Causalité 3 – Classifieurs de textes avec régularisation causale

Séminaire TALia du 13/01/2023

Séminaires TALia pertinents

- 65 – Causalité 1 - l'approche SCM de Pearl
- 67 – Causalité 2 - prédictions causales avec des textes

NeurIPS | 2021

arXiv > cs > arXiv:2106.00545

Computer Science > Machine Learning

[Submitted on 31 May 2021 (v1), last revised 2 Nov 2021 (this version, v3)]

Counterfactual Invariance to Spurious Correlations: Why and How to Pass Stress Tests

Victor Veitch, Alexander D'Amour, Steve Yadlowsky, Jacob Eisenstein

Vidéo de V. Veitch [15 min]

arXiv > cs > arXiv:2109.00725

Computer Science > Computation and Language

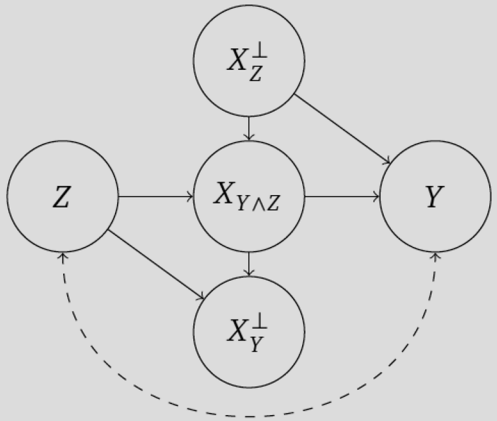
[Submitted on 2 Sep 2021 (v1), last revised 30 Jul 2022 (this version, v2)]

Causal Inference in Natural Language Processing: Estimation, Prediction, Interpretation and Beyond

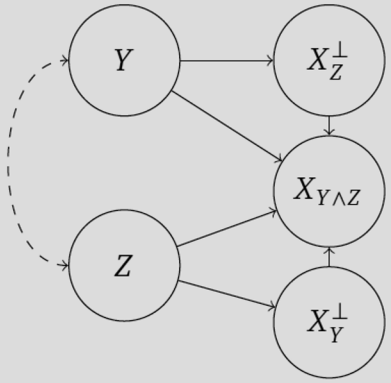
Amir Feder, Katherine A. Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E. Roberts, Brandon M. Stewart, Victor Veitch, Diyi Yang

Plan

1. Rappels des séminaires n°1 et n°2 sur la causalité
2. Deux exemples élémentaires de corrélations fictives
3. Classifieurs *Counterfactually Invariant* (CFI)
4. Le cas causal et le cas anti-causal
5. La signature expérimentale d'un classifieur CFI
6. La précision et la robustesse des prédicteurs CFI
7. Résultats expérimentaux sur *Amazon Product Review*



