Uczenie maszynowe - laboratoria 1

Wstęp do uczenia maszynowego, regresja liniowa i wielomianowa, regularyzacja

Dawid Wiśniewski oraz A.M.

12 grudnia 2020

Plan zajęć

Potrzebne narzędzia do pobrania

Python 3.x (www.python.org) + sci-kit learn

Pakiet Anaconda (www.anaconda.com) w wersji Individual Edition. Jeśli masz już Pythona zwróć uwagę na pytania w trakcie instalacji.

Jupyter Notebook (jupyter.org) lub JupyterLab niezbędny do obsługi plików *.ipynb

Pamiętaj o zaznaczeniu chęci dodania ścieżek do zmiennej PATH.

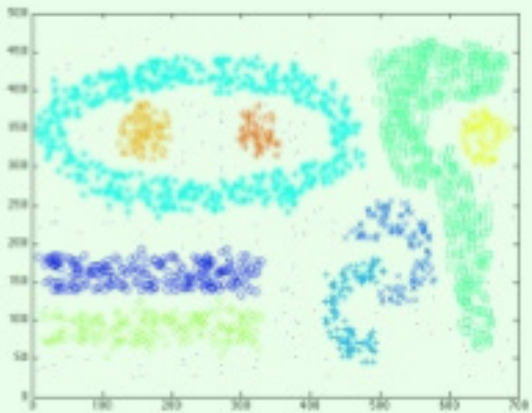
Definicja uczenia maszynowego

Definicja

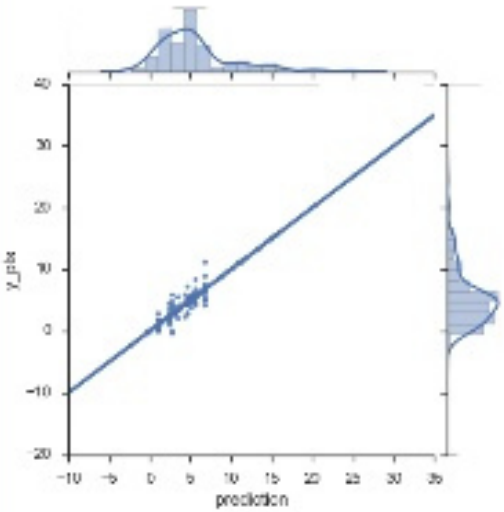
- Integralna część dziedziny sztucznej inteligencji.
- Uczenie się : autonomiczna zmiana systemu na podstawie uzyskanych doświadzeń w celu poprawy jakości jego działania.
- Uczenie maszynowe : programowanie komputerów tak, aby miały zdolność uczenia się.
- Jak ? Modelowanie wiedzy na podstawie danych.

Typy uczenia maszynowego

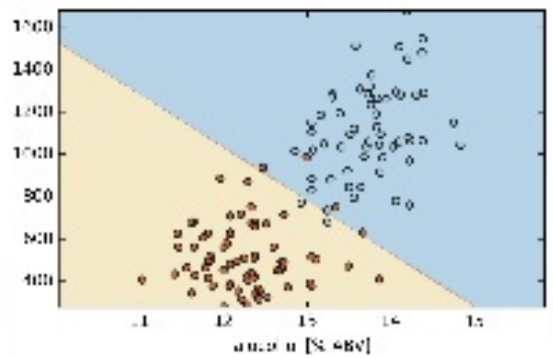
Uczenie nienadzorowane



Uczenie nadzorowane



Regresja



Klasyfikacja

Zastosowania

Udane

- Rozpoznawanie mowy (YouTube)
- Odpowiadanie na pytania (IBM Watson)
- Granie w gry (AlphaGO)
- http ://machinelearningmastery.com/inspirational-applications-deep-learning/

Oraz te mniej udane

- Microsoft Tay...

Przykład z życia

Wyobraźmy sobie Polskę początku lat 90, mamy na imię Janusz i chcemy otworzyć intratny biznes – budkę z kebabem.

Dla uproszczenia przyjmijmy, że w jednym mieście powstaje maksymalnie 1 budka z kebabem i wszyscy mieszkańcy czasami z niej korzystają.

Janusz chce przewidzieć swoje przychody z budki, ale jedyne o czym myśli, a co może mieć wpływ na zysk, to liczba mieszkańców miasta, w którym dana budka ma być otwarta.

Przykład z życia - cd

Janusz odkrył, że co roku organizowany jest zlot fanów sosu tysiąca wysp, na którym spotykają się ludzie, którzy otworzyli własne kebaby.

