2019

Lineárna a polynomiálna regresia

PB016: UMĚLÁ INTELIGENCE I

MATÚŠ ŠIKYŇA (485591)

Obsah

Úvod

Vo svete umelej inteligencie a strojového učenia existuje mnoho algoritmov, ktoré sa triedia do rôznych kategórií, závislých od problému (napr. klasifikácia, regresia, zhlukovanie, ...).

Jednou z týchto kategórií je učenie s učiteľom (ang. supervised learning), kde máme dvojice vstupných a výstupných hodnôt trénovacích dát a algoritmus sa snaží nájsť funkciu, ktorá čo najlepšie aproximuje hodnoty pre nové dáta.

V tejto práci si predstavíme regresnú analýzu, lineárnu a polynomiálnu regresiu, čo sú fundamentálne modely pre strojové učenie s učiteľom.

Regresná analýza

Označenie štatistických metód pre odhad vzťahu medzi jednou alebo viacerými premennými. Premenná, ktorá závisí od iných premenných sa nazýva závislá (ang. dependent), zvykne sa označovať Y. Premenná, na ktorej je iná premenná závislá sa nazýva nezávislá (ang. independent), alebo prediktor, zvykne sa označovať X.

Existuje mnoho typov regresie, medzi najznámejšie patria: lineárna regresia, polynomiálna regresia, SVR (Support Vector Regression), regresné rozhodovacie stromy.

Regresná analýza sa využíva hlavne na predikciu hodnôt na základe parametrov, identifikácia riskových faktorov, klasifikáciu alebo rôzne prognostiky a analýzy.

Lineárna regresia

Teória

Lineárna regresia je jeden z najčastejšie používaných typov regresných techník. Popisuje lineárny vzťah medzi závislou premennou Y (snažíme sa predikovať) a nezávislou premennou X (dostaneme na vstupe). Vzťah zapíšeme ako:

