Machine Learning

Fafer77

19 grudnia 2024

1 Definitions

1. Measures of central tendency (they describe a central position of your data):
   (a) Mean
   (b) Mode
   (c) Median
2. Measures of spread (describe how spread out your data is - clumped together or spread far apart):
   (a) Range
   (b) Quartiles
   (c) Standard deviation
   (d) Variance
3. Training set - data used for training
4. Training example (sample) - A single instance used in machine learning to train a model. It typically consists of input features and, in supervised learning, an associated label or target value.
5. Accuracy - the ratio of correctly predicted instances to the total number of instances in the dataset.

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

6. Problem regresyjny: przewidywanie wartości na podstawie cechy wejściowej

7. Wydobywanie cech (ang.feature extraction) - polega na znalezieniu skorelowanych cech i połączenie ich w jedną

8. Classification task - involves predicting a discrete label or category for a given input based on its features e.g 'spam' or 'not spam'

9. Data mining - analyzing huge amount of data to find some patterns

10. Types of learning:

    (a) Uczenie nadzorowane (ang. supervised learning) - dane uczące przekazywane algorytmowi zawierają dołączone rozwiązania problemu, tzw. etykiety (ang. labels).
    Typowe zadania systemu nadzorowanego:

        i. Klasyfikacja np. filtr spamu
        ii. Przewidywanie docelowej (ang. target) wartości numerycznej np. cena samochodu przy użyciu określonego zbioru cech. Ten typ zadania nosi nazwę regresji.

    (b) Uczenie nienadzorowane (eng. unsupervised learning) - dane uczące są nieoznakowane. System próbuje się uczyć bez nauczyciela.

        i. Analiza skupień
        ii. Algorytm wizualizujący, redukcja wymiarowości - cel uproszczenie danych bez utraty nadmiernej ilości informacji.
        iii. Wykrywanie anomalii (ang. anomaly detection) - np. nietypowe transakcje karty kredytowej w celu zapobieganiu nielegalnym operacjom, wykrywanie usterek produkcyjnych.
        iv. Wykrywanie nowości (ang. novelty detection)
        v. Uczenie przy użyciu reguł asocjacyjnych (ang.association rule learning) - analiza ogromnej ilości danych i wykrycie interesujących zależności pomiędzy atrybutami.

    (c) Uczenie półnadzorowane (ang. semisupervised learning) - część danych jest oznakowana, a większość nie, bo etykietowanie jest czasochłonne i kosztowne.

    (d) Uczenie samonadzorowane (ang.self-supervised learning) - wygenerowanie w pełni oznakowanego zestawu danych z zestawu całkowicie nieoznakowanego.

    (e) Uczenie transferowe (ang. transfer learning) - korzysta się w głębokcih sieciach neuronowych.

(f) Uczenie przez wzmacnianie (ang. reinforcment learning) - system uczący tzw. agent może obserwować środowisko, dobierać i wykonywać czynności, a także odbierać nagrody lub kary. Potem uczy się samodzielnie najlepszej strategii (ang. policy), aby uzyskiwać jak największą nagrodę. Polityka definiuje rodzja działania, jakie agent powinien wybrać w danej sytuacji.

(g) Uczenie wsadowe (ang. batch learning) - system nie jest w stanie trenować przyrostowo - do jego anuki muszą wystarczyć wszystkie dostępne dane (zużywa zwykle dużo czasu i zasobów, dlatego zwykle w trybie offline). System najpierw jest uczony, a potem wdrożony do cyklu produkcyjnego i już więcej nie jest trenowany; korzysta jedynie z dotychczas zdobytych informacji. Tzw. uczenie offline (ang. offline learning). Następuje rozkład modelu (ang. model rot) albo dryf danych (ang. data drift). Rozwiązanie: systematyczne trenowanie modelu, zależne od problemu. System trenuje się od podstaw na starym i nowym zestawie za każdym razem.

(h) Uczenie przyrostowe (ang. online learning) - trenowany jest na bieżąco poprzez sekwencyjne dostarczanie danych, które mogą być pojedyncze lub przeyjować postać tzw. minipakietów (mini-batches). Każdy krok uczący jest szybki i niezbyt kosztowny.

11. out-of-core learning - uczenie pozakorowe, czyli wykorzystywanie dużych zbiorów z poza pamięci urządzenia.

12. Współczynnik uczenia (ang. learning rate) - szybkość, z jaką dostosowują się systemy do zmieniających się danych (jeden z najważniejszych parametrów systemów uczenia przyrostowego). Wysoka wartość - szybka adaptacja, ale szybko też zapomina o starych danych.

13. Uczenie z przykładów (ang. instance-based learning) - system uczy się przykładów na pamięć i nastepnie za pomocą miary podobieństwa porównuje je z wyuczonymi przykładami (lub ich podzbiorem).

