INTERPRETOWALNOŚĆ I WYJAŚNIALNOŚĆ UCZENIA MASZYNOWEGO

Dr Robert Małysz

WYKŁAD 1

Agenda

1. Program oraz regulamin przedmiotu
2. Definicja oraz własności interpretowalności
3. Modele interpretowalne
4. Przykłady interpretacji: RuleFit, liniowa regresja, regresja logistyczna
5. Metody Post-hoc
6. Metody agnostyczne
7. Metody lokalne vs globalne

WYKŁAD 1 PROGRAM PRZEDMIOTU

1. Wprowadzenie do interpretowalności w uczeniu maszynowym

- Omówienie fundamentalnych koncepcji interpretowalności i wyjaśnialności w uczeniu maszynowym.
- Metody oceny interpretowalności modeli, w tym nowe podejścia i metryki.
- Przedstawienie scenariuszy użycia, gdzie interpretowalność jest niezbędna (np. w medycynie, finansach, prawie).

2. Interpretowalne modele uczenia maszynowego

- Analiza różnych typów modeli uczenia maszynowego pod kątem ich łatwości interpretacji.
- Case study: Implementacja modeli interpretowalnych w bankowości dla systemów scoringowych i ratingowych.

3. Zaawansowane metody interpretowalności modeli

- Techniki głębokiego wyjaśniania, w tym ulepszone wersje LIME i SHAP, oraz nowe narzędzia jak Counterfactual Explanations, Contextual Decomposition.
- Wizualizacje danych wspierające interpretowalność, takie jak zaawansowane ploty ważności cech i interaktywne wykresy zależności.

WYKŁAD 1 PROGRAM PRZEDMIOTU

4. Etyczne i społeczne aspekty interpretowalności
- Rola interpretowalności w kontekście etycznych i społecznych wyzwań AI.
- Dyskusja o odpowiedzialności i przejrzystości w projektowaniu i stosowaniu modeli AI.

5. Wprowadzenie do GPT i najnowsze modele językowe
- Przegląd ostatnich wersji modeli Generative Pre-trained
- Transformer, w tym GPT-4 i modele GPT-4o i nowsze, oraz ich wpływ na rozwój AI.
- Analiza architektury i mechanizmów nowych modeli, jak transformer XL i reformer, które przyczyniają się do lepszego zrozumienia i efektywności treningu.

6. Regulacje dotyczące intepretowalności AI, w tym AI Act (EU 2024/1495)
- Omówienie nowych regulacji i wytycznych etycznych związanych z AI.
- Regulacje m.in. z USA, Wielka Brytania, Kanada, Japonia
- AI Act (EU 2024/1495) – szczegółowe wymogi intepretowalności dla systemów wysokiego ryzyka

WYKŁAD 1 PROGRAM PRZEDMIOTU

7. Zastosowania, wyzwania i przyszłość modeli GPT

- Różnorodność zastosowań modeli GPT w nowych dziedzinach, jak tworzenie treści, programowanie i interakcje człowiek-AI.
- Przedstawienie wyzwań związanych z odpowiedzialnym stosowaniem modeli GPT, w tym zagadnień takich jak bias, prywatność danych i bezpieczeństwo.
- Perspektywy przyszłego rozwoju modeli językowych i ich wpływ na społeczeństwo.

WYKŁAD 1 PROGRAM PRZEDMIOTU

Laboratorium

Środowisko i narzędzia: Python (Scikit-learn, skops, PyTorch, TensorFlow, Keras, Fast.ai, truLens-eval – OpenAI 2024) oraz dostęp do API GPT.

WYKŁAD 1 LITERATURA

1. Christoph Molnar, Interpretable Machine Learning: A Guide for Making Black Box Models Explainable (3rd ed., 2025). https://christophm.github.io/interpretable-ml-book/
2. Scott M. Lundberg & Su-In Lee, “A Unified Approach to Interpreting Model Predictions” (NeurIPS 2017)
3. Riccardo Guidotti et al., “A Survey Of Methods For Explaining Black Box Models” (ACM Computing Surveys 51(5):1–42, 2018)
4. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). "Attention Is All You Need". Advances in Neural Information Processing Systems. Attention Is All You Need. .

WYKŁAD 1 ZALICZENIE

- W oparciu o wykład oraz laboratorium studenci w zespołach tworzą dwa programy/algorytmu w python, które podlegają ocenie. Ocena algorytmu stanowi 70% całkowitej oceny.
- Ocena punktowa na podstawie testu z tematyki wykładu stanowi 30% całej oceny.

