



Math4You

2023–2025

TODO: až nebude na sandboxu, upravit velikosti obrázku v PDF

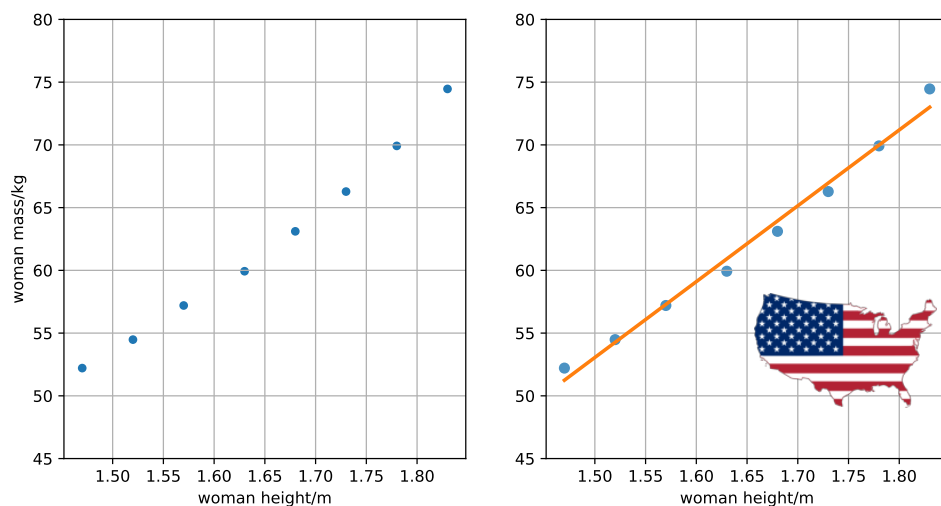
Lineární regrese

V praxi se často setkáváme s tím, že hodnotami jedné veličiny jsou dány hodnoty druhé veličiny a ze sady naměřených nebo statisticky získaných dat určíme matematický model udávající funkční závislost mezi oběma veličinami. Jako příklad můžeme uvažovat data udávající výšku amerických žen ve věku mezi 30 a 39 lety (zdroj https://en.wikipedia.org/wiki/Simple_linear_regression, 12.4.2024, pro stručnost je použita jenom polovina dat).

výška/m	1,47	1,52	1,57	1,63	1,68	1,73	1,78	1,83
hmotnost/kg	52,21	54,48	57,20	59,93	63,11	66,28	69,92	74,46

Data jsou vykreslena na obrázku vlevo. Z obrázku je patrné, že s rostoucí výškou roste i hmotnost. V takovém případě je možné najít matematický model, který udává hmotnost jako funkci výšky. Takový matematický model je znázorněn barevně na obrázku vpravo. Jedná se o model, který umožní pro zadanou výšku ženy předpovědět její hmotnost.

Co-funded by the Erasmus+ Programme of the European Union.



Obrázek 1: Vlevo jsou vykreslena data jak závisí hmotnost amerických žen (woman mass) na jejich výšce (woman height). Vpravo jsou data doplněna regresní přímkou reprezentující matematický model funkční závislosti mezi výškou a hmotností.

Popsaná úloha se nazývá *lineární regrese*.

Lineární regrese jedna ze základních metod *strojového učení*, kdy v datech odhalíme jistou funkční závislost. Tuto poté můžeme použít k tomu, abychom dávali predikce funkčních hodnot pro data, která se v zadaném souboru nevyskytují.

V následujícím textu si ukážeme, jak lineární regrese souvisí s lineární kombinací vektorů a jak je možné regresní přímkou najít pomocí operací s vektory.

- Nejprve si ukážeme, jak se řeší úlohy na zápis vektoru jako lineární kombinace zadaných vektorů.
- Poté si ukážeme, jak je možné si předchozí úlohu usnadnit, pokud je některý z vektorů kolmý k ostatním.
- Poté ukážeme, jak je možné najít přibližné řešení úlohy v případě, že přesné řešení neexistuje.
- Na závěr si ukážeme, jak se tato technika dá využít ke strojovému učení. Ze zadaných dat sestavíme matematický model odhalující trend v těchto datech a umožňující predikovat funkční hodnoty i pro hodnoty, které se v datovém souboru nevyskytují.

