Statystyka

Izabela Czarny Aleksandra Hobot Maciej Kucharski

25 stycznia 2022

Spis treści

Wprowadzenie

Nasz temat został podzielony na kilka mniejszych segmentów, jednak postanowiliśmy wybrać tylko kilka - najciekawszych - z nich. Wśród wybranych przez nas tematów znalazły się: obliczanie średniej ruchomej, przybliżenie regresji liniowej, znajdowanie wszystkich unikalnych par na liście oraz używanie łańcucha Markowa do generowania tekstu.

Obliczanie średniej ruchomej

Średnia ruchoma – metoda statystyczna używana do analizy szeregów czasowych (szereg czasowy – realizacja procesu losowego, którego dziedziną jest czas; ciąg informacji uporządkowanych w czasie). Znajduje zastosowanie w finansach, zwłaszcza w analizie technicznej.

Obliczanie średniej ruchomej

Zestawienie listy liczb w jedną reprezentatywną liczbę można wykonać, obliczając średnią. Równanie na średnią arytmetyczną polega na zsumowaniu wszystkich wartości i podzieleniu przez ich liczbę. Jeśli wartości są zbyt duże, to suma może być przepełniona. Sytuacja ta występuje dla średniej dwóch skrajnych intów: -2^{29} oraz 2^{29-1}

Obliczanie średniej ruchomej

Średnia ruchoma próbuje uniknąć tej wady. Używamy strategii wygładzania wykładniczego, co oznacza, że liczby, które widziano wcześniej, przyczyniły się do zmniejszenia wykładniczo wartości średniej kroczącej. Może być używany w sytuacjach wykrywania wahań cen lub skoków neuronów w sieci neuronowej.

Obliczanie średniej ruchomej

Równanie średniej ruchomej jest następujące:

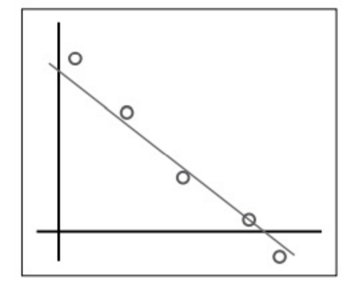
$$\boxed{\begin{aligned} &s_0 = x_0 \\ &s_t = \alpha x_{t-1} + (1-\alpha)s_{t-1},\ t > 0 \end{aligned}}$$

Przybliżenie regresji liniowej

Regresja liniowa – w modelowaniu statystycznym, metody oparte o liniowe kombinacje zmiennych i parametrów dopasowujących model do danych. Dopasowana linia lub krzywa regresji reprezentuje oszacowaną wartość oczekiwaną zmiennej y przy konkretnych wartościach innej zmiennej lub zmiennych x. W najprostszym przypadku dopasowana jest stała lub funkcja liniowa.

Przybliżenie regresji liniowej

Mając listę punktów, możemy oszacować najlepszą linię dopasowania, korzystając z przydatnej biblioteki - Statistics.LinearRegression. Oblicza różnicę najmniejszych kwadratów między punktami, aby oszacować najlepszą linię dopasowania. Przykład regresji liniowej punktów można zobaczyć na poniższym rysunku:



Przybliżenie regresji liniowej - jak to działa?

Rozwiązaniem tego problemu jest próba uzyskania tak małych, jak to możliwe, sum kwadratów „błędów" między prawą i lewą stroną tego równania, czyli znaleźć minimum funkcji.
Podstawowe obliczenia obejmują znalezienie średniej i wariancji dwóch zmiennych losowych, jak również kowariancji między nimi. Jeżeli zerkniemy do kodu źródłowego znajdziemy tam równania:

$$\alpha = \mu Y - \beta * \mu X$$

$$\beta = covar(X,Y)/\sigma 2X$$

$$f(x) = \beta x + \alpha$$

Znajdowanie wszystkich unikalnych par na liście

Unikalne pary na liście lub kombinacja bez powtórzeń – dowolny podzbiór zbioru skończonego.
Mamy zbiór n-elementowy:

$$0 \leqslant k \leqslant n$$

w którym k-elementowy podzbiór jest określany jako kombinacja z n po k.
Liczba kombinacji z n po k wyraża się wzorem:

$$C_n^k = \binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Znajdowanie wszystkich unikalnych par na liście

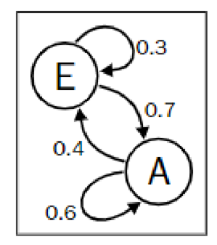
Porównywanie wszystkich par pozycji jest powszechnym zabiegiem w analizie danych. W tym przepisie Tworzymy listę par elementów z listy elementów. Na przykład, jeśli istnieje lista [1, 2, 3], stworzymy listę wszystkich możliwych par [(1, 2), (1, 3), (2, 3)].
Zauważ, że kolejność parowania nie ma znaczenia. Tworzymy listę unikalnych par, więc możemy porównać każdą pozycję z każdą inną pozycją na liście.

Używanie łańcucha Markowa do generowania tekstu

Łańcuch Markowa to system, który przewiduje przyszłe wyniki systemu w obecnych warunkach w oparciu o wyniki poprzednie. W naszym przykładzie używamy łańcuch Markowa do generowania tekstu na podstawie wcześniej załadowanego pliku tekstowego.

Używanie łańcucha Markowa do generowania tekstu

Graficzna reprezentacja łańcucha jest pokazana na poniższym rysunku:



Jak widać na rysunku, węzeł E ma 70 procent prawdopodobieństwa, że znajdzie się w węźle A i 30 procent prawdopodobieństwa, że pozostanie na miejscu. Podobnie węzeł A ma 40 procent szans na przejście do węzła E i 60 procent szans na pozostanie w miejscu.

Bibliografia

- plik 3Statistics_Haskell
- http://en.wikipedia.org/wiki/File:Correlation_coefficient.gif
- http://kfe.fjfi.cvut.cz/ kucharik/edu/PF/1/lit/Landau_Paez-CP_Python-2018.pdf
- https://pl.wikipedia.org/wiki/Statystyka
- https://zerodha.com/varsity/chapter/moving-averages/
- http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm
- https://brilliant.org/wiki/markov-chains/

The End