14. Uczenie z modelu (ang. model-based learning) - używa się go do przewidywania, prognozowania (ang. prediciton).

15. Funkcja użyteczności (dopasowania) - mówi jak dobry jest dany model, a funkcja kosztu odwrotnie.

16. Inżynieria cech (ang. feature engineering) - proces wyboru dobrego zbioru cech uczących.

(a) Dobór cechy (ang. feature selection) - dobór najprzydatniejszych spośród dostępnych cech

(b) Odkrywanie cechy (ang. feature extraction) - łączenie ze sobą istniejących cech w celu uzyskania przydatniejszech cechy (np. redukcja wymiarowości)

(c) uzyskiwanie nowych cech z nowych danych

17. Nadmierne dopasowanie (ang. overfitting) - model dobrze sprawdza się w przypadku danych uczących, ale sam proces uogólniania nie sprawuje się zbyt dobrze. Następuje gdy model jest zbyt skomplikowany w porównaniu do ilości lub zaszumienia danych uczących. Rozwiązania:

    (a) Uproszczenie modelu poprzez wybór mniejszej ilości parametrów (np. liniowego zamiast wielomianowego)

    (b) Zdobycie większej ilości danych uczących

    (c) Zmniejszenie zaszumienia danych uczących (np. usunięcie błędnych danych lub elementów odstających)

18. regularyzacja (regularization) - ograniczenie modelu w celu jego uproszczenia i zmniejszenia ryzyka przetrenowania

19. hiperparameters - parametry algorytmu uczącego. Są wyznaczane przed rozpoczeciem procesu uczenia, a nie nabywane w trakcie (pozostają niezmienne). Duża wartość hiperparametru regularyzacji, uzyskana funkjca będzie niemal stała. Strojenie hiperparametrów stanowi istotną część tworzenia systemu maszynowego.

20. Niedotrenowanie (ang. underfitting) - występuje gdy model jest zbyt prosty, aby wyuczyć się struktur danych uczących. Rozwiązania:

    (a) Wybieraj potężniejszy model z większą liczbą parametrów

    (b) Dołączaj większa liczbę cech do algorytmu uczącego (inżynieria cech)

    (c) Zmniejszaj ograniczenia modelu (np. redukując hiperparametr regularyzacji)

21. Sposoby testowania:

    (a) Możemy po prostu wypróbować model na nowych danych, gdy będą go testować użytkownicy, ale gorzej gdy nie będzie działać dobrze

(b) Podział danych na zbiór uczący (ang. training set) i zbiór testowy (ang. test set).Trenujemy na uczącym i sprawdzamy na testowym. Uzyskany współczynnik błędu nosi nazwę błędu uogólniania (generalizacji), a dzięki zbiorowi testowemu oszacowujemy jego wartość. Mówi on jak model będzie spisywał się wobec nieznanych danych. Model przetrenowany, gdy wartość błędu uczenia jest niewielka, ale błąd uogólniania jest duży, to model jest przetrenowany.

(c) Można wytrenować wiele modeli dla różnych wartości hiperparametrów i sprawdzić, który jest najlepszy. Ale wtedy można zgeneralizować hiperparametr do test setu. Zatem stosuje się tzw. sprawdzian na odłożonych danych (ang. holdout validation). Odkładamy zbiór walidacyjny/rozwojowy (ang. validation set lub dev set) i na nim wybieramy najlepszego. Potem trenujemy wybrany model na zestawie uczącym + walidacyjnym i sprawdzamy na testowym.

22. train-dev-set - pomaga odpowiedzieć na pytanie, czy problem leży w rozkładzie danych czy w samym modelu. Train-dev-set zawiera dane podobne do training set. Gdy wyniki na nim są słabe to model nie jest wystarczająco złożony lub dane są niewystarczające. Z kolei jeśli wyniki na train-dev-set są dobre ale na dev-set słabe to problem z overfittingiem lub wskazuje różnice w rozkładach między danymi treningowymi, a walidacyjnymi.

23. Przykładowa lista kontrolna projektu uczenia maszynowego:

    (a) Okreslenie problemu i przeanalizowanie go w szerszej perpsektywie

    (b) Pozyskanie danych

    (c) Analiza danych w celu wyrkycia dodaktowych informacji

    (d) Przygotwanie danych w sposób uwidaczniający wzorce wykorzystywane przez algorytmy uczenia maszynowego

    (e) Sprawdzenie wielu modeli i stworzenie krótkiej listy najwydajniejszych z nich

    (f) Dostrojenie modeli i połączenie ich w zespoły uzyskujące jeszcze lepsze wyniki

    (g) Zaprezentowanie rozwiązania

    (h) Uruchomienie, monitorowanie i utrzymywanie systemu

1.1 Math

1. RMSE(X,h) - pierwiastek błędu sredniokwadratowego (ang. Root mean Square Error). Mówi w jakim stopniu model myli się w przewidywaniach - wraz ze wzrostem wartości błędu rośnie waga tego wskaźnika.

$$\mathrm{RMSE}(X, h) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(h(x_i) - y_i)^2}$$

2. MAE(X, h) - średni błąd absolutny (ang. Mean Absolute Error, Average Absolute Deviation). Sprawdza się lepiej w danych, gdzie wiele dystryktów odstaje od reszty.

$$\mathrm{MAE}(X, h) = \frac{1}{m}\sum_{i=1}^{m}|h(x_i) - y_i|$$

3. Normy to miary odległości. RMSE wiąże się z normą euklidesową. Zapisywana w postaci $\|x\|_2$. Błąd MAE wiąże się z $\|x\|_1$. Inaczej nazywana jest normą Manhattan, taksówkową i miejską.