Eksploracja Danych

Wprowadzenie

(c) Marcin Sydow

Zawartość wykładu

- wstęp
- Data Science
- cykl eksperymentu
- uczenie maszynowe
- zasoby
- podsumowanie

Zawartość kursu

Kurs eksploracji danych można podzielić na następujące części:

1. zagadnienia związane z przygotowaniem i oceną danych
2. metody wizualizacji danych
3. przykładowe modele (algorytmy) uczenia maszynowego i ich ocena
4. wybrane przykłady specjalistycznych poddziedzin (text mining, graph mining, web mining, etc.)

Rosnąca rola Data Science

- ogromna ilość danych produkowanych w sposób ciągły
- duża potencjalna wartość ukrytej wiedzy drzemiącej w tych danych
- wzrost ogólnie dostępnej i niedrogiej mocy obliczeniowej
- rosnąca dostępność stale rozwijanego, taniego (w tym darmowego) i dobrze udokumentowanego oprogramowania do data science (np. R)
- synergia środowisk akademickich (naukowcy, matematycy, statystycy, etc.) i biznesowych oraz sektora państwowego (finanse, administracja, bezpieczeństwo, etc.)
- brak wykształconych kadr/specjalistów data science

Interdyscyplinarność

- matematyka (m.in. miary, metody statystyczne, modele, algorytmy data science)
- wiedza dziedzinowa (intuicja odnośnie modeli, interpretacji danych, etc.)
- IT (programowanie, bazy danych, big data, chmury, bezpieczeństwo, zachowanie prywatności, etc.)

Metoda Empiryczna vs Data Science

Empiryczna metoda naukowa (jednym z prekursorów był Francis Bacon 1561-1626):

- zbieranie szczegółowych danych dotyczących danego problemu
- uogólnianie obserwowanych przypadków szczególnych w celu formułowania ogólnych praw/twierdzeń/reguł

Metoda Data Science: Dane $\rightarrow$ Modele $\rightarrow$ Wnioski

- (docelowo) zbieranie możliwie wszystkich danych[1] i przechowywanie ich w postaci cyfrowej
- używanie komputerów i algorytmów do automatycznego wydobywania wiedzy z tych danych

[1] Niesie to też niestety potencjalne poważne zagrożenia społeczne: totalna inwigilacja, utrata prywatności, etc.

Odkrywanie wiedzy

Można rozróżnić 3 poziomy:

1. dane (surowe dane cyfrowe)
2. informacje (interpretacja poszczególnych danych, do interpretacji niezbędna jest wiedza dziedzinowa/ekspercka)
3. wiedza (ogólne reguły)

Ważne operacje:

- abstrahowanie
- uogólnianie

Problemy w eksploracji danych

Rozwój technologii IT i ogólnie dostępnego oprogramowania (np. R) spowodował, że zbudowanie i użycie nawet skomplikowanego modelu eksploracji danych jest równoważne z napisaniem i wykonaniem zaledwie kilku instrukcji i jest powszechnie dostępne.

To jednak nie wszystko, ponieważ większość pracy z danymi oznacza:

- zdobycie wiedzy dziedzinowej dotyczącej danego problemu
- ocena przydatności danych (np. elementy statystyki)
- wstępne przygotowanie danych (np. w R, Bash, SQL, etc.)
- dobór i odpowiednia parametryzacja modeli uczenia maszynowego
- obiektywna ocena modeli
- prezentacja/komunikacja wyników

Dwa typy analiz

- Predykcja (uzupełnienie brakujących danych, również dotyczących przyszłości).
  - W celu wyuczenia modelu należy najpierw dane oczyścić (np. zidentyfikować i usunąć wartości odstające i błędne)
  - Można użyć tu modeli, które są nieprzejrzyste (ang. black-box), czyli trudne do zinterpetowania dla analityka.
  - Ocena rozwiązania może być oszacowana za pomocą pewnych automatycznych, obiektywnych procedur
- Deskrypcja (automatyczne odkrycie ogólnych wzorców ukrytych w danych).
  - Tutaj wartości odstające mogą stanowić cenne informacje.
  - Użyte modele muszą być przejrzyste (interpetowalne) przez analityka.
  - Ocena rozwiązania zależy od przydatności dla użytkownika

Cykl eksperymentu Data Science

1. Problem: zdefiniowanie rozwiązywanego problemu
2. Dane: zgromadzenie potrzebnych danych
3. Wstępna ocena danych: ocena przydatności danych do rozwiązania problemu
4. Wstępne przygotowanie danych: czyszczenie, wzbogacanie, selekcja, etc.
5. Modelowanie: tworzenie modeli eksploracji danych (konkretne algorytmy)
6. Ewaluacja: ocena i selekcja najlepszych modeli
7. Wdrożenie: komunikacja wyników i wniosków (często częściowo graficzna)

Zwykle samo modelowanie zabiera mniejszość aktywności (najwięcej na ogół zabiera przygotowanie danych).