Lineární kombinace vektorů

Nejprve jedna snadná úloha. *Klasickou úlohou související s vektory je zapsat zadaný vektor jako lineární kombinaci zadaných vektorů.*

Úloha zapsat vektor $\vec{c} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$ jako lineární kombinaci vektorů $\vec{a} = \begin{pmatrix} 2 \\ 2 \end{pmatrix}$ a $\vec{b} = \begin{pmatrix} 3 \\ 1 \end{pmatrix}$ znamená najít čísla t_1 a t_2 s vlastností

$$t_1 \vec{a} + t_2 \vec{b} = \vec{c}.$$

Po rozepsání do souřadnic vidíme, že tato úloha vede na soustavu rovnic

$$\begin{aligned}2t_1 + 3t_2 &= 1, \\2t_1 + t_2 &= 2.\end{aligned}$$

Tato soustava má jediné řešení $t_1 = \frac{5}{4}$ a $t_2 = -\frac{1}{2}$.

Lineární kombinace s kolmými vektory

Nyní něco komplikovanějšího. Úlohu na sestavení lineární kombinace si ztížíme tím, že vektorů bude více. Vektory však budou mít speciální vlastnost, která řešení znatelně usnadní.

Úloha najít vektor $\vec{w} = \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}$ jako lineární kombinaci vektorů $\vec{u}_1 = \begin{pmatrix} 2 \\ 2 \\ 1 \end{pmatrix}$, $\vec{u}_2 = \begin{pmatrix} 3 \\ 1 \\ 2 \end{pmatrix}$, $\vec{u}_3 = \begin{pmatrix} 3 \\ -1 \\ -4 \end{pmatrix}$ vede na úlohu najít čísla t_1 , t_2 a t_3 s vlastostí

$$t_1\vec{u}_1 + t_2\vec{u}_2 + t_3\vec{u}_3 = \vec{w}. \quad (1)$$

Po dosazení dostáváme soustavu tří rovnic

$$\begin{aligned}2t_1 + 3t_2 + 3t_3 &= 1, \\2t_1 + t_2 - t_3 &= 2, \\t_1 + 2t_2 - 4t_3 &= 1.\end{aligned} \quad (2)$$

Řešení takové soustavy je poměrně nepříjemné, ale v tomto případě můžeme využít speciální vlastnosti zadaných vektorů. Vektor $\vec{u}_3$ je totiž kolmý k vektorům $\vec{u}_1$ a $\vec{u}_2$ a tedy i k rovině definované těmito vektory. Tuto skutečnost snadno ukážeme výpočtem skalárních součinů

$$\vec{u}_1 \cdot \vec{u}_3 = 2 \cdot 3 + 2 \cdot (-1) + 1 \cdot (-4) = 0$$

a

$$\vec{u}_2 \cdot \vec{u}_3 = 3 \cdot 3 + 1 \cdot (-1) + 2 \cdot (-4) = 0.$$

Díky této vlastnosti se vyplatí rovnici (1) vynásobit skalárně po řadě vektory $\vec{u}_1$ až $\vec{u}_3$. Tím dostáváme následující tři rovnice.

$$\begin{aligned}t_1(\vec{u}_1 \cdot \vec{u}_1) + t_2(\vec{u}_2 \cdot \vec{u}_1) + t_3(\vec{u}_3 \cdot \vec{u}_1) &= \vec{w} \cdot \vec{u}_1 \\t_1(\vec{u}_1 \cdot \vec{u}_2) + t_2(\vec{u}_2 \cdot \vec{u}_2) + t_3(\vec{u}_3 \cdot \vec{u}_2) &= \vec{w} \cdot \vec{u}_2 \\t_1(\vec{u}_1 \cdot \vec{u}_3) + t_2(\vec{u}_2 \cdot \vec{u}_3) + t_3(\vec{u}_3 \cdot \vec{u}_3) &= \vec{w} \cdot \vec{u}_3\end{aligned}$$

Výpočtem skalárních součinů dostáváme soustavu, která je nepoměrně jednodušší než soustava (2).

$$\begin{aligned}9t_1 + 10t_2 &= 7 \\10t_1 + 14t_2 &= 7 \\26t_3 &= -3\end{aligned}$$

Z poslední rovnice vidíme přímo jednu z neznámých a dvě první rovnice tvoří soustavu dvou rovnic o dvou neznámých. (Konkrétní řešení není v tuto chvíli podstatné. Je však zřejmé, že soustava má právě jedno řešení.)