(a) Causal direction

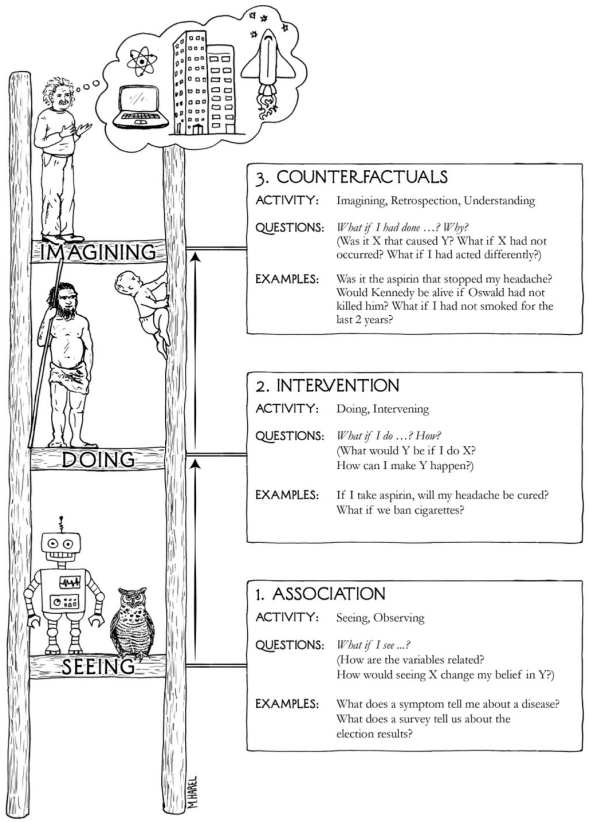


(b) Anticausal direction

Rappel des séminaires Causalité 1 et Causalité 2

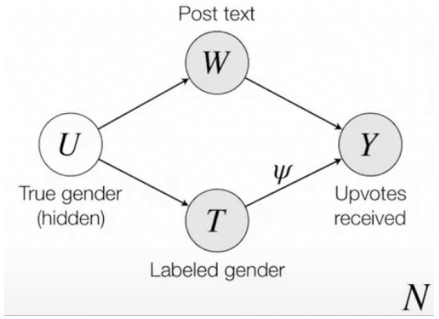
Causalité 1 (les concepts selon J. Pearl)

- Inférence causale
- Les données ne suffisent pas !
- Les SCM
- Les graphes causaux
- L'opérateur $\mathrm{do}(X = x)$
- L'identifiabilité causale
- La d-séparation
- La back-door rule
- L'ajustement des confounders
- Le do-calculus



Causalité 2 (le TAL au service de l'inférence causale)

- Impact causal de l'affichage d'une icône de genre T sur la popularité Y d'un post sur un réseau social, le texte W étant un confounder.
- Evaluation de l'ATT au moyen d'une réduction dimensionnelle du texte qui préserve l'identification causale
- Modèles C-BERT et C-ATM

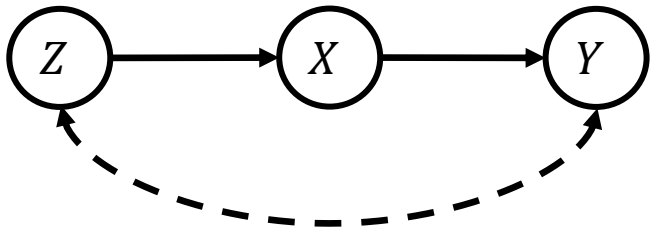


Causalité 3 (la causalité au service du TAL)

Deux problèmes de classification de texte

Prédiction causale de la *helpfullness* d'un commentaire

Prédire l'utilité Y (*helpfulness*) de l'évaluation d'un produit à partir de son texte X (cause) indépendamment de l'émotion Z exprimée (cause de X).



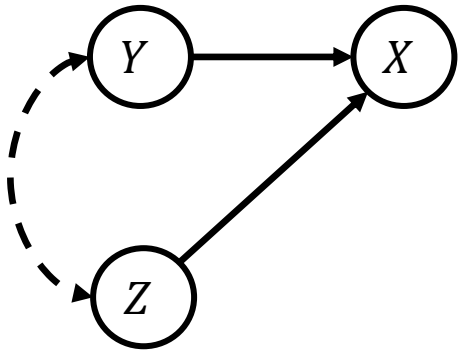
"If books tend to get more positive reviews and people who buy books are more likely to flag reviews as helpful, then the product type would be a common cause of sentiment and helpfulness."

- X texte de l'évaluation du produit
- Y *helpfulness* de l'évaluation (mesurée par vote)
- Z tonalité émotionnelle de l'évaluation

Prédiction anti-causale du nombre d'étoiles d'un film

Prédire le nombre d'étoiles Y (cause) d'un film à partir d'un commentaire X indépendamment de sa catégorie Z (cause de X).

"For example, fans of romantic comedies may tend to give higher reviews than fans of horror movies."



- X texte du commentaire du film
- Y nombre d'étoiles attribuées au film
- Z catégorie du film commenté

Le problème de corrélations fictives (*spurious correlations*)

- Un modèle de machine learning que l'on entraîne (avec ERM) à prédire Y à partir d'un texte X utilisera toute l'information disponible dans X même celle qui n'a **aucun lien causal** avec Y

Un problème élémentaire d'alignement d'une IA !

