WIKAMP FTIMS

Opracowanie artykułu

Obliczenia inteligentne są wykorzystywane wtedy gdy nie posiadamy pełnej wiedzy pozwalającej nam napisać program rozwiązujący zadane zadanie. Podejście, które się wówczas stosuje polega na wyborze pewnego modelu i automatycznym doborze jego optymalnych parametrów w oparciu o dostępną, niepełną wiedzę. Dobór ten zwykle sprowadza się do optymalizacji odpowiednio zdefiniowanej funkcji celu (ang. *objective*). Typowym przykładem niepełnej wiedzy o problemie jest zbiór treningowy par danych wejściowych i odpowiadających im oczekiwanych danych wyjściowych. W tym wypadku funkcję celu zazwyczaj definiuje się jako średnią stratę (ang. *loss*) jaką ponosimy jeśli dla zadanej danej wejściowej zamiast oczekiwanej danej wyjściowej model, przy danych parametrach, udzieli innej odpowiedzi.Należy jednak pamiętać, że celem nauki nie jest jedynie zapewnienie dobrych odpowiedzi dla danych ze zbioru treningoweg. Dla zawartych tam danych wejściowych znamy już przecież oczekiwane dane wyjściowe. Wytrenowany model powinien dawać również dobre odpowiedzi dla innych danych. Właściwość taką nazywamy zdolnością do uogólniania (ang. *generalization*). Da się ją ocenić wykorzystując zbiór testowy zawierający dane, które nie były wykorzystane podczas nauki.

Własności zbioru treningowego mają ogromny wpływ na zdolność uogólniania. Po pierwsze, powinien być on reprezentatywny dla rozwiązywanego zadania. Oznacza to, ze powinny się w nim znaleźć przykłady pokrywające w miarę równomiernie całą przestrzeń możliwych wejść. Po drugie, najlepiej by było gdyby było on bardzo liczny. W zasadzie można przyjąć, że im więcej danych tym lepiej. Niestety w praktyce często spełnienie tego warunku nie jest możliwe. Typowym przykładem zbiorów gdzie te problemy występują są dane medyczne. Po trzecie wreszcie, zbiór ten nie powinien posiadać błędów.

Pierwszym etapem podczas pracy z danymi powinno być oczywiście usunięcie błędnych danych. Jest to element tak zwanego czyszczenia danych (ang. *data cleaning*), w skład którego wchodzą też uzupełnianie brakujących danych, ich standaryzacja itp. W przypadku danych obrazowych często pomaga przejrzenie dostępnych dane.

Jeśli danych jest mało można zastosować rozszerzenie zbioru o sztucznie generowane (sensowne) dane. To podejście nosi nazwę augmentacji danych (ang. *data augmentation*) i jest szczególnie popularne w przypadku systemów wykorzystujących sieci splotowe do analizy obrazów (wsparcie dla tej opcji jest dostępne w wielu bibliotekach). Jest to też jedna z metod radzenia sobie z wrażliwością sieci splotowych na obrót i skalę rozpoznawanych struktur.

W przypadku zadań gdzie elementem rozwiązania jest przypisanie klas (etykiet, kategorii) do analizowanych struktur niedobór danych może mieć jeszcze inne konsekwencje. Może się bowiem okazać, że rozkład danych pomiędzy klasami jest nierównomierny (ang. *imbalanced data*). Wpływa to znacząco na proces nauki. Aby to sobie wyobrazić wystarczy rozważyć zbiór, który posiada 99% przykładów jednej klasy i 1% przykładów innych klas. Tu znów typowym przykładem są dane medyczne gdzie zwykle mamy wielu pacjentów zdrowych i tylko kilku chorych. Trenowanie modelu tak aby osiągnął jak najbardziej optymalną średnią wartość straty może doprowadzić do sytuacji, w której będzie on wszystko klasyfikował jako obiekty tej pierwszej klasy osiągając 99% skuteczności. Istnieje wiele sposobów radzenia sobie z tym problemem:

- Można zebrać dodatkowe dane z mniej licznych klas.
- Można modyfikować samą funkcję celu nauki tak aby więcej znaczyły dane mniej licznych klas (wsparcie dla tej opcji jest dostępne w wielu bibliotekach).
- Można losowo wyrzucać przykłady z licznej klasy (ang. *under-sampling*) lub duplikować przykłady z mniej licznych klas (ang *over-sampling*).
- Można dla mniej licznych klas generować sztuczne (sensowne) dane tak jak to ma miejsce w przypadku augmentacji.

Podobny problem występuje również w przypadku innych zadań niż klasyfikacja. Na przykład w przypadku regresji może się okazać, że pewne wartości oczekiwane nie są odpowiednio licznie reprezentowane w w dostępnych danych.