Lineární kombinace s nekonzistentní soustavou rovnic

Nyní naši úlohu modifikujeme do třetice. Jeden z vektorů se kterými pracujeme vynecháme, čímž se stane úloha v klasickém smyslu neřešitelnou. Zobecníme však rozumným způsobem pojem řešení. Nebudeme hledat hodnoty neznámých, pro které jsou si levé a pravé strany rovny. Namísto toho budeme hledat alespoň takové hodnoty neznámých, pro které se levé a pravé strany co nejméně liší.

Úloha vyjádřit vektor $\vec{w}$ z minulé podkapitoly jako lineární kombinaci vektorů $\vec{u}_1$ a $\vec{u}_2$ je ukázkou takového problému. Musíme najít čísla t_1, t_2 taková, že platí

$$t_1\vec{u}_1 + t_2\vec{u}_2 = \vec{w}$$

a v souřadnicích

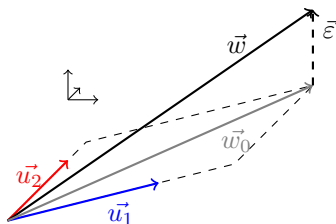
$$2t_1 + 3t_2 = 1,$$

$$2t_1 + t_2 = 2,$$

$$t_1 + 2t_2 = 1.$$

Je snadné se přesvědčit, že tato soustava je nekonzistentní a nemá řešení. Opravdu, soustavu složenou z prvních dvou rovnic jsme vyřešili v úvodu ($t_1 = \frac{5}{4}$ a $t_2 = -\frac{1}{2}$) a poslední rovnice je s touto volbou ve sporu ($\frac{5}{4} + 2 \cdot (-\frac{1}{2}) \neq 1$).

Řešení nekonzistentní soustavy rovnic. V tuto chvíli oslabíme požadavky na řešení soustavy rovnic. Řešením nekonzistentní soustavy rovnic budeme rozumět takovou volbu hodnot neznámých, pro kterou je délka vektoru vyjadřujícího rozdíl levých a pravých stran soustavy minimální.



Obrázek 2: Vektory $\vec{u}_1$ a $\vec{u}_2$ definují rovinu, v níž neleží vektor $\vec{w}$. Proto vektor $\vec{w}$ není možné zapsat jako lineární kombinaci vektorů $\vec{u}_1$ a $\vec{u}_2$. Je však možné jako lineární kombinaci zadaných vektorů zapsat kolmý průmět $\vec{w}_0$ vektoru $\vec{w}$ do uvažované roviny. Vektor $\vec{w}_0$ je ze všech vektorů, které je možno zapsat jako lineární kombinaci vektorů $\vec{u}_1$ a $\vec{u}_2$, nejbližší vektoru $\vec{w}$. Kvantitativní kritérium pro tuto vlastnost je délka vektoru $\vec{\epsilon}$.

Na obrázku si vysvětlíme, co daná soustava vyjadřuje a jak je možné si představit její řešení ve výše uvedeném oslabeném smyslu. Snažíme se pomocí lineární kombinace vektorů $\vec{u}_1$ a $\vec{u}_2$ vyjádřit vektor $\vec{w}$, který v rovině definované těmito vektory neleží. Taková úloha nemá řešení. Budeme se proto snažit najít alespoň takovou lineární kombinaci, která se od vektoru $\vec{w}$ liší co nejméně. To bude vektor $\vec{w}_0$, přičemž rozdíl mezi $\vec{w}$ a $\vec{w}_0$ je znázorněn vektorem $\vec{\epsilon}$. Snažíme se o to, aby délka vektoru $\vec{\epsilon}$ byla co nejmenší. Z geometrie je snadné nahlédnout, že toto nastane v případě, kdy je vektor $\vec{\epsilon}$ kolmý k rovině definované vektory $\vec{u}_1$ a $\vec{u}_2$. Tím se dostáváme do stejné situace, jako byla v předchozí části textu. Tam jsme také viděli trik, jak najít koeficienty u vektorů $\vec{u}_1$ a $\vec{u}_2$ bez řešení plné soustavy rovnic: vynásobili jsme soustavu skalárně vektory $\vec{u}_1$ a $\vec{u}_2$. Dokonce k tomuto výpočtu ani nemusíme znát vektor $\vec{\epsilon}$.

