Évaluer la précision des données d'entrée et du modèle résultant

Saeid Molladavoudi (Scientifique principal des données)

Division de la science des données, Statistiques Canada

Une conférence du gouvernement du Canada

Le 18 février 2021

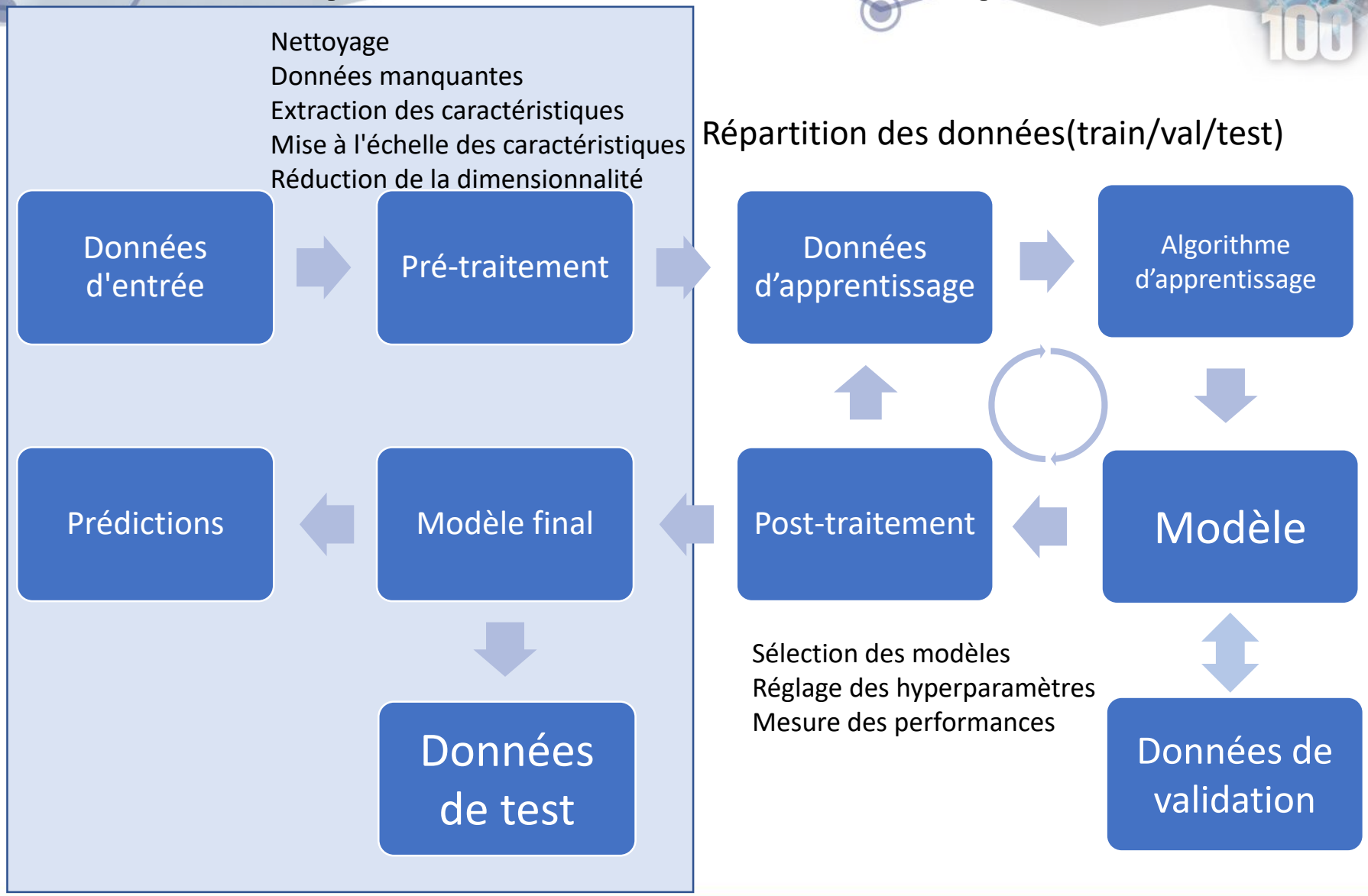
Éclairer grâce aux données pour bâtir un Canada meilleur.