Pojechał więc na ten zlot i popytał znajdujących się tam ludzi jak liczebność ich miast przekłada się na zyski generowane z budki z kebabem.
Zebrane wyniki zapisał w tabelce :

Rozmiar miasta a zyski z budki z kebabem		
Miasto	Liczba mieszkańców	Miesięczny przychód
Kakulin	198	107
Halinów	2007	997
Chodzież	24000	10708
Koszalin	215000	99500

Ile może zarobić Janusz, zakładając, że jest Januszem z Warszawy ? (2 mln mieszkańców)

Przykład z życia - cd

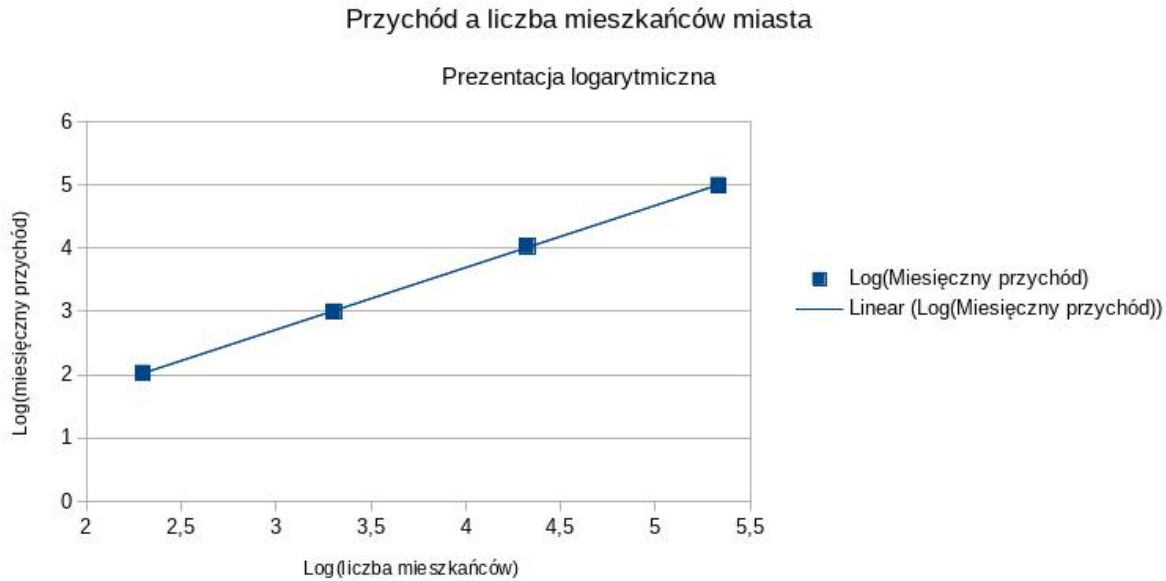
Janusz odkrył, że co roku organizowany jest zlot fanów sosu tysiąca wysp, na którym spotykają się ludzie, którzy otworzyli własne kebaby.

Pojechał więc na ten zlot i popytał znajdujących się tam ludzi jak liczebność ich miast przekłada się na zyski generowane z budki z kebabem.
Zebrane wyniki zapisał w tabelce :

Rozmiar miasta a zyski z budki z kebabem		
Miasto	Liczba mieszkańców	Miesięczny przychód
Kakulin	198	107
Halinów	2007	997
Chodzież	21000	10708
Koszalin	215000	99500
Warszawa	2000001	

Ile może zarobić Janusz, zakładając, że jest Januszem z Warszawy ? (2 mln mieszkańców)

Wizualizacja danych z tabelki



Model liniowy

Janusz widzi, że pomiędzy liczbą mieszkańców i zyskami istnieje silna zależność liniowa. Co prawda punkty wyznaczone na podstawie rozmów z innymi właścicielami nie leżą dokładnie na wyznaczonej linii, jednak widać, że linia ta najlepiej aproksymuje trend, który odkrywamy w danych.

Linia ta, użyta może zostać do wyznaczenia zysków dla liczebności miasta, którego dotąd nie obserwowano, a więc Janusz może wyznaczyć zysk podstawiając liczebność Warszawy (2000001) do najlepszej funkcji odwzorowującej dane :
$f(x) = 0,46146767x + 347,02294$

Regresja liniowa

Regresja liniowa – sprowadza się do poszukania takiej prostej, która najlepiej oddaje charakterystykę danych – najlepiej do nich pasuje.

