الذكاء الاصطناعي الأخلاقي إشارات إلى الموارد الرئيسية للخبراء التقنيين

24 سبتمبر 2018

دبـــي الذكية
SMART DUBAI

إشارات إلى الموارد الرئيسية للخبراء التقنيين

24 سبتمبر 2018

تهدف هذه الوثيقة إلى مساعدة الخبراء الفنيين (مثل، علماء البيانات، مهندسي التعليم الآلي، مهندسي الذكاء الاصطناعي) في التحقق من آلية تطبيق الذكاء الاصطناعي الأخلاقي.

يرجى الملاحظة بأن هذه الوثيقة غير شاملة. ويُعد الذكاء الاصطناعي الأخلاقي مجالاً متطوراً مع توقع المزيد من التقدم في السنوات المقبلة. وقد تم توفير قائمة بالمصادر، كمرجع فقط، بدلاً من أن تكون قائمة مراجع رسمية. ويجب أن تقوم الجهات ذات الصلة بمفردها بتقييم ملائمة ومخاطر كل طريقة.

المحتويات

الحلول لبناء وقياس العدالة في أنظمة الذكاء الاصطناعي

يمكن تقسيم الطرق المقترحة لتوفير الذكاء الاصطناعي العادل إلى فئتين: (1) اكتشاف التمييز: كشف وتقييم التمييز الذي يظهر في البيانات، (2) منع التمييز: بناء النماذج التي تميل إلى عدم اتخاذ قرارات تمييزية حتى لو كانت متدرجة من مجموعة بيانات متحيزة.

1. **كشف التحيز وتقييم العدالة** [21]

a. **تقييم "العدالة" كمعدلات خطأ متساوية / دقة عبر مجموعات مختلفة:**

على سبيل المثال، معدلات إيجابية حقيقة متساوية، معدلات سلبية زائفة متساوية، المساواة الشاملة في الدقة، إلخ.

b. **تقدير "العدالة" كقيم متوسط متساوية أو معدلات إيجابية متساوية عبر مجموعات مختلفة:**

على سبيل المثال، متوسط الفرق، الفرق الموحد، التوازن في الدرجات للفئة الإيجابية/السلبية، المعايرة، التكافؤ تحت المنحنى، إلخ.

c. **تقدير "العدالة" كاستقلال مشروط على بعض الميزات**

على سبيل المثال، الخصائص المعينة لفرد ما، يجب إن تكون القرارات المتعلقة بها مستقلة عن مجموعتها الديموغرافية (مثل، النوع الاجتماعي)

المقالات على الإنترنت:

- *Mirror Mirror" - Reflections on Quantitative Fairness"*
  /https://speak-statistics-to-power.github.io/fairness
- *Fairness Measures*
  /http://fairness-measures.org

المجلات أو الوقائع أو العمل الجاري:

- *Group Fairness Under Composition, C. Dwork and C.Ilvento, 2018*

d. **تقييم العدالة الفردية**

أي، يجب تصنيف الفردين المتشابهين في مهمة محددة بطريقة مماثلة

المجلات أو الوقائع أو العمل الجاري:

- *Learning fair representations, Zemel et al., 2013*
- *Fairness through awareness, Dwork et al., 2011*
- *Individual Fairness Under Composition, C. Dwork and C.Ilvento, 2018*

e. **طرق التعلم الآلي بغرض الكشف والتقييم**

---

[1] مقاييس التقييم المقتبسة جزئيًا من *Mirror Mirror- Reflections on Quantitative Fairness* (راجع *https://speak-statistics-to-power.github.io/fairness/*). مرخص بموجب *Creative Commons Attribution-ShareAlike 4.0 International License*

[2] مقاييس التقييم المقتبسة جزئيًا من *Fairness Measures* (راجع */http://fairness-measures.org* مرخص بموجب *CC-BY-4.0 license*

المجلات أو الوقائع أو العمل الجاري:

- *k-NN as an Implementation of Situation Testing for Discrimination Discovery and Prevention, Luong et al., 2011*
- *Combating discrimination using bayesian networks, K. Mancuhan and C. Clifton, 2014*
- *Data mining for discrimination discovery, Ruggieri et al., 2014*

f. الطرق الإحصائية للكشف والتقييم

المجلات أو الوقائع أو العمل الجاري:

- *Measuring racial discrimination, Panel on Methods for Assessing Discrimination, Blank et al., 2004.*
- *Statistical evidence of discrimination, D. Kaye, 1982*
- *The uses and misuses of statistical proof in age discrimination claims, T. Tinkham., 2010*

