Notatki rozdział 1-2

niedziela, 1 listopada 2020 13:18

Odchylenie standardowe - pierwiastek kwadratowy z wariancji - przedstawia rozrzut wartości Wariancja - średnia arytmetyczna kwadratu odchyleń od wartości średnich przykładów [65s]

Reguła 68 - 95 - 99,7 - odnosi się do funkcji Gaussa i mówi, że 68% wartości leży w odległości 1 σ od wartości oczekiwane, 95 w odległości 2 σ a 99,7 w odległości 3 σ. [65s]

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$, gdzie

μ – Średnia arytmetyczna

x_i – i-ty element

N – Ilość elementów

Rozkład długo ogonowy - wartości histogramu rozciągają się znacznie bardziej po prawej jego stronie niż po lewej - stanowi to problem dla niektórych algorytmów i dobrze jest spróbować sprowadzić te wartości do funkcji Gaussa [67s]

Losowanie warstwowe - dzielimy populację na jednorodne podgrupy zwane warstwami i przypisujemy im odpowiednią ilość przykładów odpowiadającą rzeczywistemu rozkładowi - Sprawdza się w przypadku mniejszej ilości danych, gdzie niektóre atrybuty mogą mieć znaczący wpływ na predykcji. Wtedy zamiast dzielić zbiór testowy i treningowy losowo, lepiej jest podzielić go tak, by w zbiorze testowym zachowany został rozkład znaczących atrybutów taki, jaki występuje w rzeczywistości. Po podziale powinno się usunąć utworzone warstwy. [69s]

Współczynnik korelacji liniowej (Pearsona) - korelacja między każdą parą

wartości, gdzie dla rosnącej wartości x, wartość y maleje lub rośnie (Nie uwzględnia przypadku, gdy np. wartość x zbliża się do 0 - jest to korelacja nieliniowa) [73-74s]

Kodowanie gotącojedynkowe - polega na przygotowaniu macierzy zero-jedynkowej dla atrybutów nienumerycznych, tak, aby algorytm uczenia maszynowego mógł z nimi pracować. Polega to na wyznaczeniu osobnych atrybutów dla każdej kategorii Atrybutu kategorialnego, w którym 1 oznacza się daną kategorię, a zerami resztę, i tak dla każdej kategorii osobno.
Przechowywanie takiej liczby 0 dla atrybutów które mają wiele kategorii mogłoby być marnotrawstwem, dlatego używa się macierzy rzadkich, w których przechowywana jest jedynie pozycja jedynki[80s]

Skalowanie cech - Przeważnie używane dwa:

- Standaryzacja - Skalowanie polegające na odjęciu wartości średniej i podzieleniu jej przez wariancję
    - Wady: nie skaluje do określonego zakresu co może utrudnić zadanie niektórym algorytmom
    - Zalety: nie jest tak wrażliwa na elementy odstające
- Normalizacja (przekształcenie min-max) - proste przekształcenie odejmujące od danej wartości min i podzielenie wyniku przez max-min. Sprowadza wyniki do zakresu 0-1.
    - Wady i zalety odwrotne do Standaryzacji[82s]

Pipeline (potoki przekształcające) - Pozwalają na przeprowadzanie w łatwy sposób wielu wcześniej opracowanych operacji przekształcających w odpowiedniej kolejności.[83-84s]

MSE - błąd średniokwadratowy w którym liczymy sumę kwadratów różnic między wartościami przewidywanymi a wartościami faktycznymi i dzielimy przez ilość elementów. Pozwala nam wyznaczyć RMSE który jest pierwiastkiem z błędu średniokwadratowego i pokazuje, o ile przewidywania średnio różnią się od rzeczywistych wartości. Czym mniej tym lepiej.[85-86s]

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - y_i^p)^2$$

y_i^p – Wartość prognozowana

Radzenie sobie z niedotrenowaniem modelu:

- Wybór innego algorytmu
- Zmniejszenie ograniczeń modelu
- Wybór lepszych cech

Dzięki sprawdzaniu wyników uzyskanej predykcji z użyciem zbioru treningowego, możemy lepiej wykryć przetrenowanie modelu. Pomocna w tym może być walidacja krzyżowa. Jeśli RMSE znacząco różni się w przypadku walidacji krzyżowej i zwykłego obliczenia RMSE dla zbioru testowego, na korzyść zwykłego, to model najprawdopodobniej uległ przetrenowaniu [88s].

Radzenie sobie z przetrenowaniem modelu:

- Uproszczenie modelu
- Regularyzacja
- Uzyskanie większej ilości danych

Regulowanie modelu z użyciem metody przeszukiwania siatki - polega na odpowiednim dobraniu wartości hiperparametrów i szukaniu dobrej ich kombinacji.

Losowe przeszukiwanie - w przeciwieństwie do metody przeszukiwania siatki, nie sprawdza wszystkich możliwych kombinacji w celu dobrania wartości hiperparametrów.

Dzięki analizie najlepszych uzyskanych modeli możemy uzyskać wiele informacji na temat problemu jaki próbujemy rozwiązać, np. wskazać istotność atrybutów.

Jeśli zbyt intensywnie stroimy hiperparametry, możemy pogorszyć działanie modelu.