- On souhaiterait **garantir l'invariance des prédictions** vis-à-vis de certaines variations du texte X :

  - "**Keanu Reeves** stars in this excellent thriller" → ★★★★
  - "**Kevin Costner** stars in this excellent thriller" → ★

Le nombre d'étoiles ne devrait pas dépendre de la catégorie du film, peut-être implicitement révélée par le nom de l'acteur.

- Mesurer l'invariance des prédictions à l'aide de « ***stress tests*** » qui modifient le texte d'une manière qui ne devrait avoir aucun impact

Easy
"**He** is a doctor" → "**She** is a doctor"

"This great movie stars **Keanu Reeves**" → "This great movie stars **Kevin Costner**"

Hard
"**Terrible** jeans, **poorly** made" → "**Excellent** jeans, **great craftsmanship**"

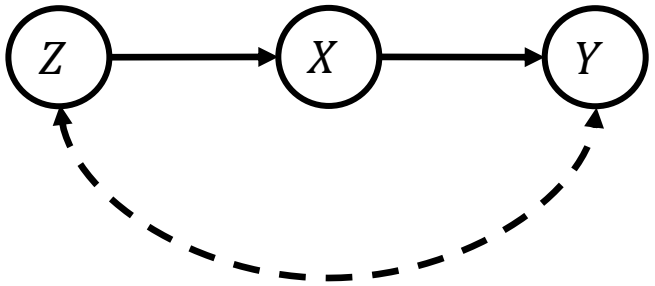
Q1 : quel est l'impact de l'invariance :

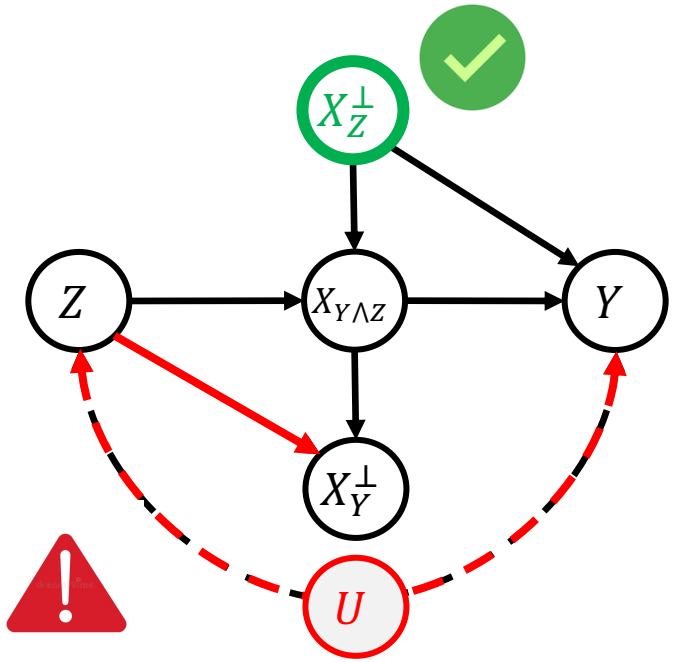
- Sur la précision des prédictions ?
- Sur la robustesse vis-à-vis d'une distribution cible différente de la distribution d'entraînement (OOD) ?

Q2 : comment créer des stress tests dans les cas difficiles ?

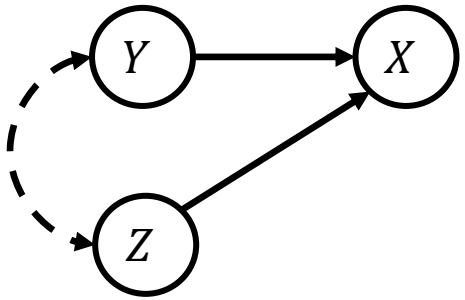
Zoom sur le cas causal et sur le cas anti-causal