WYKŁAD 1

DEFINICJA INTERPRETOWALNOŚCI

“Ability to explain or to present a model in understandable terms to humans “

„Zdolność do wyjaśniania lub prezentowania modelu w sposób zrozumiały dla ludzi"

(Doshi-Velez 2017)

WYKŁAD 1

DEFINICJA INTERPRETOWALNOŚCI

Definicja interpretowalności jako "Zdolność do wyjaśniania lub prezentowania modelu w sposób zrozumiały dla ludzi" (Doshi-Velez 2017) jest szeroko akceptowana i uważana za istotną w kontekście modeli uczenia maszynowego i sztucznej inteligencji. Oto kilka punktów na temat tej definicji:

Zalety:

- Uniwersalność: Definicja jest na tyle szeroka, że może być stosowana w różnych kontekstach i dziedzinach.
- Skupienie na Ludziach: Podkreśla ważność zrozumienia modelu przez ludzi, co jest kluczowe dla zaufania i akceptacji modelu.
- Praktyczność: W kontekście biznesowym i naukowym, zdolność do wyjaśniania modelu jest niezbędna, aby wyniki analiz były użyteczne i mogły być stosowane w podejmowaniu decyzji.

Wyzwania:

- Subiektywność: „w sposób zrozumiały" mogą być różnie interpretowane przez różne osoby, w zależności od ich tła i doświadczenia.
- Kompleksowość: Niektóre modele, zwłaszcza te bardziej zaawansowane, mogą być trudne do przedstawienia w sposób zrozumiały, bez utraty istotnych szczegółów.
- Granice Interpretowalności: Nie zawsze jest jasne, jak głęboko model powinien być wyjaśniany, aby był uważany za "zrozumiały".

Definicja jest użyteczna i stanowi solidną podstawę, ważne jest, aby pamiętać o tych wyzwaniach i dostosowywać podejście do interpretowalności w zależności od konkretnego przypadku użycia i publiczności. Różne sytuacje mogą wymagać różnych poziomów i form interpretowalności, a co za tym idzie, różnych technik i narzędzi do jej osiągnięcia.

WYKŁAD 1

DEFINICJA INTERPRETOWALNOŚCI

1. Zrozumiałość vs. Zaufanie

"Interpretowalność to stopień, w jakim człowiek może zrozumieć model, a zaufanie to stopień, w jakim człowiek myśli, że model działa poprawnie" (Kim et al., 2016).

2. Zrozumiałość i Przejrzystość

"Interpretowalność to miara, jak łatwo ludzie mogą zrozumieć przyczyny i skutki w trakcie działania modelu, a przejrzystość to miara, jak łatwo ludzie mogą zrozumieć, jak działa cały model" (Doshi-Velez i Kim, 2017).

3. Zrozumiałość dla Ekspertów

"Model jest interpretowalny, jeśli praktykujący ekspert może zrozumieć przyczynowość modelu" (Rudin, 2018).

4. Zrozumiałość przez Ludzi

"Model jest interpretowalny, jeśli jego działania i decyzje mogą być zrozumiane przez ludzi" (Lipton, 2016).

5. Zrozumiałość i Użyteczność

"Interpretowalność to stopień, w jakim człowiek może zrozumieć i używać modelu oraz jak łatwo można go eksplorować" (Hohman et al., 2018).

6. Zrozumiałość i Zgodność

"Interpretowalność to miara, jak dobrze model uczenia maszynowego może być zrozumiany w kontekście konkretnego zadania" (Carvalho et al., 2019).

7. Zrozumiałość i Wyjaśnialność

"Interpretowalność to zdolność do opisania wewnętrznych mechanizmów modelu lub przyczyn i skutków przewidywań modelu" (Gilpin et al., 2018).

WYKŁAD 1

WŁASNOŚCI INTERPRETOWALNOŚCI

- Wierność - Jak dostarczyć wyjaśnień, które dokładnie reprezentują prawdziwe uzasadnienie za decyzją końcową modelu.
- Wiarygodność – Czy wyjaśnienie jest poprawne lub czy możemy uwierzyć, że jest prawdziwe, biorąc pod uwagę naszą aktualną wiedzę na temat problemu?
- Zrozumiałość – Czy mogę to wyrazić w terminach, które końcowy użytkownik bez dogłębnej wiedzy o systemie może zrozumieć?
- Stabilność – Czy podobne instancje mają podobne interpretacje?

WYKŁAD 1

Interpretowalność i wyjaśnialność to dwa kluczowe pojęcia w dziedzinie uczenia maszynowego, zwłaszcza gdy chodzi o modele "czarnej skrzynki", które są trudne do zrozumienia dla ludzi.

Interpretowalność odnosi się do zdolności modelu do przedstawienia swoich decyzji w sposób zrozumiały dla ludzi. Jest to właściwość modelu, która pozwala na zrozumienie, jakie cechy danych są ważne dla modelu podczas podejmowania decyzji i jak te cechy wpływają na wynik. Interpretowalność jest kluczowa w przypadkach, gdy potrzebujemy zrozumieć, jak model "myśli" i jakie mechanizmy kierują jego decyzjami.