Protože délka vektoru $\vec{\epsilon}$ se vyjadřuje pomocí druhých mocnin souřadnic tohoto vektoru, má metoda název *metoda nejmenších čtverců*.

Celý postup si ukážeme na příkladě v následující části.

Lineární regrese

Uvažujme data z následující tabulky.

x	2	3	4
y	1	5	7

Hledáme přímku $y = ax + b$, která co nejlépe vystihuje trend v tomto souboru a byla by vhodným matematickým modelem pro tato data. Dosazením každého ze tří bodů do modelu dostáváme tři rovnice.

$$2a + b = 1$$

$$3a + b = 5$$

$$4a + b = 7$$

To je úloha s nekonzistentní soustavou rovnic, která nemá řešení v klasickém slova smyslu. Vektorový tvar této soustavy je následující.

$$a \begin{pmatrix} 2 \\ 3 \\ 4 \end{pmatrix} + b \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 5 \\ 7 \end{pmatrix}$$

Po vynásobení po řadě vektory $\begin{pmatrix} 2 \\ 3 \\ 4 \end{pmatrix}$ a $\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$ dostáváme soustavu dvou rovnic.

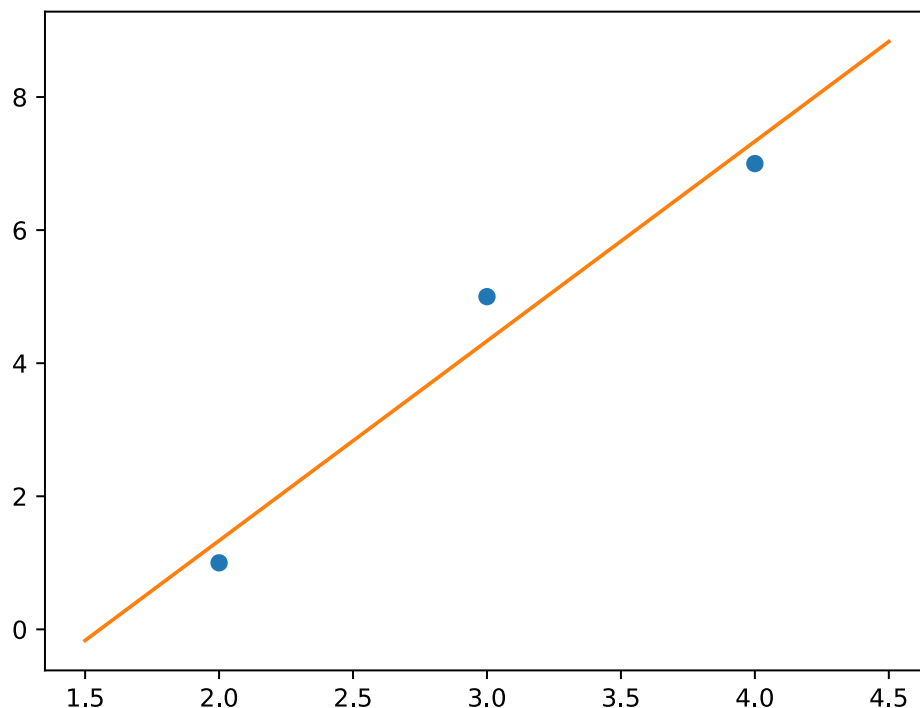
$$29a + 9b = 45$$

$$9a + 3b = 13$$

Řešení této soustavy je $a = 3$ a $b = -\frac{14}{3}$. Regresním modelem pro zadaná data je tedy přímka

$$y = 3x - \frac{14}{3}.$$

Graf obsahující uvedená data a regresní přímku je na obrázku.



Obrázek 3: Tři body neležící je jedné přímce a přímka, která je regresním modelem pro zadanou trojici bodů.

Regrese pro větší datové soubory

Postup uvedený v předchozím textu pro tři body je možné zobecnit na libovolný počet bodů. Není neobvyklé pracovat s datovým souborem obsahujícím stovky bodů.

