O interpretowalności

Mateusz Staniak

Warszawa, 29 X 2018

References

- Doshi-Velez, Finale, i Been Kim. „Towards A Rigorous Science of Interpretable Machine Learning”.
- Doshi-Velez, Finale, Mason Kortz, Ryan Budish, Chris Bavitz, Sam Gershman, David O’Brien, Stuart Schieber, James Waldo, David Weinberger, i Alexandra Wood. „Accountability of AI Under the Law: The Role of Explanation”.
- Goodman, Bryce, i Seth Flaxman. „European Union Regulations on Algorithmic Decision-Making and a «Right to Explanation»”.
- Lipton, Zachary C. „The Mythos of Model Interpretability”.

Dlaczego interpretowalność? i

1. Zaufanie
    - Czym jest zaufanie?
    - Pojedyncza metryka to za mało
    - Niemierzalne cele (np. równość)
    - Dla których przykładów model się myli?
2. Przyczynowość
    - Wyjaśnienia $\rightarrow$ hipotezy badawcze
    - counterfactual faithfulness
3. Przekładalność wyników (transferability)
    - Dryf konceptu
    - Modele wpływają na własne otoczenie
    - Ataki
4. Informatywność
    - Główne źródło informacji: odpowiedź modelu

Dlaczego interpretowalność? ii

- Alternatywne źródła: m.in. przykłady (podobne obserwacje)

5. Sprawiedliwe / etyczne decyzje (fairness)
   - Obciążenie danych
   - Regulacje prawne: *right to explain*
   - Prawo do wyjaśnienia $\rightarrow$ podważanie i poprawianie decyzji
   - zasada niedyskryminacji (uncertainty bias!)

Czym jest wyjaśnienie? i

1. Zrozumiały dla człowieka opis procesu decyzyjnego od zmiennych (wejście) do decyzji (wyjście)
    - zrozumiałość zależy od odbiorcy
    - zrozumiałość nie wymaga zagłębienia się w algorytm
2. Interpretowalność = możliwość przedstawienia w sposób zrozumiały dla ludzi
3. Nie ma jednak pełnej, formalnej definicji
    - autorzy zwykle unikają precyzyjnego określenia
    - Lundberg: wyjaśnienie = model (explainer jest modelem)
4. Fong, Ruth, i Andrea Vedaldi. „Interpretable Explanations of Black Boxes by Meaningful Perturbation”
    - explainer = meta-model (meta-predyktor), którego celem jest przewidywanie zachowania modelu predykcyjnego

Czym jest wyjaśnienie? ii

- kryterium oceny jakości explainerów: bliskość decyzji meta-modelu i modelu (+regularyzacja)

5. Wyjaśnienie powinno odpowiadać na jedno z pytań:
   - które zmienne wpłynęły na decyzję?
   - jaka zmiana i których cech zmieniłaby decyzję?
   - dlaczego dwie podobne obserwacje mają różne predykcje (lub odwrotnie)
6. Alernatywy dla wyjaśnień:
   - teoretyczne gwarancje
   - dowody statystyczne

Wbudowana wyjaśnialność modeli i

1. Symulowalność: znając dane i parametry, możemy dokładnie opisać odpowiedź modelu
2. Rozkładalność: każda składowa (np. parametr) modelu ma konkretną interpretację
    - wymaga interpretowalnośći danych wejściowych
3. Przejrzystość algorytmu: zrozumiałość samej metody uczenia
    - ludzie nie mają tej własności

Wyjaśnialność post-hoc i

1. Wizualizacja
2. Lokalne wyjaśnienia
3. Wyjaśnienia poprzez przykłady
4. Wyjaśnienia tekstowe (reinforcement learning)

Podsumowanie

1. Bogatsze metryki (uwzględniające np. fairness)
2. Przeniesienie interpretowalności do nowych dziedzin (reinforcement learning: modelowanie interakcji modelu ze światem)
3. Porównywanie interpretowalności metod wymaga
   - ustalenia definicji
   - przyjęcia mierzalnego kryterium

Kryteria do uwzględnienia

- sprawiedliwość (fairness / unbiasedness)
- prywatność
- odporność
- przyczynowość
- użyteczność

Niemierzalnośc tych kryteriów i trudność ich uwzględnienia $\rightarrow$ niekompletność sformułowania problemu $\rightarrow$ potrzebne są wyjaśnienia.

Przeszkody dla interpretowalności

1. Prywatność danych i modeli
2. Brak wiedzy technicznej odbiorców
    - w szczególności: potrzeba tworzyć interpretowalne (zrozumiałe dla ludzi) cechy na bazie danych wejściowych i pośrednich reprezentacji w algorytmach
3. Złożoność algorytmów (black box)
4. Dodatkowo, wyjaśnienia kosztują i mogą szkodzić systemom (utrata zaufania, błędne wyjaśnienia, ludzka skłonność do naginania faktów przy wyjaśnieniach post-hoc)

Wg Doshi-Velez i in. idealną odpowiedzią są osobne systemy generujące wyjaśnienia $\rightarrow$ istotność metod model-agnostic. Lokalność i *counterfactual faithfulness* (AKA what-if) nie wymagają dostępu do tajnych elementów systemu.

Jak oceniać wyjaśnienia? i

Wyjaśnienia można oceniać na bazie eksperymentów, które kwalifikują się do trzech kategorii.

1. Eksperci i prawdziwe zadania
    - Eksperyment polega na wykonania rzeczywistego zadania z dziedziny danych ekspertów
    - Wyjaśnienia są oceniane na bazie tego, czy pomogły w redukcji liczby błędów, odkryciu nowych faktów lub zmniejszeniu dyskryminacji
2. Ludzie i uproszczone zadania
    - Taki eksperyment służy zbadaniu bardziej ogólnych własności wyjaśnien
    - Uczestnicy wykonują proste, nie rzeczywiste zadanie (np. wybór bardziej przemawiającego do nich wyjaśnienia, przewidywanie zachownia modelu)

Jak oceniać wyjaśnienia? ii

3. Automatyczna ewaluacja na sztucznych zadaniach
    - Metoda symulacyjna
    - Zwykle opiera się na zastosowaniu wyjaśnienia na modelu powszechnie uważanym za interpretowalny
    - Wymaga wyboru miary do porównan

Zakończenie

- Dziękuję za uwagę
- Zapraszam do dyskusji