Wyjaśnialność odnosi się do zdolności modelu do dostarczenia jasnych, konkretnych i zrozumiałych powodów dla swoich decyzji. Gdy mówimy o wyjaśnialności, chcemy, aby model nie tylko był zrozumiały, ale także dostarczał konkretne wyjaśnienia dla indywidualnych prognoz. Wyjaśnialność jest kluczowa w sytuacjach, gdy musimy uzasadnić decyzje modelu przed innymi ludźmi, na przykład w medycynie, prawie czy finansach.

W skrócie, interpretowalność koncentruje się na ogólnym zrozumieniu działania modelu, podczas, gdy wyjaśnialność skupia się na dostarczaniu konkretnych powodów dla indywidualnych decyzji modelu.

WYKŁAD 1

- Interpretowalność odnosi się do zdolności modelu do przedstawienia swoich decyzji w sposób zrozumiały dla ludzi.

- Chodzi o to, abyśmy mogli zrozumieć, jak model "myśli" i jakie cechy danych są dla niego ważne podczas podejmowania decyzji.

- Z drugiej strony, wyjaśnialność odnosi się do zdolności modelu do dostarczenia jasnych i zrozumiałych powodów dla swoich decyzji. Wyjaśnialność jest często w praktyce traktowana jako narzędzie wizualizacyjne, które pozwala na badanie, jak czarna skrzynka działa wewnętrznie.

WYKŁAD 1

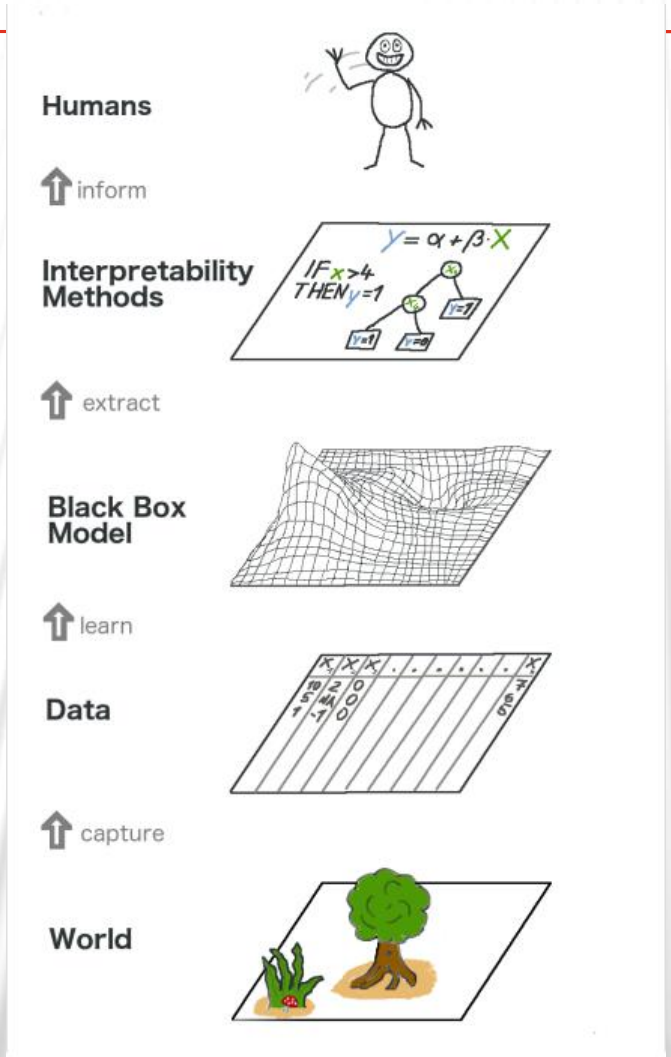
Interpretowalność w uczeniu maszynowym stała się kluczowym zagadnieniem, zwłaszcza w kontekście zastosowań w dziedzinach takich jak:

- medycyna,
- finanse,
- prawo,

gdzie decyzje podejmowane przez modele mają bezpośredni wpływ na ludzkie życie. W tych dziedzinach nie wystarczy, że model jest dokładny - musi być również zrozumiały dla ludzi.

WYKŁAD 1

Interpretowalność modeli jest kluczowym elementem w dziedzinie nauki o danych i uczenia maszynowego, zwłaszcza w kontekście zastosowań w rzeczywistym świecie, gdzie zrozumienie, dlaczego model podejmuje pewne decyzje, jest często wymagane ze względów prawnych lub etycznych.


Humans
inform
Interpretability Methods
$Y = \alpha + \beta \cdot X$
IF x>4
THEN y=1
extract
Black Box Model
learn
Data
capture
World

WYKŁAD 1

DLACZEGO INTERPRETOWALNOŚĆ JEST WAŻNA?

Podstawowe problemy dla ML (Machine Learning - Uczenia Maszynowego)

- Ludzie nie ufają modelom
- Nie wiemy, co dzieje się w skrajnych przypadkach
- Błędy mogą być bardzo kosztowne
- Czy modele popełniają podobne błędy co ludzie?
- Jak zmieniać modele, gdy coś „idzie nie tak”?