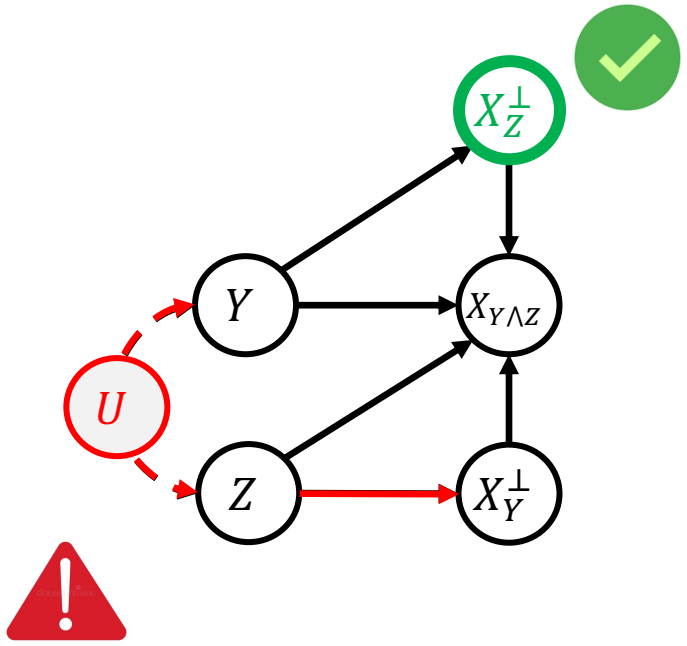
Prédiction **causale**



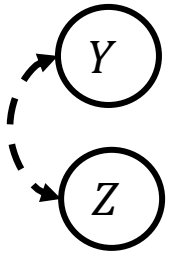


Prédiction **anti-causale**

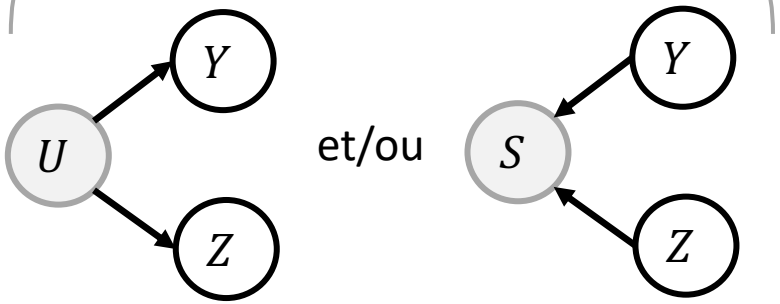




Deux mécanismes à l'origine de la corrélation entre Y et Z.



=



cause commune non observée

processus de sélection

$$P(X,Y,Z) = \sum_U P(X,Y,Z,U|S=1)$$

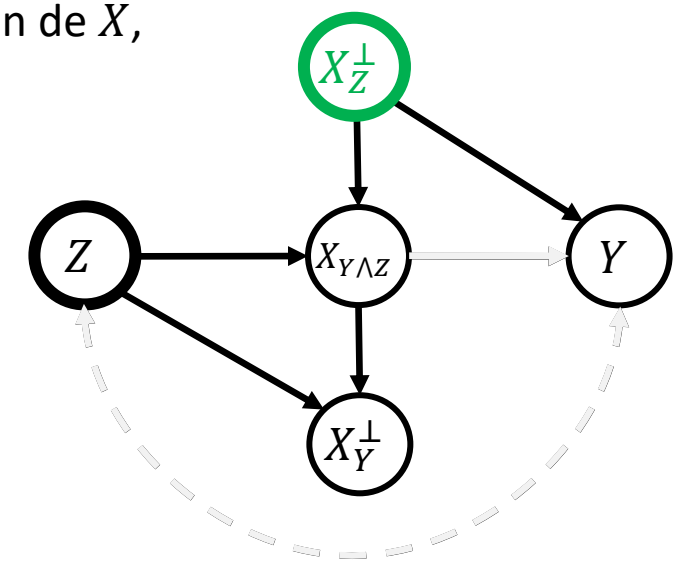
Formalisation avec les outils de la causalité

L'invariance causale est formulée (dans le cas où Z est binaire) à l'aide de **variables contrefactuelles** $X(z=0)$ et $X(z=1)$ **non observables**.