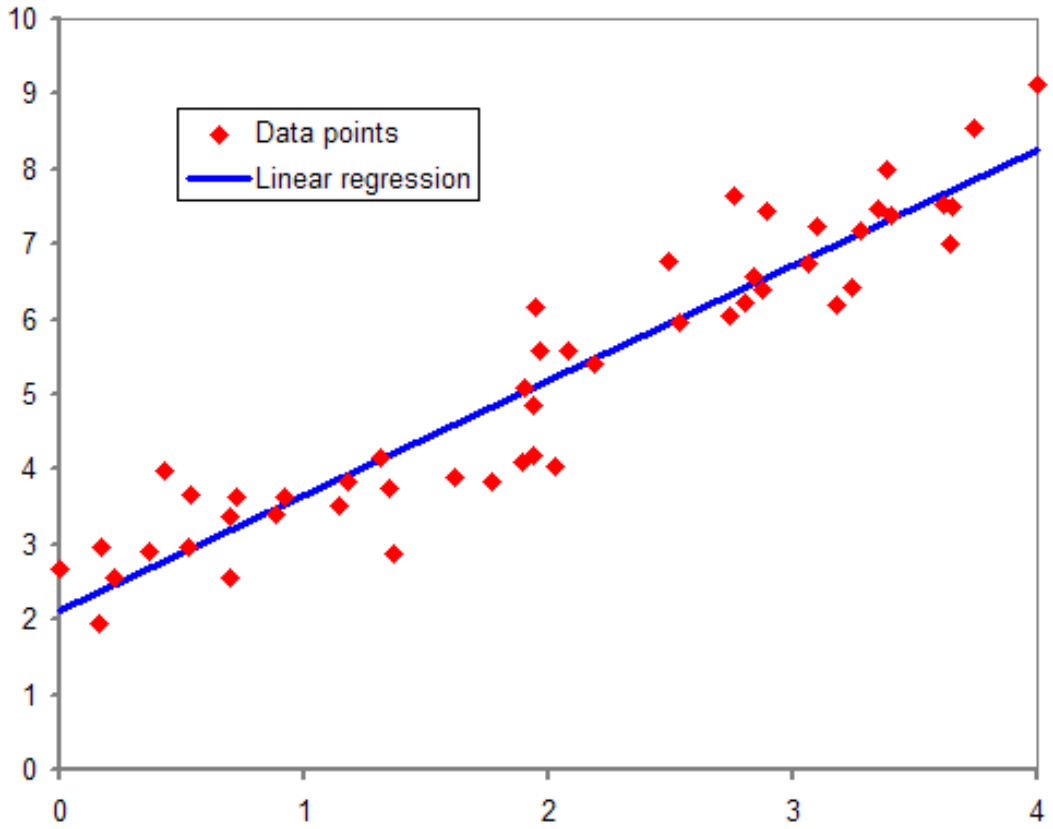
$$Y = a + bX + \varepsilon$$

Premenné a a b sú modelové koeficienty a sú to práve tieto hodnoty, ktoré sa učia. Hodnota ε je náhodná chyba, ktorá zachytáva ostatné faktory, ktoré majú vplyv na Y a sú nezávislé od X. Hodnoty koeficientov sa môžu vypočítať použitím rôznych optimalizačných algoritmov, často sa však používa metóda Gradient Descent.

Cieľom trénovania je nájdenie takých koeficientov, aby bola účelová funkcia minimálna a to minimalizovaním rozdielu medzi predikovanou hodnotou a skutočnou hodnotou pri trénovaní. Najčastejšia účelová funkcia je MSE (Mean Squared Error), ktorá berie predikovanú a skutočnú hodnotu a vypočíta ich štvorcový rozdiel.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - Y'_i)^2$$

Hodnota n je počet predikcií, Y je vektor skutočných hodnôt, Y' je vektor predikovaných hodnôt.



Použitie

Lineárna regresia je široko používaná v biologických, behaviorálnych a sociálnych vedách, ako aj v epidemiológií, finančnom svete, ekonomike alebo ekológií.

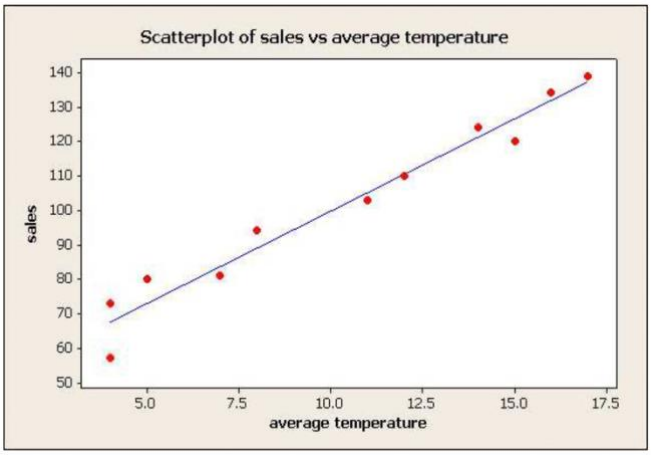
Príklady použitia v ekonomike:

- Predikcia spotreby zdrojov súkromných domácností, podnikov alebo štátov, čo je dôležitá súčasť pre výpočet hrubého domáceho produktu (HDP).
- Výpočet exportu a importu pre krajiny pre rôzne komodity.
- Sledovanie vývoja ponuky a dopytu na pracovnom trhu.

Iné príklady použitia:

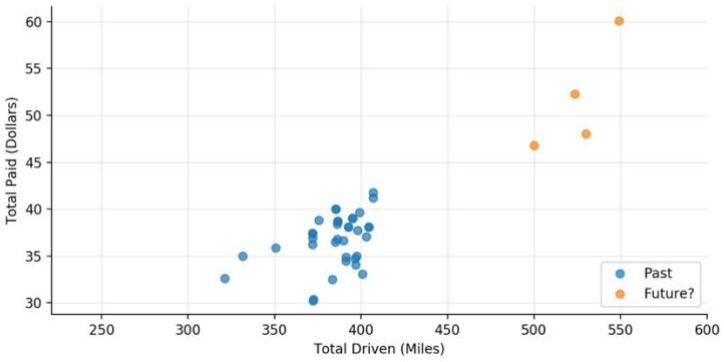
- Počet predaných zmrzlín vzhľadom na teplotu