Prawo i Etyka

GDPR, Dyrektywa UE – algorytmy powinny wyjaśniać swoje wyniki.

Modele typu "czarne skrzynki" mogą być nielegalne bez interpretowalności.

WYKŁAD 1

Metody Wbudowane (Intrinsic)

- Metody te są częścią samego modelu i zazwyczaj są stosowane w trakcie procesu uczenia.
- Modele interpretowalne to takie, które umożliwiają łatwe zrozumienie relacji między zmiennymi wejściowymi i prognozowanymi wyjściami.

WYKŁAD 1

Kilka przykładów modeli, które są często uważane za interpretowalne:

1. Regresja Liniowa

Prosta Regresja Liniowa: Model z jedną zmienną niezależną.

Wielokrotna Regresja Liniowa: Model z wieloma zmiennymi niezależnymi.

2. Drzewa Decyzyjne

Drzewo Decyzyjne: Proste drzewa są zazwyczaj łatwe do interpretacji i wizualizacji.

Model drzewa klasyfikacyjnego i regresyjnego (CART): Podobnie jak drzewa decyzyjne, ale mogą być stosowane do problemów regresji.

3. Model Liniowy Generalizowany (GLM)

Regresja Logistyczna: Specjalny przypadek GLM stosowany do klasyfikacji binarnej.

Poisson Regression: Używane do modelowania liczby zdarzeń, które mają miejsce w określonym czasie.

4. K-Najbliższych Sąsiadów (K-NN)

Choć model K-NN jest prosty, jego interpretowalność może zależeć od kontekstu i wymiarowości danych.

5. Naiwny Klasyfikator Bayesowski

Prosty i interpretowalny model, często stosowany w klasyfikacji tekstów.

WYKŁAD 1

6. Reguły Decyzyjne

Reguły Asocjacyjne: Takie jak algorytmy Apriori lub FP-growth.

RuleFit: Model, który łączy reguły z modelami liniowymi.

7. Model Addytywny Generalizowany (GAM)

Pozwala na modelowanie nieliniowych zależności przy jednoczesnym zachowaniu interpretowalności.

8. LASSO (Least Absolute Shrinkage and Selection Operator)

Choć jest to technika regularyzacji stosowana w regresji, LASSO może być również interpretowalne, ponieważ prowadzi do modeli, które mają mniej niezerowych współczynników.

9. Analiza Głównych Składowych (PCA)

Choć jest to technika redukcji wymiarowości, a nie model predykcyjny, PCA jest często używane w celu zrozumienia struktury zmiennych w zestawie danych.

10. Analiza Dyskryminacyjna

Liniowa Analiza Dyskryminacyjna (LDA): Technika używana do znajdowania liniowych kombinacji cech, które najlepiej rozróżniają dwie lub więcej klas obiektów.

WYKŁAD 1

MODELE INTERETOWALNE

Model	Liniowość	Monotoniczność	Interakcja	Zadanie
Regresja liniowa	Tak	Tak	Nie	regr
Regresja logistyczna	Nie	Tak	Nie	class
Drzewa decyzyjne	Nie	Czasami	Tak	class, regr
RuleFit	Tak	Nie	Tak	class, regr
Naiwny Bayesowski	Nie	Tak	Nie	class
K-Najbliższych Sąsiadów (K-NN)	Nie	Nie	Nie	class, regr

WYKŁAD 1

MODELE BEZ INTERACJI

Modele bez interakcji to takie, które nie uwzględniają efektów współdziałania między różnymi zmiennymi wejściowymi. Oto kilka przykładów:

1. Prosta Regresja Liniowa

Model prostej regresji liniowej zakłada liniową zależność między zmienną zależną a jedną zmienną niezależną, bez żadnych interakcji. Model ma postać: $$Y = \beta_0 + \beta_1 X_1 + \epsilon$$

2. Wielokrotna Regresja Liniowa bez Interakcji

Wielokrotna regresja liniowa może również być modelem bez interakcji, jeśli uwzględnia tylko główne efekty zmiennych niezależnych, bez ich kombinacji. Model ma postać: $$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \epsilon$$

3. Modele Logistyczne bez Interakcji

Podobnie jak w przypadku regresji liniowej, modele logistyczne mogą być bez interakcji, jeśli uwzględniają tylko główne efekty zmiennych. Model ma postać: $$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$$

4. Niektóre Modele Drzew Decyzyjnych

Choć drzewa decyzyjne są naturalnie zdolne do modelowania interakcji, pewne drzewa, zwłaszcza te prostsze, mogą nie wykazywać wyraźnych interakcji między zmiennymi, jeśli dane tego nie wymagają.

5. Niektóre Modele Naiwnego Klasyfikatora Bayesowskiego

Naiwne klasyfikatory Bayesowskie zakładają, że wszystkie cechy są warunkowo niezależne, co implikuje brak interakcji między zmiennymi.