Considérations générales

- **Chaque projet d'AA commence par une discussion avec les personnes concernées et les propriétaires des données.**
- **Les points suivants doivent être clairs dès le début :**
  1. **Problème :** identifier le besoin des entreprises
  2. **Champ d'application :** Approches d'Apprentissage Automatique supervisées et non supervisées
  3. **Objectifs :** Classification, Régression, Regroupement ...
  4. **Contraintes :** Préoccupations concernant les données (collecte, montant, ...)
  5. **Produits livrables :** Solutions consommables pour les machines et solutions consommables pour l'homme
  6. **Critères de réussite :** Économie d'heures de travail manuel ou en temps de traitement, augmentation de la qualité, etc.

https://www.jeremyjordan.me/ml-projects-guide/

Flux d'exploration de modèles supervisés



Les données d'entrée : la partie la plus importante d'un projet d'AA

Principaux défis	• Manque de données • Mauvaise qualité • Données d'apprentissage non représentatives de la situation • Pas assez d'éléments pertinents • Trop de caractéristiques non pertinentes
Questions auxquelles nous devons répondre	• Y a-t-il "suffisamment" de données ? • Combien de sources de données existent ? • Combien de données provenant de chaque source ? • Est-il possible d'acquérir davantage de données ? • Les données sont-elles structurées (par exemple via l'API) ou non structurées (par exemple des images) ?
Par exemple, si le problème est un cas d'utilisation supervisée d'AA	• Est-ce qu'il y a des labels ? • Quelle est la quantité de données avec des labels ? • Est-il possible d'acquérir d'autres labels ?

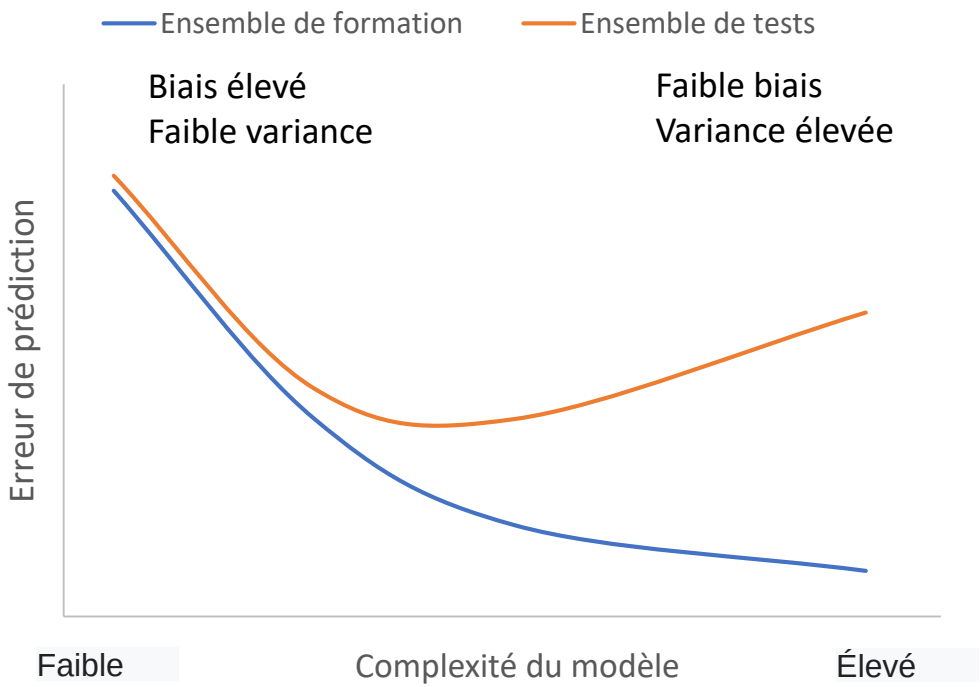
Analyse exploratoire des données (pré-traitement)