C'est un concept défini à l'origine dans l'**approche Potential Outcome** de la causalité mais on peut le définir dans l'**approche de Pearl** comme la v.a. induite par une opération $\mathrm{do}(Z=0)$ ou $\mathrm{do}(Z=1)$ sur le SCM original.

Définition : Un predicteur $f(X)$ de Y est **contrefactuellement invariant** (CFI) si $f(X(z=0)) = f(X(z=1))$ p.s.

Lemme : Dans le cas où Z est discrète il existe une variable $X_Z^{\perp}$, fonction de X, telle que f est CFI ssi $f(X) = f(X_Z^{\perp})$.



Définition : Une association entre Y et Z est dite **purement fictive** (*purely spurious*) si

$$Y \perp\!\!\!\perp X | X_Z^{\perp}, Z$$

Intuition : une fois éliminée l'association non causale entre Z et Y, la variable $X_Z^{\perp}$ est la seule utile pour prédire de Y

La propriété CFI est invérifiable à partir des données !

Remarque : Puisque la définition de la **propriété CFI** d'un prédicteur s'exprime à l'aide de variables contrefactuelles non-observables elle est, au sens strict du terme, **invérifiable** !

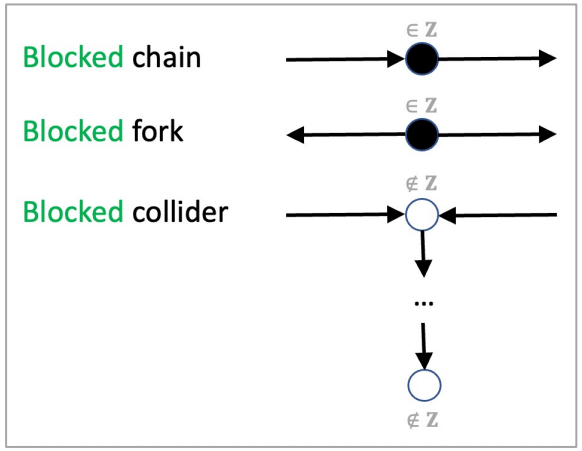
Stratégie :

1. On identifie des relations d'indépendance **expérimentalement vérifiables** qui sont des **signatures de CFI**. Ce sont des conditions **nécessaires** mais **insuffisantes** pour garantir CFI.
2. On définit une **régularisation** de ERM qui favorise les prédicteurs $f(X)$ qui vérifient approximativement ces conditions d'indépendance.
3. On **vérifie expérimentalement l'impact** de la régularisation sur la robustesse vis-à-vis de stress tests et sur la capacité de généralisation OOD.

La signature d'un prédicteur CFI et la régularisation causale

Théorème : Si $f(X)$ est un $\overbrace{\text{prédicteur CFI}}^{\text{non vérifiable expérimentalement}}$ alors :

1. dans le cas **causal** si Y et Z ne sont soumis à aucune sélection : $\overbrace{f(X) \perp\!\!\!\perp Z}^{\text{indép. marginale vérifiable}}$
2. dans le cas **causal** si Y et Z n'ont pas de confounder U et si leur association est purement fictive : $\overbrace{f(X) \perp\!\!\!\perp Z \mid Y}^{\text{indép. conditionnelle vérifiable}}$
3. dans le cas **anti-causal** : $\underbrace{f(X) \perp\!\!\!\perp Z \mid Y}_{\text{indép. conditionnelle vérifiable}}$



On favorise une signature correcte à l'aide d'une régularisation causale du prédicteur $f(X)$ qui pénalise les prédicteurs qui s'écartent trop de ces relations d'indépendance.

$$f^* := \arg\min_f \left(\mathbb{E}_P[\ell(Y, f(X))] + \lambda \, \mathrm{reg}_{\ldots}[f] \right)$$ où

C'est le mieux qu'on puisse faire !