WYKŁAD 1

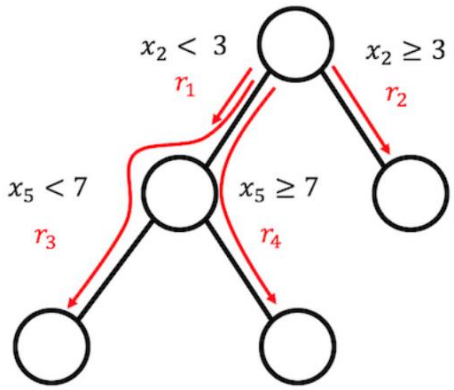
MODELE BEZ INTERACJI

- Modele bez interakcji mogą być bardziej interpretowalne, ale mogą również być mniej elastyczne i dokładne, jeśli w danych występują istotne interakcje.
- Wybór między modelem z interakcjami a modelem bez interakcji często zależy od celu analizy, dostępności danych i wymagań dotyczących interpretowalności.
- Nawet jeśli model jest początkowo zbudowany bez interakcji, warto przeprowadzić analizę, aby sprawdzić, czy dodanie czynników interakcyjnych może poprawić wydajność modelu.

WYKŁAD 1

RULEFIT

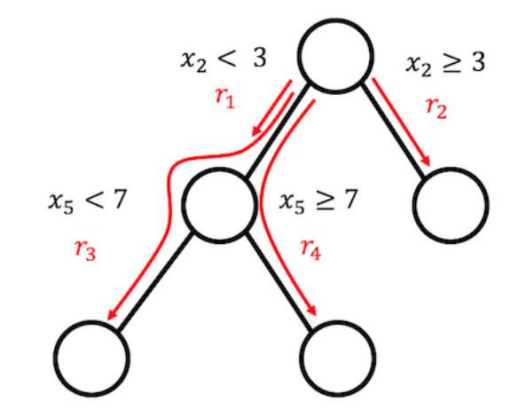
"RuleFit" to technika modelowania predykcyjnego, która łączy podejście oparte na drzewach decyzyjnych z modelem liniowym, starając się uzyskać kompromis między interpretowalnością a precyzją predykcji. Metoda ta została zaproponowana przez Friedmana i Popescu (2008-"Predictive learning via rule ensembles") i jest często stosowana w analizie danych, gdzie zarówno dokładność predykcji, jak i zdolność do interpretacji modelu są ważne.



WYKŁAD 1

RULEFIT

- RuleFit dopasowuje rzadki model liniowy z oryginalnymi cechami oraz zestawem nowych cech, które są regułami decyzyjnymi.
- Nowe cechy „przechwytują” interakcje między oryginalnymi cechami.
- Nowe cechy są generowane automatycznie z drzew decyzyjnych.



4 reguły mogą być wygenerowane z drzewa mającego 3 węzły końcowe.

WYKŁAD 1

GŁÓWNE KONCEPCJE RULEFIT

1. **Generowanie Reguł**:

   RuleFit zaczyna od wygenerowania zestawu reguł, używając drzew decyzyjnych. Drzewa są trenowane na danych, a następnie każda ścieżka od korzenia do liścia jest przekształcana w regułę (np. "jeśli $X1>5$ i $X2\leq10$, to...").

2. **Dopasowanie Modelu Liniowego**:

   Następnie, dla każdej obserwacji, RuleFit tworzy nowe zmienne binarne, które wskazują, czy dana obserwacja spełnia każdą z wygenerowanych reguł.

   Te nowe zmienne binarne są używane jako zmienne objaśniające w modelu liniowym (często z regularyzacją, taką jak LASSO), aby przewidzieć zmienną zależną.

3. **Interpretowalność**:

   Współczynniki modelu liniowego wskazują na ważność każdej reguły w kontekście predykcji.

   Reguły i ich współczynniki mogą być przedstawione w czytelnej formie, umożliwiając interpretację modelu.

WYKŁAD 1

RULEFIT

Zalety RuleFit:

- **Interpretowalność**: Dzięki reprezentacji modelu jako zestawu reguł, RuleFit jest stosunkowo łatwy do zinterpretowania, nawet dla osób nieznających technicznych aspektów modelowania.
- **Precyzja**: RuleFit jest w stanie uchwycić nieliniowe zależności i interakcje między zmiennymi, co może prowadzić do modeli o wysokiej precyzji.
- **Automatyzacja**: Proces generowania reguł i dopasowywania modelu liniowego jest zautomatyzowany, co sprawia, że RuleFit jest stosunkowo łatwy w użyciu.

Ograniczenia RuleFit:

- **Złożoność**: Mimo że RuleFit jest interpretowalny, może generować wiele reguł, co czasem utrudnia pełne zrozumienie modelu.
- **Wydajność**: Generowanie reguł i dopasowywanie modelu liniowego może być czasochłonne dla dużych zestawów danych.
- **Nadmierna Dopasowanie**: Jeśli RuleFit generuje zbyt wiele skomplikowanych reguł, model może stać się nadmiernie dopasowany do danych treningowych.