Je-li vektor $\vec{X}$ vektorem obsahujícím hodnoty nezávislé proměnné¹ a $\vec{Y}$ vektor obsahující hodnoty závislé proměnné, budeme uvažovat model²

$$\vec{Y} = a\vec{X} + b.$$

Koeficienty a a b určíme tak, že tuto rovnici přepíšeme na vektorovou rovnici

$$\vec{Y} = a\vec{X} + b\vec{1},$$

kde $\vec{1}$ je vektor sestavený z jedniček. Tuto rovnici vynásobíme skalárně vektorem $\vec{X}$ a vektorem $\vec{1}$.

¹Použijeme zápis běžně užívaný při zpracování dat, kdy datové soubory (vektory) označujeme velkými písmeny a vektor, který má všechny komponenty rovny stejnému číslu zapisujeme jako dané číslo se šipkou pro označení vektoru.

²Přísně vzato tato operace nemá matematicky smysl, protože sčítáme vektor s reálným číslem. Tuto operaci je nutno interpretovat po složkách, kdy toto sčítání znamená, že se reálné číslo změní na vektor příslušné dimenze tak, aby operace byla definována. toto přizpůsobení se nazývá *broadcasting*. Ve výsledku ke každé komponentě vektoru $a\vec{X}$ přičítáme hodnotu b .

Dostáváme tak soustavu

$$\begin{aligned}a(\vec{X} \cdot \vec{X}) + b(\vec{X} \cdot \vec{1}) &= \vec{X} \cdot \vec{Y} \\ a(\vec{1} \cdot \vec{X}) + b(\vec{1} \cdot \vec{1}) &= \vec{1} \cdot \vec{Y}\end{aligned}\tag{3}$$

Pro více než tři body takto pracujeme s vektory větší dimenze než tři. V důsledku toho přicházíme o názornou geometrickou představu. Až na tuto skutečnost se však na práci nic nemění. Skalární součin dvou se stále počítá tak, že se vynásobí odpovídající komponenty a tyto součiny se poté sečtou.

Po dosazení dat o amerických ženách z úvodu tohoto textu dostáváme následující vztahy.

$$\begin{aligned}\vec{X} \cdot \vec{X} &= 1,47^2 + 1,52^2 + 1,57^2 + \dots + 1,83^2 = 21,9257 \\ \vec{X} \cdot \vec{Y} &= 1,47 \cdot 52,21 + 1,52 \cdot 54,48 + 1,57 \cdot 57,20 + \dots + 1,83 \cdot 74,46 = 828,4568 \\ \vec{1} \cdot \vec{X} &= 1,47 + 1,52 + 1,57 + \dots + 1,83 = 13,21 \\ \vec{1} \cdot \vec{Y} &= 52,21 + 54,48 + 57,20 + \dots + 74,46 = 497,59 \\ \vec{1} \cdot \vec{1} &= 1 + 1 + 1 + \dots + 1 = 8\end{aligned}$$

Po dosazení hodnot skalárních součinů do (3) dostáváme soustavu dvou rovnic o dvou neznámých