$$\mathrm{reg}_{\text{marginal}}[f] := \mathrm{MMD}[P(f(X)|Z=0),\, P(f(X)|Z=1)]$$

$$\mathrm{reg}_{\text{conditional}}[f] := \sum_{y \in \{0,1\}} \mathrm{MMD}[P(f(X)|Z=0, Y=y),\, P(f(X)|Z=1, Y=y)]$$

Interlude sur la notion de distance MMD

$\chi \ni x \to \phi(x) \in \mathscr{H}$ l'idée de feature mapping avec un produit scalaire $\langle \phi(x), \phi(y) \rangle_{\mathscr{H}} := k(x, y)$

$\text{RKHS}: \text{ si } f \in \mathscr{H} \text{ alors } f(x) = \langle f, \phi(x) \rangle_{\mathscr{H}}$ un espace « de luxe » avec une version soft de la fonction de Dirac

$p \to \mu_p := \mathbb{E}_{x \sim p}[\phi(x)] \in \mathscr{H}$ une distribution p sur χ est représentée par l'espérance des représentations

$$\mathbb{E}_{x \sim p}[f(x)] = \mathbb{E}_{x \sim p}[\langle f, \phi(x) \rangle_{\mathscr{H}}] = \langle f, \mu_p \rangle_{\mathscr{H}}$$

L'article original sur MMD (2012)

$$\begin{aligned} \text{MMD}[p, q] &:= \|\mu_p - \mu_q\|_{\mathscr{H}} \\ &= \sup_{\|f\|_{\mathscr{H}} \leq 1} \langle f, \mu_p - \mu_q \rangle_{\mathscr{H}} \\ &= \sup_{\|f\|_{\mathscr{H}} \leq 1} |\mathbb{E}_{x \sim p}[f(x)] - \mathbb{E}_{y \sim q}[f(y)]| \end{aligned}$$

le plus grand « contraste » qu'on peut révéler entre p et q.

$$\widehat{\text{MMD}}[p, q]^2 := \frac{1}{m(m-1)} \sum_{i=1}^{m} \sum_{j \neq i}^{m} k(x_i, x_j) - 2 \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} k(x_i, y_j) + \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} k(y_i, y_j)$$

Performances des prédictions hors domaine (OOD 1)

Question : Que gagne-t-on du point de vue des performances à effectuer une régularisation causale d'un classifieur $f(X)$?

Pour répondre à cette question il faut spécifier quels sont les changements de domaine que l'on décide d'admettre.

Définition : les distributions P et Q sont **causalement compatibles** si :

1. Elles sont associées au même graphe causal (même processus génératif).
2. Elles ont les mêmes distributions marginales $P(Y) = Q(Y)$.
3. Les corrélations entre Y et Z peuvent varier
   (induite par un confounder U ou par une sélection S)

$$P(X,Y,Z) = \sum_U P(X,Y,Z|U,S=1)\,\tilde{P}(U)$$

$$Q(X,Y,Z) = \sum_U P(X,Y,Z|U,\tilde{S}=1)\,\tilde{Q}(U)$$

Performances des prédictions hors domaine (OOD 2)

Théorème : Soit $\mathcal{F}^{\mathrm{CFI}}$ l'ensemble des prédicteurs $f(X)$ qui sont CFI. Soit $\ell(Y, \hat{Y})$ une fonction de coût qui est soit quadratique, soit la cross-entropie. Supposons que P et Q soient causalement compatibles.

Si l'une des conditions suivantes est vérifiées :

- Le graphe est **anti-causal**
- Le graphe est **causal** mais il n'y a pas de confounder U et l'association entre Y et Z est purement fictive

Alors

$$f^* := \arg\min_{f \in \mathscr{F}^{\mathrm{CFI}}} \mathbb{E}_P[\ell(Y, f(X))] = \arg\min_{f \in \mathscr{F}^{\mathrm{CFI}}} \mathbb{E}_Q[\ell(Y, f(X))]$$

Conclusion : Même sans données dans le domaine cible on peut trouver un bon prédicteur en cherchant un prédicteur CFI sur les données d'entraînement.