RuleFit może być szczególnie przydatny w scenariuszach, gdzie interpretowalność modelu jest ważna, ale jednocześnie chcemy uwzględnić nieliniowe zależności i interakcje między zmiennymi.

WYKŁAD 1

LINIOWA REGRESJA - INTERPRETACJA

Interpretacja Cechy Numerycznej
Zwiększenie wartości cechy x_k o jednostkę zwiększa prognozę dla y o β_k jednostek, gdy wszystkie inne wartości cech pozostają niezmienione.

Interpretacja Cechy Kategorycznej
Zmiana cechy x_k z kategorii odniesienia na inną kategorię zwiększa prognozę dla y o β_k, gdy wszystkie inne cechy pozostają niezmienione.

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \epsilon$$

$$\hat{\boldsymbol{\beta}} = \arg\min_{\beta_0,\ldots,\beta_p} \sum_{i=1}^{n} \left(y^{(i)} - \left(\beta_0 + \sum_{j=1}^{p} \beta_j x_j^{(i)} \right) \right)^2$$

WYKŁAD 1

LINIOWA REGRESJA - INTERPRETACJA

R-kwadrat - jak duża część całkowitej wariancji zmiennej zależnej jest wyjaśniana przez model.

$$R^2 = 1 - SSE/SST$$

SSE (Suma Kwadratów Błędów) to kwadratowa suma błędów. Określa, ile wariancji pozostaje po dopasowaniu modelu liniowego, co jest mierzone przez kwadraty różnic między przewidywanymi a rzeczywistymi wartościami zmiennej zależnej.

$$SSE = \sum^{n}(y^{(i)} - \hat{y}^{(i)})^2$$

SST (Suma Kwadratów Całkowitych) to kwadratowa suma wariancji danych. Jest to całkowita wariancja zmiennej zależnej względem średniej.

$$SST = \sum_{i=1}^{n}(y^{(i)} - \bar{y})^2$$

Ważność Cechy/**Feature Importance** (wartość bezwzględna statystyki t)
Statystyka t to oszacowana waga przeskalowana za pomocą jej błędu standardowego.

$$t_{\hat{\beta}_j} = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}$$

WYKŁAD 1

REGRESJA LOGISTYCZNA - INTERPRETACJA

$$\hat{y}^{(i)} = \beta_0 + \beta_1 x_1^{(i)} + \ldots + \beta_p x_p^{(i)}$$

$$P(y^{(i)} = 1) = \frac{1}{1 + exp(-(\beta_0 + \beta_1 x_1^{(i)} + \ldots + \beta_p x_p^{(i)}))}$$

$$ln\left(\frac{P(y=1)}{1-P(y=1)}\right) = log\left(\frac{P(y=1)}{P(y=0)}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

$$\frac{P(y=1)}{1-P(y=1)} = odds = exp\left(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p\right)$$

$$\frac{odds_{x_j+1}}{odds} = exp\left(\beta_j(x_j+1) - \beta_j x_j\right) = exp\left(\beta_j\right)$$

WYKŁAD 1
REGRESJA LOGISTYCZNA - INTERPRETACJA

Interpretacja Cechy Numerycznej
Zwiększenie wartości cechy x_k o jednostkę zmienia szacowane szanse (odds) o współczynnik $\exp(\beta_k)$, gdy wszystkie inne wartości cech pozostają stałe.

Interpretacja Cechy Kategorycznej Binarniej
Zmiana cechy x_k z kategorii odniesienia na inną kategorię zmienia szacowane szanse o współczynnik $\exp(\beta_k)$, gdy wszystkie inne cechy pozostają stałe.

Interpretacja Cechy Kategorycznej z Więcej niż Dwiema Kategoriami
Jednym z rozwiązań radzenia sobie z wieloma kategoriami jest kodowanie "one-hot", co oznacza, że każda kategoria ma własną kolumnę.

Interpretacja $\boldsymbol{\beta_0}$
Gdy wszystkie cechy numeryczne są równe zero, a cechy kategoryczne przyjmują wartości kategorii odniesienia, szacowane szanse wynoszą $\exp(\beta_0)$

WYKŁAD 1

Metody Post-hoc (1/2)

Te metody są stosowane po wytrenowaniu modelu, aby wyjaśnić jego predykcje.

Przykłady:

- LIME (Local Interpretable Model-agnostic Explanations): Tworzy lokalne, liniowe przybliżenia modelu wokół punktu predykcji, aby wyjaśnić, dlaczego model podjął konkretną decyzję.
- SHAP (SHapley Additive exPlanations): Opiera się na teorii gier i rozkłada predykcję modelu na wartości Shapleya, które wskazują, jak każda cecha przyczynia się do predykcji.
- Permutacyjna ważność cech: Mierzy wpływ każdej cechy na wydajność modelu poprzez obserwowanie, jak zmienia się wydajność modelu, gdy wartości danej cechy są losowo permutowane.
- Ceteris Paribus / Profile Ceteris Paribus: Analizuje, jak zmienia się predykcja modelu, gdy wartość jednej zmiennej jest zmieniana, a wszystkie inne są utrzymywane stałe.