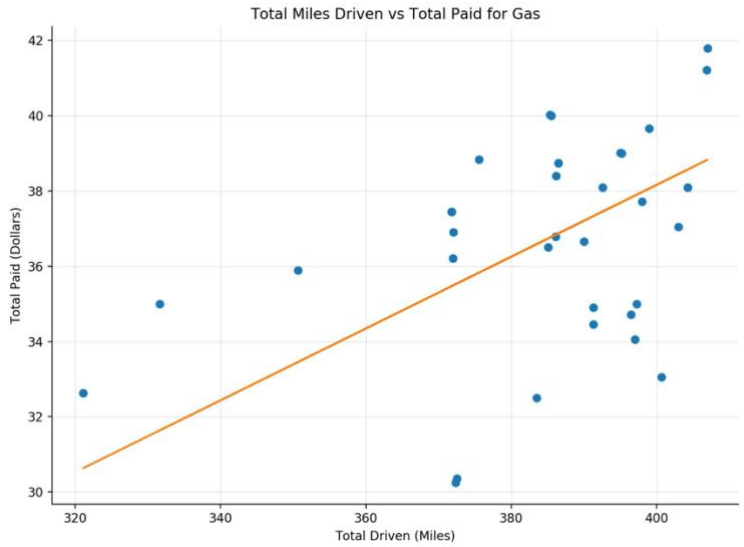
Month	Average Temp (°C)	Sales (£ 000's)
January	4	73
February	4	57
March	7	81
April	8	94
May	12	110
June	15	124
July	16	134
August	17	139
September	14	124
October	11	103
November	7	81
December	5	80



- Koľko peňazí vymedziť na benzín na cestu?

	A	B
1	Total Payed	Total Miles
2	36.66	390
3	37.05	403
4	34.71	396.5
5	32.5	383.5
6	32.63	321.1
7	34.45	391.3
8	36.79	386.1
9	37.44	371.8
10	38.09	404.3
11	38.09	392.6
12	38.74	386.49
13	39	395.2
14	40	385.50
15	36.21	372
16	34.05	397
17	41.79	407
18	30.25	372.33
19	38.83	375.6
20	39.66	399





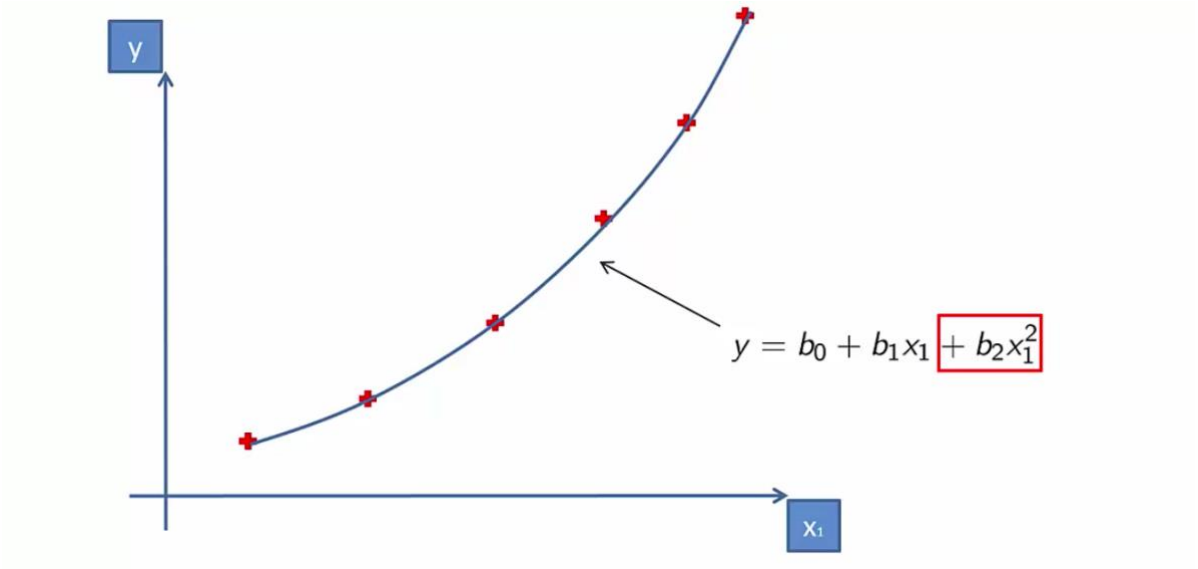
Polynomiálna regresia

Teória

Polynomiálna regresia popisuje polynomiálny vzťah medzi závislou premennou Y (snažíme sa predikovať) a nezávislou premennou X (dostaneme na vstupe). Tento polynóm môže byť n-tého stupňa. Všeobecný vzťah zapíšeme ako:

$$Y = a + bX + cX^2 + dX^3 + \cdots X^n + \varepsilon$$

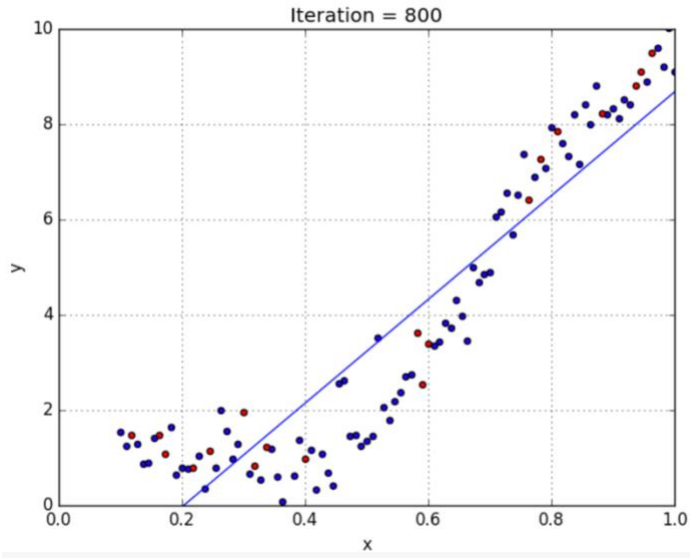
Princíp výpočtu modelových koeficientov je ako pri lineárnej regresií.

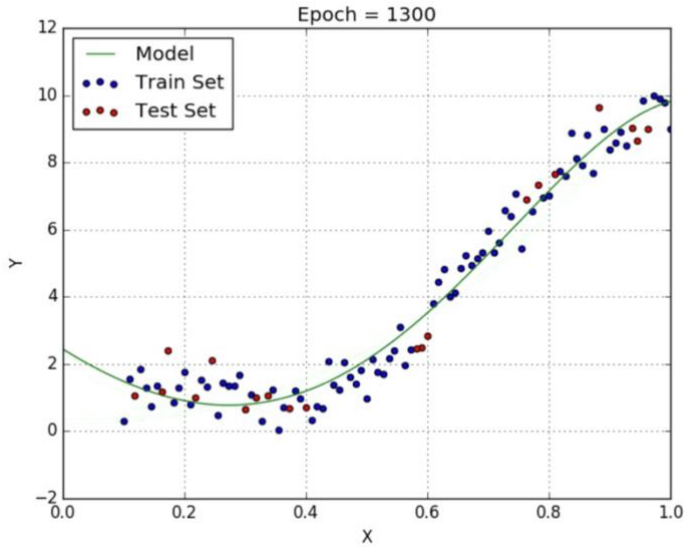


Polynomiálna regresia poskytuje najlepšiu aproximáciu vzťahu medzi závislou a nezávislou premennou. Dokáže sa prispôsobiť rôznym dátam. Medzi nevýhody patrí nepresnosť, ktorá môže byť spôsobená pri jednom alebo viacerých neobvyklých výkyvoch v dátach.

Veľmi časté je použitie kvadratickej regresie, čo je polynomiálna regresia druhého stupňa tvaru:

