rozmiaru $n \times m$ taka, że $\sum_{i=1}^{n} \sum_{j=1}^{m} p_{i,j} = 1$ oraz posortowane rosnąco wektory liczbowe `x` (n-elementowy) i `y` (m-elementowy). Trójka $(\mathtt{x}, \mathtt{y}, P)$ opisuje rozkład prawdopodobieństwa pewnej dwuwymiarowej zmiennej losowej dyskretnej (X, Y), tak jak w poniższym podzadaniu.

W pewnej szkole rozkład prawdopodobieństwa uzyskania ocen z Filozofii Bytu i Wychowania Fizycznego przez tego samego studenta przedstawia się następująco.

		WF			
		2	*3*	*4*	*5*
FB	*2*	*0*	*0,01*	*0,1*	*0,2*
	3	*0,01*	*0,05*	*0,03*	*0,1*
	4	*0,1*	*0,03*	*0,05*	*0,01*
	5	*0,2*	*0,1*	*0,01*	*0*

Rysunek 1: Tabela prawdopodobieństw

a) Zmienne losowe X i Y są niezależne wtedy i tylko wtedy, gdy dla każdego i, j zachodzi $p_{i,j} = (\sum_{k=1}^{n} p_{k,j})(\sum_{l=1}^{m} p_{i,l})$. Napisz funkcję `niezaleznosc()`, która sprawdza, czy zachodzi ta własność dla danych $(\mathtt{x}, \mathtt{y}, P)$ (zwracamy wartość logiczną).

b) Ponadto napisz funkcję `podststat()`, która dla $(\mathtt{x}, \mathtt{y}, P)$ zwróci wektor liczbowy (z ustawionym atrybutem `names` – dowolne, lecz czytelne dla użytkownika etykiety) zawierający wartości podstawowych charakterystyk (X, Y):

- Wartości oczekiwane: $\mathbb{E}\,X = \sum_{i=1}^{n} x_i \sum_{j=1}^{m} p_{i,j}$, $\mathbb{E}\,Y = \sum_{j=1}^{m} y_j \sum_{i=1}^{n} p_{i,j}$,
- Wariancje: $\operatorname{Var} X = \mathbb{E}\,X^2 - (\mathbb{E}\,X)^2$, gdzie $\mathbb{E}\,X^2 = \sum_{i=1}^{n} x_i^2 \sum_{j=1}^{m} p_{i,j}$ oraz $\operatorname{Var} Y = \mathbb{E}\,Y^2 - (\mathbb{E}\,Y)^2$, gdzie $\mathbb{E}\,X^2 = \sum_{j=1}^{m} y_j^2 \sum_{i=1}^{n} p_{i,j}$,
- Kowariancję: $\operatorname{Cov}(X,Y) = \mathbb{E}\,(XY) - \mathbb{E}\,X\,\mathbb{E}\,Y$ dla $\mathbb{E}\,(XY) = \sum_{i=1}^{n}\sum_{j=1}^{m} x_i y_j p_{i,j}$,
- Współczynnik korelacji: $\varrho(X,Y) = \operatorname{Cov}(X,Y)/\sqrt{\operatorname{Var} X \operatorname{Var} Y}$.

Zadanie 3.6 [MG] Napisz funkcję `movingavg()`, która dla wektora liczbowego `x` o n elementach (reprezentującego pewien szereg czasowy) oraz nieparzystej liczby naturalnej k wyznaczy k-średnią ruchomą, $k < n$, tj. ciąg wartości $(w_1, \ldots, w_{n-k+1})$ takich, że $w_i = \sum_{j=1}^{k} x_{i+j-1}/k$.

Zadanie 3.7 [MG] Wykonaj następujące polecenia.

1. Wybierz 100 losowych wierszy z ramki danych `iris` (ramka danych jest dostępna bez konieczności wczytywania).
2. Wybierz 5% wierszy w sposób losowy.
3. Wybierz 100 pierwszy wierszy.
4. Wybierz 100 ostatnich wierszy.

Zadanie 3.8 [MG] Do ramki danych `vehicles` z pakietu `fueleconomy` (`fueleconomy::vehicles`) dodaj nową kolumny `z_cty` oraz `z_hwy`, takie, że są to ustandaryzowane (z-scores) zmienne `cty` (city-) and `hwy` (highway-fuel economy, in mpg) w grupach wyznaczonych przez kolumne `class`. Innymi słowy, standaryzujemy (średnia = 0, odchylenie standardowe = 1) zmienne w każdej podgrupie. Na poczatku jednak zamień jednostki z `mile-per-galon` na `1l/100km`.

Zadanie 3.9 [AO] Napisz funkcję `rozwin()`, która przekształca daną macierz rozmiaru $n \times m$ (niekoniecznie liczbową) z ustawionym atrybutem `dimnames` na ramkę danych zawierającą nm obserwacji i trzy kolumny o nazwach zadanych przy użyciu odpowiedniego argumentu funkcji. Wartości z macierzy mają znajdować się w pierwszej kolumnie, a w kolejnych dwóch – kombinacje nazw wierszy i kolumn odpowiadające podanym poziomom czynnika.

Na przykład obiekt `WorldPhones` (wbudowany) zawiera dane o liczbie telefonów (w tysiącach) w różnych regionach świata w wybranych latach. Wynikiem wywołania `rozwin(WorldPhones, c("ile", "gdzie", "kiedy"))` może być:

```
     ile    gdzie kiedy
...
2  60423   N.Amer  1956
3  64721   N.Amer  1957
...
9  29990   Europe  1956
10 32510   Europe  1957
...
```

Zadanie 3.10 [MG] Napisz funkcję odwrotną do funkcji z zad. 3.9. Dana jest ramka danych zawierająca nm wierszy oraz 3 kolumny (pierwsza – dowolnego typu, druga i trzecia – typu czynnikowego, odpowiednio o n i m poziomach). Obserwacje zawierają wszystkie możliwe kombinacje poziomów dwóch czynników, ale nie możemy założyć, że są one koniecznie ułożone w jakimś określonym porządku (funkcja ma działać dla dowolnej permutacji obserwacji). Wynikiem ma być macierz rozmiaru $n \times m$ o elementach pochodzących z pierwszej kolumny ramki danych. Atrybut `dimnames` ustawiamy na podstawie wartości poziomów pierwszego i drugiego czynnika.