

درک چرایی تصمیم‌گیری سیستم‌های هوش مصنوعی

سید محمد امین دادگر
دانشجوی هوش مصنوعی ارشد دانشگاه اصفهان

معرفی

سید محمد امین دادگر
دانشجوی مهندسی کامپیوتر ارشد دانشگاه اصفهان



فعال در حوزه‌های:

تفسیر مدل‌های هوش مصنوعی

تحلیل داده

نویسنده مقالات medium

۱. وضعیت مدل‌های هوش مصنوعی

مقایسه هوش مصنوعی از گذشته تا حال و آینده

سه دوره هوش مصنوعی [۱]

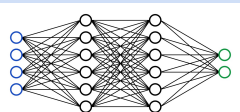
دوره سوم
زمان حال - اواسط ۲۰۳۰

- یادگیری: یادگیری مفهومی^۳
- تصمیم‌گیری: با جمع‌آوری و مقایسه مفهومی و ارائه دلیل تصمیم

¹ Hand-Crafted Learning

دوره دوم
۲۰۰۰-زمان حال



- یادگیری: با مدل‌های ریاضی^۲ و به کمک داده
- تصمیم‌گیری: به کمک مدل‌های ریاضیاتی



² Statistical Learning

دوره اول
۱۹۷۰-۲۰۰۰

- یادگیری: فرموله‌سازی اطلاعات به صورت دستی^۱
- تصمیم‌گیری: به کمک قوانین به دست آمده

³ Concept Learning

سه دوره هوش مصنوعی (ادامه)

بیان تاریخچه

راه حل
مشکلات

مثال راه حل

هوش مصنوعی
توضیح پذیر

کارایی بالای مدل های دوره دوم:

خودروهای خودران

سیستم پردازش زبان طبیعی Google

سیستم تشخیص چهره

دلیل نیاز به دوره سوم:

عدم شفافیت در تصمیم گیری

نیاز به داشتن دلیل در سیستم های حیاتی

5/18

مشاهده برخی تصمیمات دوره دوم هوش مصنوعی

بیان تاریخچه

راه حل
مشکلات

مثال راه حل

هوش مصنوعی
توضیح پذیر



6/18

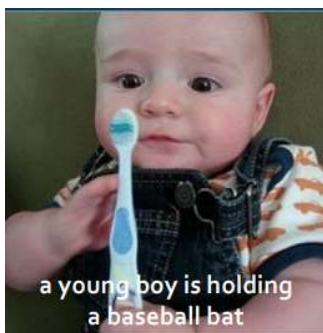
مشاهده برخی تصمیمات دوره دوم هوش مصنوعی

بیان تاریخچه

راه حل
مشکلات

مثال راه حل

هوش مصنوعی
توضیح پذیر



(ب)



(الف)

7/18

۲. راهکار ابتدایی برای دوره دوم هوش مصنوعی

تفسیر مدل های هوش مصنوعی و مدل تفسیر پذیر

تفسیر مدل های هوش مصنوعی

بیان تاریخچه

راه حل
مشکلات

مثال راه حل

هوش مصنوعی
توضیح پذیر

- تفسیر نتایج مدل براساس ورودی
- بررسی تاثیر تغییرات ورودی در خروجی
- بررسی روند طی شده در یک مدل
- ساخت مدل ساده تر با کارایی مشابه

9/18

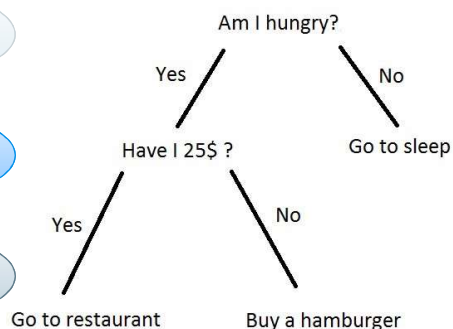
مدل تفسیر پذیر

بیان تاریخچه

راه حل
مشکلات

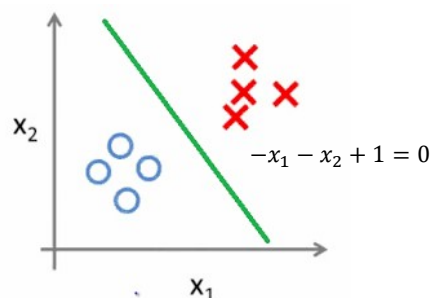
مثال راه حل

هوش مصنوعی
توضیح پذیر



(ب) درخت تصمیم - منبع

مشخص بودن نتایج مدل
قابل فهم بودن نتایج از پیش



(الف) دسته بندی دو دسته ای به کمک رگرسیون لجستیک - منبع

10/18

تفسیر مدل های هوش مصنوعی – مثال تصاویر [۲]

بیان تاریخچه

راه حل
مشکلات

مثال راه حل

هوش مصنوعی
توضیح پذیر



(a) Original Image
(الف)



(f) ResNet Grad-CAM 'Cat'
(ب)



(l) ResNet Grad-CAM 'Dog'
(ج)