WYKŁAD 1

Metody Post-hoc (2/2)

Przykłady:

- Analiza czułości: Badanie, jak różne wartości wejściowe wpływają na wyjście modelu.
- Przeciwdziałać wyjaśnieniom: Zmiana wartości cech, aby zobaczyć, jakie zmiany muszą zajść, aby zmienić decyzję modelu.
- Drzewa decyzyjne surrogatowe: Trenowanie drzewa decyzyjnego na predykcjach modelu, aby uzyskać przybliżone wyjaśnienie decyzji modelu.
- Mapy ciepła i konturowe: Wizualizacja, jak różne kombinacje wartości cech wpływają na predykcje modelu.
- Analiza reszt: Badanie różnic między prognozowanymi a rzeczywistymi wartościami.

WYKŁAD 1

W praktyce istnieje wiele metod oceny interpretowalności modeli, w tym narzędzia wizualizacyjne, które pozwalają na badanie, jak model działa wewnętrznie.

Przykładem jest "**Prediction difference analysis**", która pozwala na wizualizację decyzji podejmowanych przez głębokie sieci neuronowe. Metoda ta koncentruje się na identyfikacji obszarów na obrazie wejściowym, które są kluczowe dla modelu podczas podejmowania konkretnej decyzji. PDA jest szczególnie przydatne w kontekście modeli klasyfikacji obrazów, gdzie interpretowalność modelu jest kluczowa dla zrozumienia, dlaczego model dokonuje pewnych predykcji.

PDA jest często używane w analizie medycznej (np. identyfikacja obszarów na obrazach medycznych, które są kluczowe dla diagnozy) oraz w innych dziedzinach, gdzie wizualizacja decyzji modelu jest ważna.

Inne metody, takie jak "rule extraction", pozwalają na wyodrębnienie symbolicznych reguł z wyszkolonych modeli, takich jak sieci neuronowe, co może pomóc w zrozumieniu ich działania.

WYKŁAD 1

METODY AGNOSTYCZNE

Ribeiro, Singh, Guestrin 2016

Elastyczność Modelu: Metoda interpretacji może pracować z dowolnym modelem uczenia maszynowego, takim jak lasy losowe czy głębokie sieci neuronowe.

Elastyczność Wyjaśnień: Nie jesteś ograniczony do pewnej formy wyjaśnienia. W niektórych przypadkach przydatny może być liniowy wzór, w innych przypadkach grafika z ważnościami cech.

Elastyczność Reprezentacji: System wyjaśnień powinien być w stanie używać innej reprezentacji cech niż model, który jest wyjaśniany. Dla klasyfikatora tekstu, który używa abstrakcyjnych wektorów osadzeń słów, preferencyjne może być używanie obecności poszczególnych słów dla wyjaśnienia.

WYKŁAD 1

METODY LOKALNE VS GLOBALNE

Metody lokalne i globalne to różne podejścia do interpretacji modeli uczenia maszynowego, które koncentrują się na różnych aspektach prognoz modelu:

Metody Globalne

1. **Ogólny Wgląd**: Metody globalne dostarczają ogólnego wglądu w działanie modelu na całym zbiorze danych.
2. **Całość Modelu**: Skupiają się na interpretacji całego modelu, nie tylko pojedynczych prognoz.
3. **Ważność Cech**: Często oceniają ważność cech na poziomie globalnym, pokazując, które zmienne są generalnie najważniejsze dla modelu.

WYKŁAD 1

METODY LOKALNE VS GLOBALNE

Metody Lokalne

1. **Indywidualne Prognozy:** Metody lokalne koncentrują się na wyjaśnianiu indywidualnych prognoz, nie całego modelu.

2. **Szczegółowe Wyjaśnienia**: Dostarczają szczegółowych wyjaśnień dotyczących decyzji podjętej dla konkretnej obserwacji.

3. **Wrażliwość na Zmienne**: Pokazują, jak zmiany w wartościach cech wpływają na prognozę dla konkretnej obserwacji.

WYKŁAD 1

METODY AGNOSTYCZNE

Globalne Metody Niezależne od Modelu

1. Wykres Częściowej Zależności (PDP - Partial Dependence Plot)
2. Wykres Skumulowanych Lokalnych Efektów (ALE - Accumulated Local Effects)

   PDP i ALE są technikami wizualizacji, które pomagają zrozumieć, jak poszczególne cechy wpływają na prognozę modelu na poziomie globalnym.
3. Interakcja Cech
4. Ważność Cechy Permutacyjnej

   Interakcja Cech i Ważność Cechy Permutacyjnej są technikami oceny ważności cech i ich wpływu na model.
5. Globalny Model Zastępczy (Surrogate) - jest modelem, który stara się naśladować prognozy bardziej skomplikowanego modelu w sposób, który jest łatwiejszy do zrozumienia.