Question : Est-ce que f^* est le meilleur prédicteur tout court pour Q ?

Réponse : Non en général sauf si l'association entre Y et Z est purement fictive + conditions techniques.

Résultats expérimentaux – robustesse aux corrélations fictives

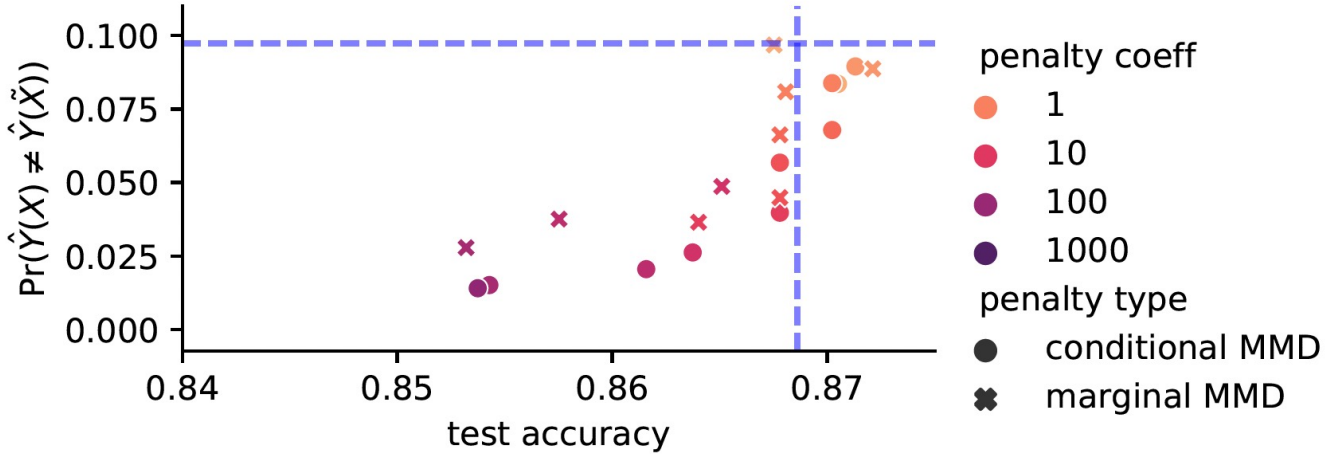
Prédiction causale

X = texte de l'évaluation du produit
Y = « *helpfulness* » de l'évaluation
Z = tonalité émotionnelle du texte (exprimée par le nb d'étoiles)

Prédiction anti-causale

X = texte de l'évaluation du produit
Y = catégorie du produit (∃ tag = « *clothing* »)
Z = tonalité émotionnelle du texte (exprimée par le nb d'étoiles)

- Idéalement la tonalité émotionnelle ne devrait avoir aucun impact sur la prédiction de Y
- On fait varier cette tonalité en permutant des adjectifs positifs et négatifs : $X \leftrightarrow \tilde{X}$
- On fixe $P(Y = Z) = 0.3$ par sélection des données

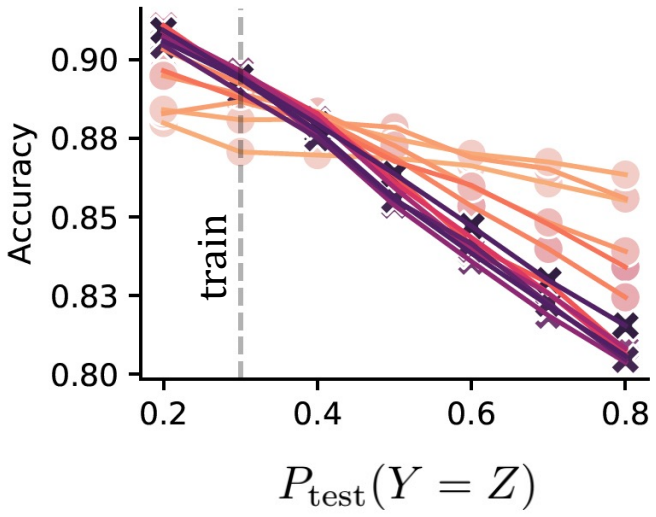


Théorème : Si $f(X)$ est un prédicteur CFI alors :