A group of people flying kites on a beach A man is sitting at a table with a pizza

(a) Image captioning explanations
(د)

11/18

تفسیر مدل های هوش مصنوعی – مثال داده عددی [۳]

بیان تاریخچه

راه حل
مشکلات

مثال راه حل

هوش مصنوعی
توضیح پذیر

Prediction probabilities

False_Transport 0.07

True_Transport 0.93

False_Transport

True_Transport

Feature

Value

Feature	Value
CryoSleep_True	1.00
HomePlanet_Europa	0.00
CryoSleep_False	0.00
Age	17.00
Destination_55 Cancri e	0.00
HomePlanet_Earth	0.00
VIP_True	0.00
Destination_PSO J318.5-220.0	0.00
HomePlanet_Mars	1.00
Destination_TRAPPIST-1e	1.00

12/18

تفسیر مدل های هوش مصنوعی – مثال داده متنی [۴،۵]

بیان تاریخچه

راه حل
مشکلات

مثال راه حل

هوش مصنوعی
توضیح پذیر

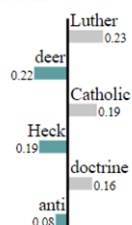
Text with highlighted words

Heck, I remember reading a quote of Luther as something like: "Jews should be shot like deer." And of course much Catholic doctrine for centuries was extremely anti-Semitic.

Prediction probabilities

christian	0.40
mid-east	0.38
atheism	0.10
guns	0.06
Other	0.07

NOT christian christian



13/18

۳. هوش مصنوعی در دوره سوم

هوش مصنوعی توضیح پذیر (Explainable AI)

هوش مصنوعی توضیح پذیر (XAI)

بیان تاریخچه

راه حل
مشکلات

مثال راه حل

هوش مصنوعی
توضیح پذیر

نتایج قابل فهم توسط افراد غیرمتخصص

ارائه دلایل خروجی مدل

دارای قابلیت اعتماد

شفافیت در نتایج

نتایج قابل فهم توسط افراد غیرخبره

15/18

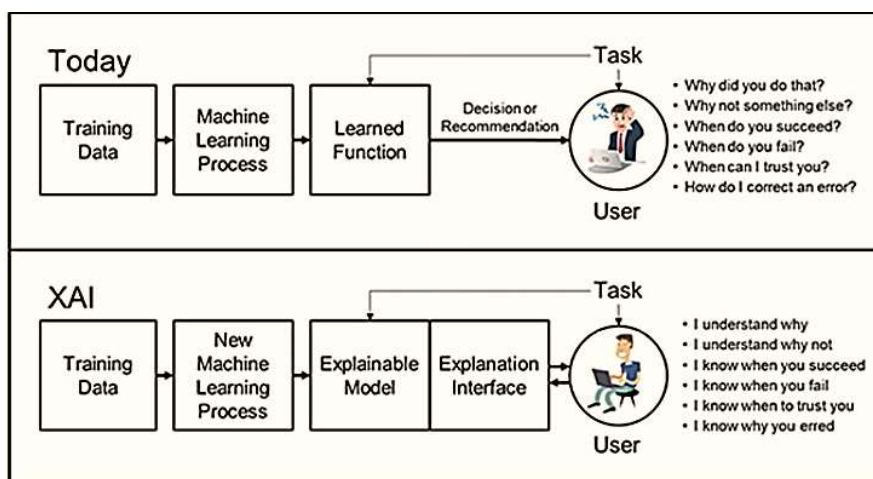
مدل های امروزه و هوش مصنوعی تفسیر پذیر [۶]

بیان تاریخچه

راه حل
مشکلات

مثال راه حل

هوش مصنوعی
توضیح پذیر



Toward Explainable Artificial Intelligence Through Fuzzy Systems Book, Jose Maria et. al.

16/18

مراجع

- [1] <https://www.darpa.mil/about-us/darpa-perspective-on-ai>
- [2] Selvaraju, R.R., Cogswell, M., Das, A. *et al.* Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *Int J Comput Vis* 128, 336–359 (2020). <https://doi.org/10.1007/s11263-019-01228-7>
- [3] <https://github.com/amindadgar/Spaceship-titanic-prediction>
- [4] Ribeiro, M., Singh, S. and Guestrin, C., 2022. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. [online] arXiv.org. <https://doi.org/10.48550/arXiv.1602.04938>.
- [5] Lime examples at <https://marcotcr.github.io/lime/tutorials/Lime%20-%20multiclass.html>
- [6] Alonso Moral, J.M., Castiello, C., Magdalena, L., Mencar, C. (2021). Toward Explainable Artificial Intelligence Through Fuzzy Systems. In: Explainable Fuzzy Systems. Studies in Computational Intelligence, vol 970. Springer, Cham. https://doi.org/10.1007/978-3-030-71098-9_1

17 / 18

ممنون از توجه همگی

سوال؟

راه ارتباطی:

تلگرام:

@mramin22

ایمیل:

amin.dadgar@eng.ui.ac.ir

18 / 18