Lokalne Metody Niezależne od Modelu

1. Lokalne Modele Zastępcze (LIME - Local Interpretable Model-agnostic Explanations)
2. SHAP (SHapley Additive exPlanations)

   LIME i SHAP są technikami, które starają się wyjaśnić prognozy pojedynczych obserwacji przez aproksymację lokalnego modelu lub rozkładanie prognozy na wkład indywidualnych cech.
3. Wartość Shapleya - jest koncepcją z teorii gier, która została zaadaptowana do wyjaśniania prognoz modeli uczenia maszynowego.

WYKŁAD 1

Interpretowalność vs. Dokładność: Często istnieje kompromis między interpretowalnością a dokładnością modelu. Prostsze modele, takie jak drzewa decyzyjne, są łatwiejsze do interpretacji, ale mogą nie osiągać tak wysokiej dokładności jak bardziej skomplikowane modele, takie jak głębokie sieci neuronowe.

Etyka i Prawo: Interpretowalność jest ważna nie tylko dla zrozumienia modelu, ale także dla spełnienia wymogów prawnych (np. rozporządzenie GDPR w Unii Europejskiej) i etycznych.

Zastosowanie w Praktyce: Wybór metody interpretowalności może zależeć od specyfiki problemu, rodzaju modelu i danych oraz od wymagań interesariuszy projektu.

WYKŁAD 1

1. Medycyna i Opieka Zdrowotna

Diagnostyka: Lekarze muszą rozumieć, dlaczego model AI zaleca określoną diagnozę lub plan leczenia, aby móc podejmować świadome decyzje i ewentualnie wyjaśnić je pacjentowi.

Badania Kliniczne: Zrozumienie, jakie cechy wpływają na prognozy modelu, może pomóc naukowcom odkrywać nowe zależności i mechanizmy w danych medycznych.

2. Finanse

Kredyty: Banki i inne instytucje finansowe muszą być w stanie wyjaśnić, dlaczego wniosek kredytowy został zaakceptowany lub odrzucony, aby spełniać wymogi regulacyjne i zapewniać sprawiedliwość (brak dyskryminacji).

Algorytmy Tradingowe/Handlowe: Inwestorzy i regulatorzy mogą wymagać zrozumienia, jak algorytmy handlowe podejmują decyzje, aby zapewnić uczciwość i stabilność rynku.

3. Prawo

Systemy Wykrywania Przestępczości: Zrozumienie, dlaczego system AI flaguje pewne działania jako podejrzane, jest kluczowe dla dalszych śledztw i unikania fałszywych alarmów.

Prognozowanie Ryzyka Przestępczości: Sędziowie i pracownicy systemu sądowego mogą potrzebować wyjaśnień dotyczących ocen ryzyka przestępczości, aby podejmować sprawiedliwe decyzje.

WYKŁAD 1

4. Automatyka Procesów Przemysłowych

Bezpieczeństwo: Inżynierowie muszą rozumieć, dlaczego system AI podejmuje pewne decyzje, zwłaszcza w kontekście bezpieczeństwa i awarii, aby móc interweniować i optymalizować procesy.

Optymalizacja Produkcji: Zrozumienie, jakie zmienne wpływają na prognozy modelu, może pomóc w identyfikacji obszarów do poprawy i optymalizacji.

5. Rekrutacja i Zarządzanie Zasobami Ludzkimi

Selekcja Kandydatów: HR musi być w stanie wyjaśnić, dlaczego pewni kandydaci są preferowani przez system AI, aby unikać dyskryminacji i zapewniać sprawiedliwość w procesie rekrutacji.

Ocena Wydajności: Pracownicy mogą wymagać wyjaśnień dotyczących ocen wydajności generowanych przez modele AI, aby zrozumieć, jak mogą się poprawić.

6. Edukacja

Systemy Rekomendacji: Edukatorzy i studenci mogą chcieć wiedzieć, dlaczego pewne materiały lub kursy są rekomendowane, aby dostosować ścieżki edukacyjne.

Ocena Automatyczna: Nauczyciele i uczniowie mogą potrzebować wyjaśnień dotyczących automatycznie przyznawanych ocen, aby zapewnić sprawiedliwość i jakość edukacji.

WYKŁAD 1

7. Marketing i Reklama

Personalizacja: Marketerzy mogą chcieć zrozumieć, dlaczego pewne produkty lub reklamy są rekomendowane konkretnym klientom, aby optymalizować kampanie.

Segmentacja Klientów: Zrozumienie, jakie cechy wpływają na segmentację klientów, może pomóc w tworzeniu bardziej skutecznych strategii marketingowych.