$$Y = a + bX + cX^2 + \varepsilon$$





Použitie

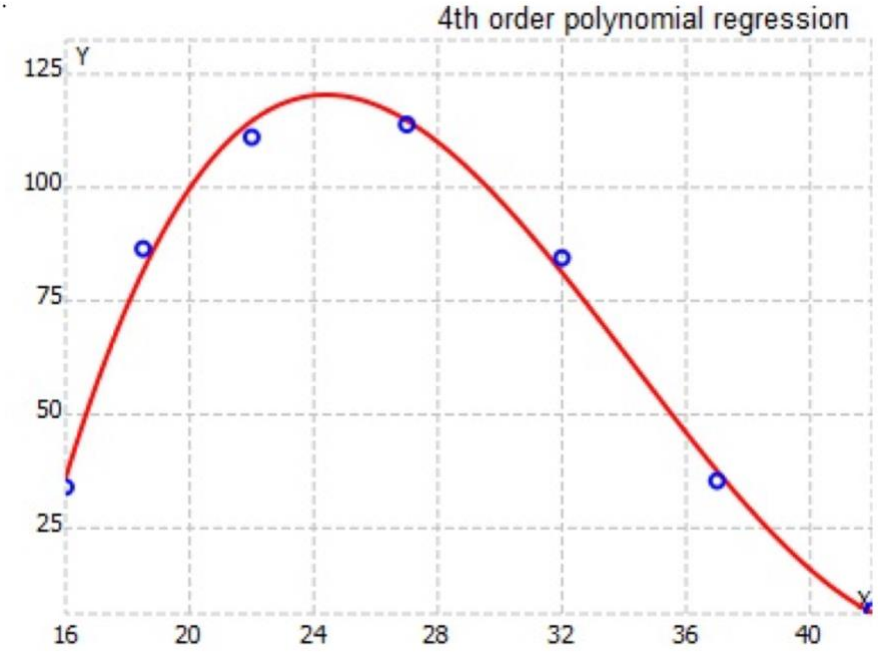
Polynomiálna regresia má podobné využitie ako lineárna regresia. Pri viac rozptýlených hodnotách je vhodnejšie použiť polynomiálnu, pretože lineárna krivka môže byť v tomto prípade veľmi nepresná. Polynóm môže byť taktiež nepresný, najmä pri použití vysokého stupňa, kde môže nastať že sa naučí určiť hodnoty trénovacích dát až príliš presne a nebude môcť dobre predikovať iné hodnoty (overfitting).

Príklady použitia:

- Meranie progresu epidémie alebo pandémie.
- Distribúcia izotopov uhlíka v sedimentoch jazier.

- Priemerný počet narodení detí u žien rôzneho veku

Age	Births/1000 women
16	34
18,5	86,5
22	111,1
27	113,9
32	84,5
37	35,4
42	6,8
45	We will calculate this value



$$\text{polyfit}(xz, 4) = -5.368\text{e-}5\, x^4 + 0.038\, x^3 - 3.479\, x^2 + 105.811\, x - 916.689$$

$$\text{replace symbols}\left(\text{polyfit}(xz, 4), x, 45\right) = 2.601$$

The average number of births per 1000 women of age 45 is 2.6.

Záver

Stručne sme si predstavili lineárnu a polynomiálnu regresiu, či už po teoretickej stránke alebo poukázaním na použitie. Dané metódy sú vhodné pre použitie keď hľadáme vzťah medzi premennými, určitú postupnosť alebo to z pohľadu na graf uznáme za vhodné.

Lineárnu regresiu som si vyskúšal v jednom projekte pre predmet IV127: Seminář laboratoře adaptabilní výuky. Ukázalo sa, že to v mojom prípade nebol najlepší prístup a nakoniec som zvolil iný algoritmus.

Zdroje

https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/

https://statistikapspp.sk/linearna-regresna-analyza/

https://medium.com/datadriveninvestor/regression-in-machine-learning-296caae933ec

https://towardsdatascience.com/introduction-to-linear-regression-and-polynomial-regression-f8adc96f31cb

https://en.wikipedia.org/wiki/Gradient_descent

https://www.freecodecamp.org/news/machine-learning-mean-squared-error-regression-line-c7dde9a26b93/

http://matematikabezproblemov.webjet.sk/domov/studijne-materialy/matematika-vs/pravdepodobnost-statistika/korelacna-regresna-analyza/

https://en.wikipedia.org/wiki/Linear_regression#Applications

https://towardsdatascience.com/polynomial-regression-bbe8b9d97491

https://www.jeremyjordan.me/polynomial-regression/

https://medium.com/coinmonks/polynomial-regression-11bec9262d64

https://www.theanalysisfactor.com/regression-modelshow-do-you-know-you-need-a-polynomial/

https://www.geeksforgeeks.org/python-implementation-of-polynomial-regression/

http://www.mas.ncl.ac.uk/~nag48/teaching/MAS1403/notes7slr.pdf

https://statisticsbyjim.com/regression/predictions-regression/

https://towardsdatascience.com/linear-regression-in-real-life-4a78d7159f16

https://www.quora.com/Whats-the-point-of-polynomial-regression-if-I-can-just-use-multiple-linear-regression

https://labdeck.com/examples/curve-fitting/curve-fitting-in-real-life.pdf?39fd30&39fd30