2. **إزالة الظلم**

a. إزالة الظلم من خلال المعالجة المسبقة لمجموعات البيانات التدريبية

المجلات أو الوقائع أو العمل الجاري:

- *Certifying and removing disparate impact, Feldman et al., 2015*
- *Quantifying explainable discrimination and removing illegal discrimination in automated decision making, Kamiran et al., 2013*
- *Combating discrimination using bayesian networks, K. Mancuhan and C. Clifton, 2014*

b. إزالة الظلم في النماذج من خلال إضافة مُنظم

المجلات أو الوقائع أو العمل الجاري:

- *Fairness-aware classifier with prejudice remover regularizer, Kamishima et al., 2012*
- *Learning fair representations, Zemel et al., 2013*
- *Controlling attribute effect in linear regression, Calders et al., 2013*
- *Discrimination aware decision tree learning, Kamiran et al., 2010*
- *Learning fair representations, Dwork et al., 2013*

c. إزالة الظلم من خلال المعالجة اللاحقة للنماذج التدريبية

المجلات أو الوقائع أو العمل الجاري:

- *A methodology for direct and indirect discrimination prevention in data mining, S. Hajian and J. Domingo-Ferrer, 2013*

d. إزالة الظلم من خلال المعالجة اللاحقة لنتائج النماذج

المجلات أو الوقائع أو العمل الجاري:

- *Discrimination aware decison tree learning, F. Kamiran, T. Calders, and M. Pechenizkiy, 2010*

مصادر أخرى بشأن عدالة الذكاء الاصطناعي:

مواد المساقات:

- *CS294: Fairness in Machine Learing (Lectures), UC Berkeley*
  https://fairmlclass.github.io/

الأدوات:

- تدابير العدالة - مستودع الشفرات
  https://github.com/megantosh/fairness_measures_code/tree/master
- *التعلم الآلي للعدالة: تدقيق نماذج توقع الصندوق الأسود*
  https://github.com/adebayoj/FairML

الحلول لتوثيق أنظمة الذكاء الاصطناعي

3 طرق محتملة لتوثيق وتفسير البيانات التي تم عرضها:

1. **جداول البيانات الخاصة بمجموعات البيانات**
   - طريقة لتوثيق كيفية وسبب إنشاء مجموعة بيانات معينة والمعلومات التي تحتوي عليها والمهام التي يجب أو لا يجب استخدامها وإذا كانت ستؤدي إلى إثارة أي مخاوف أخلاقية أو قانونية
2. **مخطط البيان الخاص بالبيانات**
   - تم بناؤه من أجل أنظمة معالجة اللغة الطبيعية
   - يسعى لمعالجة قضايا الأخلاقيات والاستثناءات والتحيز في أنظمة معالجة اللغة الطبيعية
3. **تصنيف المعلومات المدخلة في مجموعة البيانات**

المجلات أو الوقائع أو العمل الجاري:

- *Datasheets for datasets, Timnit Gebru et al., 2018*
- *Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science, Emily M. Bender and Batya Friedman,*
- *The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards, Sarah Holland et al., 2018*

طرق التوثيق العام:

1. **إقرارات الموردين بشأن المطابقة**
   - طريقة تقدم قائمة شاملة للبنود والأسئلة والأجوبة حول أنظمة الذكاء الاصطناعي التي يتعين توثيقها
2. **بيان الأثر الاجتماعي للخوارزميات**
   - طريقة التوثيق المقترحة لضمان توفر الذكاء الاصطناعي الأخلاقي

المجلات أو الوقائع أو العمل الجاري:

- *Increasing Trust in AI Services through Supplier's Declarations of Conformity, Hind et al., 2018*

المقالات على الإنترنت:

- *Principles for Accountable Algorithms and a Social Impact Statement for Algorithms, FAT/ML* http://www.fatml.org/resources/principles-for-accountable-algorithms

الحلول لتوفير ذكاء اصطناعي قابل للتفسير

توجد بشكل عام 3 أنواع من الطرق لتوفير الذكاء الاصطناعي القابل للتفسير[3]:

1. **استخدام النماذج القابلة للتفسير مباشرة**
   - على سبيل المثال، نموذج خطي، انحدار لوجستي، شجرة القرارات، قواعد اتخاذ القرارات (قواعد إذا-فإن)، قاعدة المواءمة، طرق بايزي البسيطة، خوارزمية تصنيفية، خوارزمية الغابات العشوائية / التعزيز، نماذج ماركوف المخفية، إلخ.
   - يجب تقديم التوضيحات بواسطة، مثلًا سرد الميزات المحددة وأوزانها.