Polega na wyznaczeniu wag prostej :
$y = ax + b$: : w przypadku jak wyżej – kiedy jedna cecha jest użyta do wnioskowania (liczba mieszkańców)

lub w ogólności $y = a_1 x_1 + a_2 x_2 + ... + a_n x_n + b = \vec{a}^T \vec{x} + b$ – kiedy liczba cech do wnioskowania jest = n (poza liczbą mieszkańców możemy użyć także dodatkowych cech, takich jak ilość innych restauracji, ilość studentów itp)

Wyznaczanie parametrów regresji - Funkcja kosztu

Zacznijmy od zdefiniowania tzw. funkcji kosztu (błędu), która informuje nas o tym jak bardzo nasza aktualna aproksymacja się myli. Funkcja taka ma tym większą wartość, im gorzej dopasowuje się do istniejących punktów.

Jedną z najpopularniejszych funkcji kosztu, która dobrze sprawdza się w przypadku regresji jest błąd średniokwadratowy :

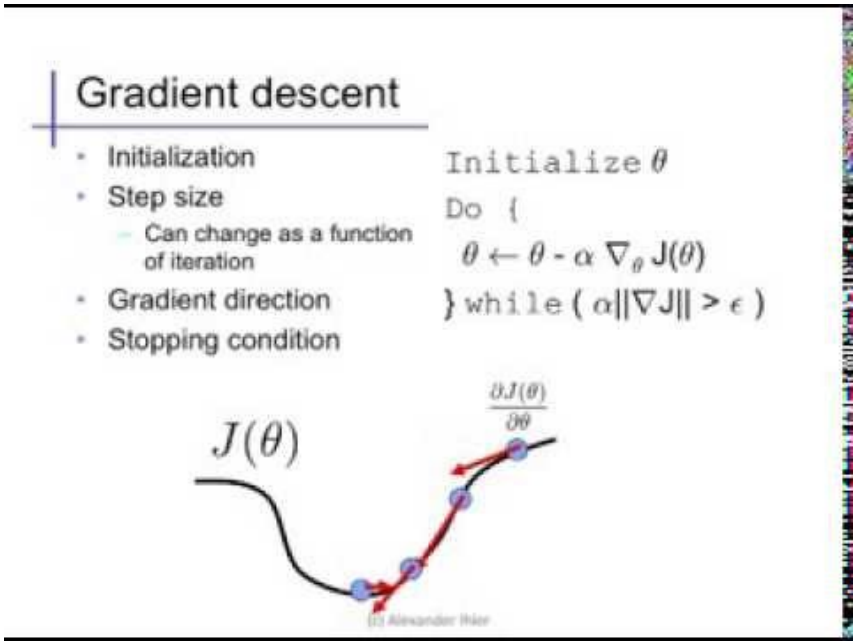
$$J(\vec{a}, b) = \frac{1}{2m} \sum_{i=1}^{m} (\vec{a}^T \vec{x^{(i)}} + b - y^{(i)})^2 \tag{1}$$

Funkcja kosztu jest dobra, kiedy :

- jest różniczkowalna
- w miarę możliwości nie wprowadza minimów lokalnych

Naszym celem jest minimalizacja funkcji $J(\vec{a}, b)$

Wyznaczanie parametrów regresji - Algorytm spadku gradientowego

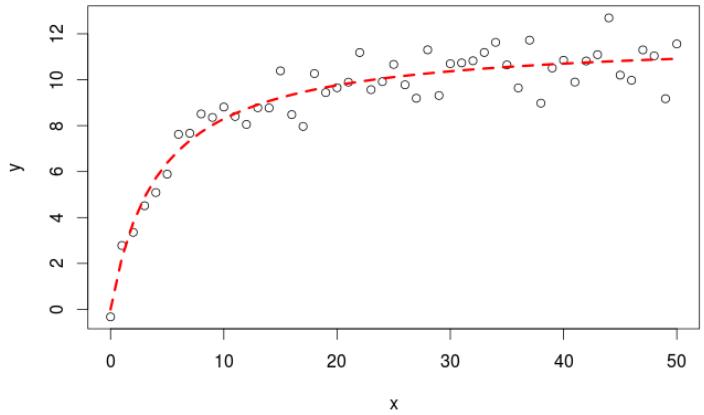


Podążaj iteracyjnie w kierunku minimum używając gradientu (aktualizuj wszystkie wagi w jednym kroku !)
W modelu liniowym z jedną cechą, wagi aktualizowane w następujący sposób (w każdym kroku aktualizujemy zarówno parametr a jak i b) :

$$a = a - \alpha \frac{1}{m} \sum_{i=1}^{m} (ax^{(i)} + b - y^{(i)}) x^{(i)} \tag{2}$$