Niestety okazuje się, że nawet jeśli dane są odpowiednio przygotowane, nie oznacza to, że wytrenowany model będzie na pewno posiadal zdolność uogólniania. Zależy to też od jego właściwości. Jeśli jest on zbyt elastyczny (na przykład posiada za dużo parametrów) istnieje szansa, że osiągnie on świetny wynik na danych treningowych (zapamięta te dane), ale dla innych danych nie będzie działał prawidłowo. Mówimy wówczas o tak zwanym przeuczeniu modelu (ang. *overfitting*). Ryzyko przeuczenia jest większe gdy danych treningowych jest bardzo mało (łatwiej jest je zapamiętać i nie ma z czego uogólnić). Istnieje kilka sposobów aby sobie z tym poradzić, z których dwa opisane są poniżej:

- Zazwyczaj podczas nauki modelu, która to nauka jest zwykle procesem iteracyjnym, uczy się on wpierw generalizować, a dopiero potem ulega przeuczeniu. Oznacza to, że jeśli będzie można odkryć kiedy następuje moment utraty zdolności uogólniających to będzie można przerwać naukę wcześniej (ang. *early stopping*) lub co jest równoważne wybrać rozwiązanie z iteracji wcześniejszej niż końcowa. Aby móc to odkryć trzeba umieć oceniać model na poszczególnych etapach nauki.
- Powodem przeuczenia jest zbytnia elastyczność modelu (w przypadku sieci neuronowych może ona oznaczać zbyt dużą liczbę warstw lub neuronów w warstwach). Aby ją ograniczyć można starać się eksperymentować z różnymi modelami (różnymi architekturami) tak aby wybrać tę optymalną (warto zauważyć, że model nie może być też zbyt prosty bo niczego się nie nauczy). Proces ten nosi nazwę walidacji (ang. *validation*). Wymaga on również możliwości oceniania różnych modeli. W przypadku mało licznych zbiorów danych stosuje się tak zwaną walidację krzyżową (ang. *cross-validation*).

Oba te podejścia wymagają oceny zdolności uogólniających modelu, do której wykorzystuje się tak zwany zbiór walidacyjny. Zbiór ten powinien być różny zarówno od zbioru treningowego (bo ma oceniać zdolność uogólniania), jak i od zbioru testowego (bo zbiór testowy nie może być użyty do wyboru modelu). Gdy dysponujemy licznym zbiorem danych treningowym wydziela się zbiór walidacyjny z tego zbioru (w przypadku walidacji krzyżowej stosuje się wiele takich

podziałów).

Dla porządku warto też dodać, że oprócz przeuczenia możliwe jest również wystąpienie problemu niedouczenia (ang. *underfitting*), które może mieć miejsce jeśli wybrany model będzie zbyt prosty (na przykład rozważymy zbyt prostą architekturę sieci). Wówczas model nie tylko nie będze w stanie uogólniać ale nawet nie nauczy się poprawnie odpowiadać dla danych treningowych.

Zadanie polega na przeprowadzeniu badań w ramach ustalonego z prowadzącym zadania badawczego i opisania wyników w postaci krótkiego raportu o strukturze artykułu naukowego:

- streszczenie
- opis innych prac, które rozwiązywały dane zadanie
- opis metody
- wyniki zawierające porównanie z wynikami dostępnymi w innych pracach
- dyskusja i podsumowanie
- spis literatury

Kluczowym elementem umożliwiającym wykonanie zadania są dane, które posłużą do treningu, walidacji i testowania przygotowanego rozwiązania. Z tego względu, wybierając tematykę badań, należy zadbać o ich dostępność.

Podczas rozwiązywania wszystkich tych części zadania należy wziąć pod uwagę zagadnienia omówione we wprowadzeniu do tego tego zadania (generalizacja). W takcie opracowywania artykułu należy dbać o zachowanie formalizmów matematycznych (precyzja i jednoznaczność notacji matematycznej we wzorach) oraz o zadbanie o to, aby na podstawie opisu dało się zreplikować przeprowadzone badania (opisanie niezbędnych szczegółów).

Jako wynik prac należy przesłać dwa pliki:

- `article.pdf` - artykuł opisujący przeprowadzone badania (maksymalnie 5 stron)
- `source.zip` - program użyty do przeprowadzenia eksperymentów (kod źródłowy)

Submission status

This assignment requires submission in groups. You are not a member of any group, so you cannot create a submission. Please contact your teacher to be added to a group.

Group	Not a member of any group
Attempt number	This is attempt 1.

Submission status	Nothing has been submitted for this assignment
Grading status	Not graded
Due date	Wednesday, 8 June 2022, 1:30 PM
Time remaining	7 days 3 hours
Last modified	-

« PREVIOUS ACTIVITY
Analiza artykułu

NEXT ACTIVITY
Tematyka badań »