Raport – uczenie maszynowe wprowadzenie

Michał Dubiel

1. **Jakie są typy uczenia maszynowego? Jakie można wyróżnić grupy? Proszę wskazać różnicę oraz zastosowanie każdego typu.**

* Uczenie nadzorowane – przy treningu model otrzymuje dane oraz dołączone do nich etykiety, które będzie musiał przewidywać dla nowych danych. Na przykład problemy klasyfikacji czy regresji. Implementowane między innymi w przewidywaniu cen mieszkań lub przydzielanie kredytów
* Uczenie nienadzorowane – zadaniem modelu jest grupowanie danych, tzw. Klusteryzacja. Używana np. przy tworzeniu baniek internetowych, czyli grupowania ludzi z podobnymi zainteresowaniami
* Uczenie przez wzmacnianie (reinforcment learning) – uczymy model zachowań poprzez system nagrody i kary. Używany np. przy treningu botów w grach komputerowych albo samochodów autonomicznych

1. **Jakie są róznice pomiedzy Machine Learningiem a AI? dlaczego preprocessing danych jest istotny w procesie trenowania modelu uczenia maszynowego.**

* Machine learning (uczenie maszynowe) jest dziedziną AI (sztucznej inteligencji). Zadaniem dziedziny sztucznej inteligencji jest tworzenie algorytmów przypominających ludzką inteligencję. Uczenie maszynowe korzysta z zaawansowanych funkcji i modeli matematycznych które w kolejnych iteracjach zmieniają swoje atrybuty (uczą się) aby osiągnąć ten cel. Przykładem algorytmu sztucznej inteligencji mogą być wczesne algorytmy przetwarzania języka naturalnego takie jak bazy semantycznych powiązań lub gramatyki bezkontekstowe.
* Preprocessing czyści dane. Częstymi problemami związanymi z danymi są:
  + Korelacja danych: pewna para danych jest ze sobą skorelowana tzn. widzimy proporcjonalne zmiany w obu przypadkach. Prowadzi to do marnowania zasobów na dane które nie poprawiają wyników działania modelu. Poprzez preprocessing możemy to rozwiązać usuwając jedne ze skorelowanych danych
  + Brak normalizacji: normalizacja ułatwia analizę i interpretację danych oraz poprawia wydajność modelu. W preprocessingu często normalizuje się dane
  + Szum: dane często posiadają niepożądane informacjie które zakłócają proces uczenia. Preprocessing może usuwać szum
* Dodatkowe plusy preprocessingu to:
  + Zapobieganie przetrenowaniu
  + Redukcja wymiarowości

1. **Kiedy sieć neuronową możemy nazwać głęboką ?**

Jeśli sieć neuronowa posiada warstwę ukrytą możemy nazwać ją głęboką

1. **Co oznacza bias/variance trade off? Czym cechuje się model posiadający dużą wariancję?**

* Model posiadający dużą wariancje jest przetrenowany, czyli jest za bardzo przystosowany(przyzwyczajony) do danych treningowych. Osiąga wspaniałe wyniki na tych danych ale nie radzi sobie z nowymi
* Bias/variance trade off oznacza zależność wariancji do bias’u czyli założeń. Na początku model posiada praktycznie same założenia i nie jest dokładny, tzn. nie generuje zadowalających (celnych) etykiet, ale odpowiedzi są blisko siebie (są spójne). Ze zwiększaniem się treningu zwiększa się wariancja czyli zwiększa się rozrzut odpowiedzi ale zmniejszają się założenia modelu co wpływa na jego celność. Za duży trening sprawia że wyniki są zbyt rozrczucone ponieważ model za bardzo przystosował się do danych treningowych. Celem dobrego treningu jest znaleźć złoty środek pomiędzy wariancją i bias’em tak aby były jak najmniejsze.

1. **Jakie warunki muszą spełnić macierze aby je pomnożyć ?**

* Liczba kolumn w jednej, musi być równa liczbie wierszy w drugiej

1. **Do czego uzyte moga byc narzedzia redukcji wymiarowosci np. PCA? Jaki inny algorytm poza PCA moze posłuzyc do redukcji wymiarowosci?**

* W przypadku posiadania za dużej liczby cech danych, używamy algorytmów redukcji wymiarowości do ich zmniejszenia
* Inne algorytmy:
  + Autoenkodery, czyli sieci neuronowe przystosowane do zmniejszania lub sumowania cech do treningu
  + Usuwanie Korelacji i Szumu

1. **Podaj 3 algorytmy rozwiązujące problem uczenia nadzorowanego.**

* Regresja liniowa
* Regresja Logistyczna
* sieci neuronowe

1. **Co to jest funkcja straty ? (loss function)**

* Funkcja oceny (jakości) modelu. Za jej pomocą oceniamy trening i wyniki modelu i dzięki niej możemy model trenować. Zazwyczaj bazuje ona na błędach modelu czyli różnicy wyjścia modelu z faktyczną etykietą danych.

1. **Opisz problem przeuczenia modelu. W jaki sposób możemy go uniknąć ?**

* Problem polega na za dużym przystosowaniu modelu do danych treningowych. Wynika on w głównej mierze ze złej metody treningowej. Model który będzie w prosty iteratywny sposób przechodził wiele razy po WSZYSTKICH danych dostępnych, za bardzo przystosuje się do nich i nie będzie w stanie poprawnie przewidywać etykiet dla nowych danych. Najlepszym rozwiązaniem jest zastosowanie preprocessing i dzieleniu danych dzięki czemu symulujemy ich różnorodność. To pozytywnie wpłynie na wyniki modelu przy nowych danych.

1. **Jakie mozna przyjac metryki oceny skutecznosci modelu regresji, a jakie dla modeli klasyfikacji?**

* Dla regresji:
  + Mean Squared Error
  + Squared Error
  + Root Mean Squared Error
* Dla klasyfikacji
  + Dokładność
  + F1 Score
  + Log Loss
  + Macierz pomyłek

1. **Jakim wynikiem na zbiorze testowym oraz jakim wynikiem na zbiorze treningowym cechują sie modele jakie cierpia na problem underfittingu, a jakim te cierpiace na overfitting?**

* Under Fitting
  + Złe wyniki na zbiorze testowym (praktycznie równe losowym etykietom)
  + Złe wyniki na zbiorze treningowym (praktycznie równe losowym etykietom)
* Over Fitting
  + Złe wyniki na zbiorze testowym
  + Nienaturlnie dobre wyniki na zbiorze treningowym

1. **Jakie według Ciebie ma zalety, a jakie wady rozwój sztucznej inteligencji, np. GPT? Jakie widzisz zastosowania GPT dla siebie?**

* Zalety:
  + Rozwiązanie problemów, które normalne podejście algorytmiczne nie byłoby wstanie rozwiązać
  + Nowe spojrzenie na różne dziedziny nauki
* Wady:
  + Rozleniwianie społeczeństwa
  + Bańka AI (giełdowa)
  + Nadmierne zaufanie
* GPT może mnie wesprzeć w dużej mierze w nauce i poznawaniu nowych technologii. Pomimo problemów przy szczegółowych faktach, sprawdza się przy powierzchownych i prostych informacjach co pozwala mi na szybkie opanowywanie i wstęp do nowych treści.