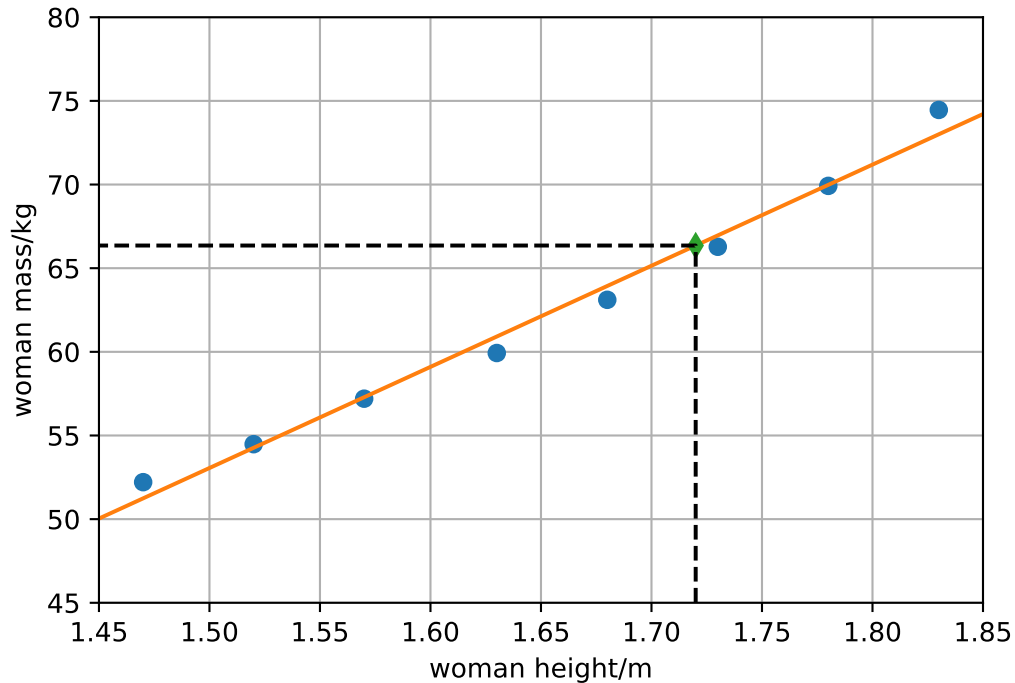
$$\begin{aligned}21,9257a + 13,21b &= 828,4568, \\ 13,21a + 8b &= 497,59,\end{aligned}$$

která má jediné řešení. Tímto řešením³ je $a = 60,44$ a $b = -37,61$. Modelem, který udává závislost hmotnosti žen y na jejich výšce x , je vztah

$$y = 60,44x - 37,61.$$

Obrázek ukazuje použitá data, regresní závislost a predikci pro hmotnost ženy o výšce 1,72 metru.

³Pozor, úloha je poměrně citlivá na zaokrouhlování.



Obrázek 4: Data použitá v úloze, lineární regrese a predikce pro výšku 1,72 metru.

Úlohy pro samostatnou práci

Úloha 1. Najděte regresní přímku pro zadaná data.

x	1,0	2,0	3,0	4,0
y	2,3	2,5	3,1	3,3

Úloha 2. Lineární regrese je jednou ze základních inženýrských metod. Jako taková je prakticky v každém nástroji na zpracování dat. Lineární regresi obsahuje většina vyspělejších kalkulátor, lineární regresi najdete v programech MS Excel a OpenOffice Calc, najdete ji i v programovacích jazycích pro práci s daty, jako je Python nebo Octave. Zjistěte, jestli tato funkce je ve výpočetním prostředí, který používáte a zopakujte výpočet se závislostí hmotnosti žen na výšce. Hledejte pojmy jako lineární regrese nebo metoda nejmenších čtverců.

Závěrečné poznámky

- Ve statistice je uvedená metoda jednou ze základních metod při predikci, zda jedna veličina má vliv na hodnoty druhé veličiny. Proto jsou k dispozici metody, které hodnotí kvalitu aproximace a

také to, zda je pro daný soubor bodů uvažovaná aproximace vhodná či nikoliv.

- Existují také vícerozměrné verze metody nejmenších čtverců, kdy predikovanou hodnotu určujeme ne z jedné ale z několika nezávislých veličin.
- Úloha najít řešení nekonzistentní soustavy lineárních rovnic se vyskytuje i při rekonstrukci obrazu v *akustické tomografii*. To umožňuje studovat složení geologických vrstev nebo zdravotní stav dřeva či stromu na základě informace o rychlosti, s jakou materiálem prochází vlny elastické deformace. Jako úvod do problematiky může posloužit série článků na blogu <https://tomroelandts.com/>.
- Je možné sestavit přímo vztahy pro výpočet koeficientů lineární regrese ze zadaných dat a vynechat tak výpočet skalárních součinů a řešení soustavy rovnic. Viz například https://en.wikipedia.org/wiki/Simple_linear_regression#Expanded_formulas.

Literatura a odkazy

Literatura

- Wikipedie, Simple linear regression, https://en.wikipedia.org/wiki/Simple_linear_regression, 12.4.2024
- Tom Roelandts, <https://tomroelandts.com/articles/tomography-part-1-projections>, <https://tomroelandts.com/articles/sirt-algorithm>, 13.4.2024

Zdroje obrázků

- https://commons.wikimedia.org/wiki/File:Flag-map_of_the_United_States.pdf