$$b = b - \alpha \frac{1}{m} \sum_{i=1}^{m} (ax^{(i)} + b - y^{(i)}) \tag{3}$$

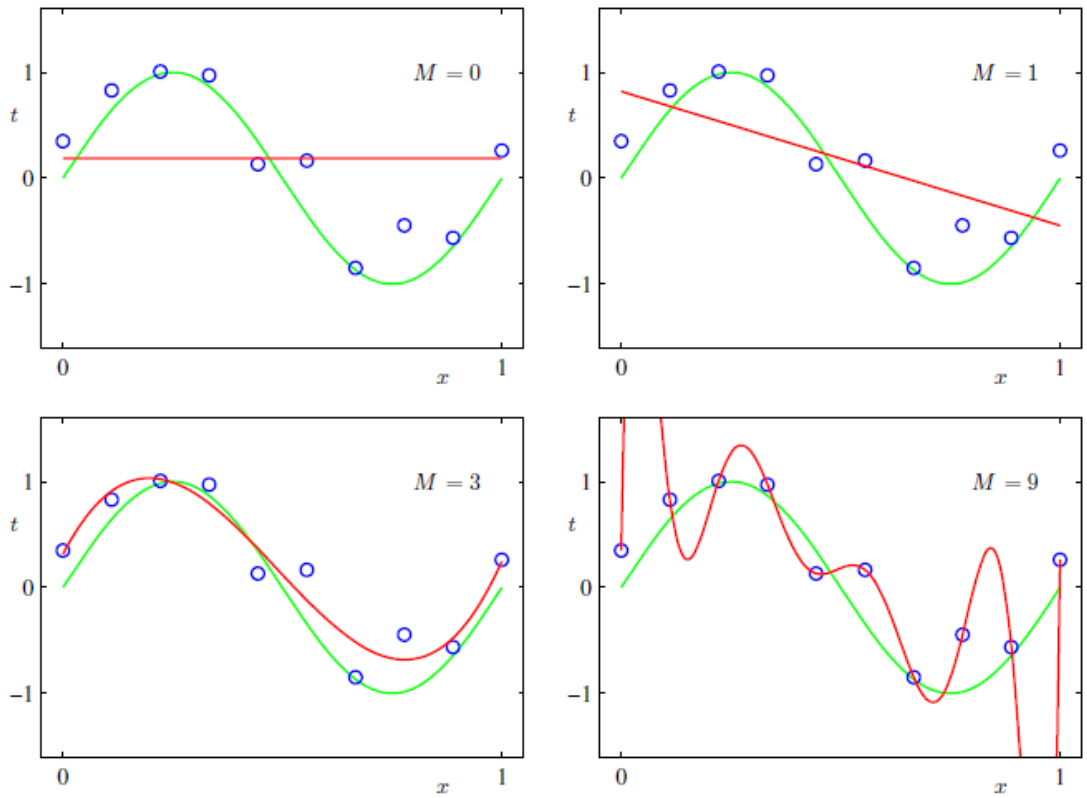
Przypadek nieliniowy



Nie zawsze model liniowy jest odpowiednim do naszych danych. Kiedy widzimy, że funkcja liniowa się nie sprawdzi, możemy aproksymować zbiór danych wielomianem k-tego stopnia, którego model wygląda następująco :

$$y = a_1 x^1 + a_2 x^2 + ... + a_k x^k + b = \sum_{i=1}^{k} a_i x^i + b \tag{4}$$

Czym jest przeuczenie



Kiedy model staje się zbyt skomplikowany (np. poprzez wybór zbyt wysokiego stopnia wielomianu) może dojść do przeuczenia – stanu, w którym model uczy się szumu z danych (odchyleń nie mających wpływu na realny trend).

Taki model bardzo kiepsko sprawdzi się na nieobserwowanych dotąd danych. Potrzeba mechanizmu, który potrafi zapobiegać przeuczeniu.

Regularyzacja - regresja Ridge

Aby zminimalizować ryzyko przeuczenia stosuje się mechanizm regularyzacji. Wprowadza on dodatkową karę za duże wartości wyuczonych wag, co sprawia, że wygenerowana funkcja staje się lepiej dopasowana do rzeczywistego trendu w danych.

Funkcja kosztu z użyciem regularyzacji (typu Ridge Regression) dla modelu liniowego przyjmuje postać :

$$J(\vec{a}, b) = \frac{1}{2m} \sum_{i=1}^{m} (\vec{a}^T \vec{x^{(i)}} + b - y^{(i)})^2 + \lambda \vec{a}^T \vec{a} \tag{5}$$