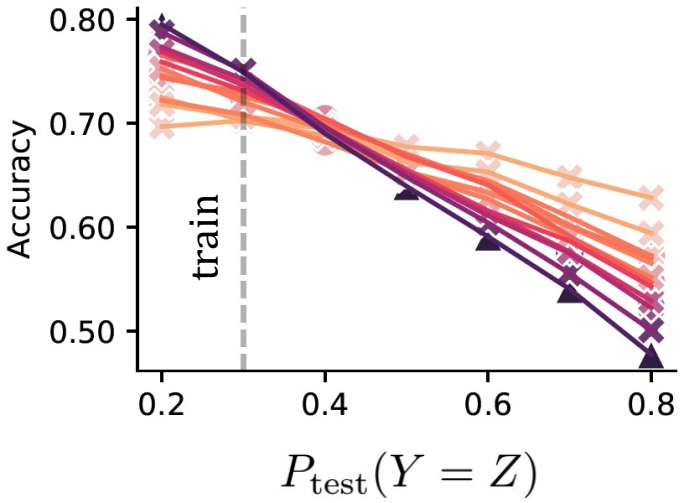
1. cas causal si Y et Z et si aucune sélection : $f(X) \perp\!\!\!\perp Z$
2. cas anti-causal : $f(X) \perp\!\!\!\perp Z \mid Y$

Résultats expérimentaux – robustesse OOD

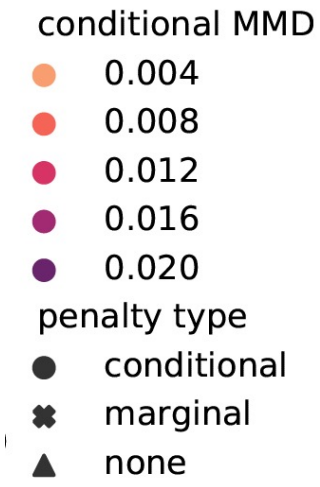
- On évalue la précision des prédictions sur la cible
- On fait varier la corrélation $P_{\text{test}}(Y = Z)$ entre le label Y et le confounder Z
- On fait varier l'amplitude λ de la régularisation MMD



Données **anti-causales**



Données **causales**



Pistes de recherche…

1. Autres structures causales plus complexes
2. Traiter le cas où Z a beaucoup de valeurs ou est continu
3. Peut-on renforcer les résultats si l'on se restreint à des modèles paramétriques ?
4. Dans quelles situations pourrait-on exploiter les idées présentées pour dire quelque chose de la structure causale lorsque celle-ci est inconnue ou incertaine ?
5. Généralisation au cas $P(Y) \neq Q(Y)$
6. Identifier des connections avec l'adaptation de domaine
7. Création de stress tests pour les cas difficiles (quels prompts pour faire faire le travail par ChatGPT ?) ou créer des paires minimales avec un modèle génératif causal
8. Explorer les connexions entre la causalité et les questions de *fairness* (*demographic parity*, *equality of odds*)
9. Applications pratiques ?
10. Relation avec la détection d'anomalies (le cas des entités dans le problème avec Aurélien) ?
11. Toutes les questions auxquelles je n'ai pas pensé (un ensemble infini non dénombrable)

Résumé

1. **Formalisation** de la notion de stress test
2. Mise en évidence d'une **signature observable** de CFI d'un classifieur à l'aide de conditions d'indépendance
3. Définition d'une **procédure de régularisation**, nécessaire mais insuffisante pour garantir CFI
4. La **régularisation** dépend de la **structure causale** des données analysées
5. La régularisation **améliore** en pratique la **robustesse aux stress tests**
6. La régularisation **améliore** en pratique la **robustesse aux changements de domaine** qui préservent la structure causale.