الكتب:

- *Interpretable Machine Learning - A Guide for Making Black Box Models Explainable, Molnar. C. , 2018*
  /https://christophm.github.io/interpretable-ml-book

التقارير:

- *Explainable AI - Driving business value through greater understanding, PricewaterhouseCoopers LLP. , 2018*
  https://www.pwc.co.uk/audit-assurance/assets/explainable-ai.pdf

2. **استخدام طرق التفسير المحايدة للنموذج**
   - يمكن تطبيق الطرق التي لا تحدد نماذج التعلم الآلي الأساسية بشكل عام على نماذج الصندوق الأسود
   - على سبيل المثال، قطعة أرض مستقلة جزئياً، استثناء مشروط لفرد، تفاعل ميزة، أهمية ميزة، نموذج بديل عالمي، مخطط آثار محلية متراكمة، نموذج بديل محلي، تفسيرات شابلي للقيمة، إلخ.

الكتب:

- *Interpretable Machine Learning - A Guide for Making Black Box Models Explainable, Molnar. C. , 2018*
  /https://christophm.github.io/interpretable-ml-book

المجلات أو الوقائع أو العمل الجاري:

- *Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models, Apley, D. W., 2016*
- *Why Should I Trust You?" Explaining the Predictions of Any Classifier, Ribeiro et al., 2016"*

[3] *Types of methods adapted from Interpretable Machine Learning - A Guide for Making Black Box Models Explainable, Molnar. C. , 2018 (see https://christophm.github.io/interpretable-ml-book/). Licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.*

الأدوات:

- *LIME Python package & R package*

https://github.com/marcotcr/lime

https://cran.r-project.org/web/packages/lime/index.html

3. **استخدام طرق التفسير من خلال المثال على حالة**
   - على سبيل المثال، التفسيرات المغايرة والنموذج الأصلي والانتقادات والحالات المؤثرة، إلخ.

الكتب:

- *Interpretable Machine Learning - A Guide for Making Black Box Models Explainable, Molnar. C. , 2018*

/https://christophm.github.io/interpretable-ml-book

المجلات أو الوقائع أو العمل الجاري:

- *Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR, S. Wachter, B. Mittelstadt and C. Russell, 2018*
- *Inverse Classification for Comparison-based Interpretability in Machine Learning, Laugel et al., 2017*
- *Examples are not Enough, Learn to Criticize! Criticism for Interpretability", B. Kim, R. Khanna, " and O. Koyejo, 2016*

الطرق التي تم اقتراحها وفقاً للمراحل الثلاث المختلفة لتطوير نظام الذكاء الاصطناعي:

1. **ما قبل البناء**

مثل، التخيل، تحليل البيانات الاستكشافية

2. **البناء**

مثل، قائم على قاعدة، قائم على مفهوم لكل ميزة، قائم على الحالة، طريقة التبعثر، طريقة الرتابة

3. **ما بعد البناء**

مثل، تحليل الحساسية، الطرق القائمة على التدرج، نماذج المحكاة/البديلة، استقصاء الطبقات المخفية

البرامج التعليمية:

التعلم الآلي القابل للتفسير: The fuss, the concrete and the questions, B.Kim and F. Doshi-Velez, 2017

https://people.csail.mit.edu/beenkim/papers/BeenK_FinaleDV_ICML2017_tutorial.pdf

مصادر أخرى بشأن قابلية الذكاء الاصطناعي للتفسير:

المجلات أو الوقائع أو العمل الجاري:

- *Towards A Rigorous Science of Interpretable Machine Learning, F. Doshi-Velez and B. Kim, 2017*

المصادر الأخرى

المنظمات والمؤتمرات

(*ACM Conference on Fairness, Accountability, and Transparency (ACM FAT

/https://fatconference.org

ACM SIGKDD

/http://www.kdd.org

NIPS Conference

ICML Conference

قوائم القراءة حول الذكاء الاصطناعي الأخلاقي

Toward ethical, transparent and fair AI/ML: a critical reading list

https://medium.com/@eirinimalliaraki/toward-ethical-transparent-and-fair-ai-ml-a-critical-reading-list-d950e70a70ea