Cele eksperymentu data science

Eksperyment data science może mieć różne cele, np.:

- wyjaśnienie możliwych przyczyn problemu
- określenie możliwych rozwiązań problemu (i porównanie ich potencjalnej jakości)
- oszacowanie ryzyka

Ocena przydatności danych

- podsumowania (statystyki pozycyjne i rozrzutu)
- rozkład częstości zmiennych (atrybutów)
- wykrycie pewnych korelacji między zmiennymi

Wstępne przygotowanie danych

- uzupełnianie brakujących wartości
- wykrywanie i poprawianie błędnych wartości
- wartości odstające (ang. outliers)
- normalizacja
- dyskretyzacja
- uogólnianie
- numerowanie stanów
- selekcja atrybutów
- redukcja wymiarów (np. PCA)

Wzbogacanie danych

- równoważenie danych
  - usuwanie niektórych przypadków klas większościowych
  - nadpróbkowanie
- transformacja zmiennych (np. liniowa lub logarytmiczna, etc.)
- dodawanie nowych zmiennych (np. sum, różnic lub iloczynów istniejących zmiennych, etc.)
- podział danych (różny w zależności od typu modelu)
  - dane treningowe
  - dane ewaluacyjne
  - dane testowe

Modelowanie

Umiejętność tworzenia uproszczonych modeli rzeczywistości i obserwacja przypadków w celu wyodrębnienia pewnych wzorców:

1. zdefiniowanie obiektów (np. użytkownik)
2. zdefiniowanie zdarzeń (np. atak hakerski, podejrzana operacja finansowa, etc.)
3. zdefiniowanie reguł (tym silniejsze im dokładniejszy model i im więcej danych)

Ocena modeli

- Kryteria oceny modeli
  - interpretowalność
  - dokładność
  - wiarygodność
  - skalowalność i wydajność
  - przydatność dla użytkownika
- ocena modeli klasyfikacyjnych
  - macierz omyłek
  - dokładność, precyzja, pełność, f-miara
  - wykresy: ROC, precyzja vs czułość, zysk
- ocena regresji (miary błędu)
- ocena grupowania (np. optymalna liczba klastrów, etc.)
- walidacja krzyżowa
- poprawa jakości modeli

Rola uczenia maszynowego

(ang. machine learning: ML)

Niektórych problemów nie można łatwo rozwiązać za pomocą dokładnych algorytmów. Dzieje się tak z rozmaitych powodów, np:

- dokładne algorytmy (rozpatrujące wszystkie możliwe niuanse i przypadki danych wejściowych) dla niektórych problemów byłyby zbyt skomplikowane, aby je stosować (a nawet opisać)
- problem braku wiedzy nt pewnych zjawisk
- problem zmienności zjawisk
- problem skalowalności

Idea i ograniczenia uczenia maszynowego

- gromadzić dane opisujące analizowane obiekty i zjawiska (dane treningowe)
- użyć tych danych do automatycznego wyuczenia odpowiednich modeli

Uczenie maszynowe nie jest jednak magicznym rozwiązaniem pozwalającym rozwiązać wszystkie problemy. Np. nie dostarczy gotowego rozwiązania jak zmniejszyć bezrobocie, ale np. odpowie z jakimi innymi czynnikami jest ono skorelowane, i od czego może zależeć jego poziom, co może pozwolić podjąć właściwe dezycje.

Typy modeli uczenia maszynowego

- klasyfikacja
- regresja
- analiza skupień (grupowanie)
- rekomendacja
- prognozowanie (szeregi czasowe)

Problemy uczenia maszynowego

Generalnie im więcej danych tym więcej informacji można wydobyć, ale też tym więcej szumu, który trzeba odfiltrować.

- przetrenowanie (ang. overfitting): zbyt sztywne dostosowanie modelu do konkretnych danych, niemożność uogólniania na nowe przypadki (spoza zbioru treningowego)
- niedouczenie (ang. underfitting): zbyt uproszczony model nie wychwytujący nawet zależności w zbiorze treningowym

Procesem uczenia maszynowego można sterować poprzez:

- odpowiednie przygotowanie danych
- dobór modeli
- parametryzację modeli

Przykładowe narzędzia

Pakiet R:

`https://cran.r-project.org/`

(polecane są dodatkowe pakiety “tidyverse”, “ggplot2”)

Środowisko graficzne RStudio:
`https://www.rstudio.com/`

Do wielu operacji bardzo wygodna jest też powłoka Linuxa (Bash) z dziesiątkami wbudowanych wspaniałych narzędzi (sort, cut, tr, etc.) i mini-języków (awk, sed, etc.)

Przykładowe repozytoria danych

- http://archive.ics.uci.edu./ml/datasets.html
- http://www.rdatamining.com/resources/data
- http://www.gapminder.org/data/
- http://www.kdnuggets.com/datasets/index.html
- http://www.kaggle.com
- http://www.openintro.org/stat

Podsumowanie

- Data Science
- cykl eksperymentu
- uczenie maszynowe
- zasoby

Przykładowe pytania/zadania/problemy

- Data Science a metoda empiryczna
- wymień możliwe zagrożenia związane z rozwojem Data Science
- problemy Data Science
- fazy cyklu eksperymentu Data Science
- przykładowe cele eksperymentu Data Science
- na czym polega uczenie maszynowe
- problemy uczenia maszynowego
- dwa typy analiz

Dziękuję za uwagę.