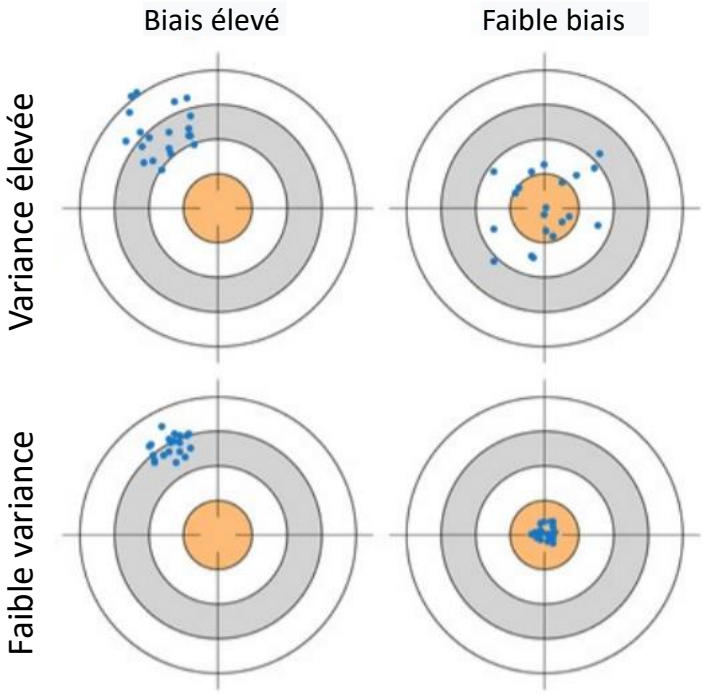
A considérer	Détails
Types de données	Hétérogènes (continues, catégoriques, textes, images, ...)
Transformations	Ingénierie des caractéristiques, catégorisation à un seul coup, normalisation, intégration de mots, matrice de termes de documents, ...
Extraction	Pipeline s'il s'agit de données non structurées, par exemple vidéo, audio, pdfs, images ...
Labels	Classification (binaire, multi-classe, multi-label) vs labels de régression
Quantité de données	Existe-t-il "suffisamment" de données pour créer et entraîner, faire la validation des groupes d'essai ?
Déséquilibre des classes	Distribution des labels (déséquilibre de classe, évolution dans le temps)
Visuels	Outils visuels, par exemple diagrammes de dispersion et histogrammes
Statistiques de synthèse	En particulier pour les données structurées, par exemple pour trouver les corrélations entre les caractéristiques
Valeurs manquantes	Imputation ou retrait ?
Dimensionnalité	Réduction de la dimensionnalité des données d'entrée, par exemple l'analyse en composantes principales
Autres types de tâches	Réseaux, séries chronologiques, regroupement, détection des anomalies ....

Principe de formation, de validation et d'essai

Ensemble de données	Objectifs
Ensemble de formation	- L'objectif de l'AA est de minimiser la perte attendue (minimisation empirique du risque). - Les modèles s'inspirent d'exemples de formation.
Ensemble de validation	- La validation est utilisée pour trouver les meilleurs paramètres du modèle (réglage des hyperparamètres).
Ensemble de tests	- Le test est utilisé pour simuler ce qui se passera lorsque nous appliquerons le modèle final à de nouvelles données invisibles provenant de la même distribution que les données de formation (c'est-à-dire le monde réel). - Il donne une estimation non biaisée de l'erreur de généralisation du modèle.

Compromis biais-variance





https://www.variancejournal.org/articlespress/articles/Bias-Variance_Brady-Brockmeier.pdf

Classification supervisée AA

Taille de l'ensemble de données	Labels	Stratégies d'atténuation
Petit	Peu de références pour les classes minoritaires.	• Utilisez la validation croisée en k. • Ajoutez une régularisation au risque empirique pour éviter le sur-ajustement. • Utiliser des modèles d'ensemble. • Utiliser des méthodes d'augmentation des données. • Transférer l'apprentissage. • Réduction de la dimensionnalité.
Moyen	Peu de références pour les classes minoritaires.	• Utilisez la validation croisée en k. • Ajoutez moins de régularisation. • Utiliser des modèles d'ensemble. • Utiliser des méthodes d'augmentation des données. • Transférer l'apprentissage. • Courbes d'apprentissage à rechercher en cas de sous-ajustement ou de surajustement. • Outil d'annotation.
Grand	Petite quantité de données labellisées.	• Apprentissage semi-supervisé. • Apprentissage actif. • Outil d'annotation.

Évaluation des performances

Tâche d'AA	Perte	Mesures
Classification	- Entropie croisée (binaire) - Probabilité logarithmique négative (perte logarithmique) - Perte de la charnière - Divergence Kullback-Leibler - Entropie relative - Perte focale	- Exactitude - Précision - Rappeler - F-score - Courbe ROC - AUC
Régression	- Erreur quadratique moyenne (EMS) ou perte quadratique - (Lissage) Erreur Absolue Moyenne (EAM) - Perte de bois	- EMS - EAM
Regroupement	- somme des carrés à l'intérieur d'un groupe	- Indice Rand - Score d'information mutuelle - Coefficients de silhouette
Classement	- Perte de classement des marges - Triple perte de marge	- Rang réciproque moyen - Précision @ k - Gain cumulé normalisé et actualisé

Model Refinements:

- Utiliser la recherche par grille ou l'échantillonnage aléatoire pour le réglage des hyperparamètres.

Assurance de la qualité

Éléments à prendre en considération	Détails
Incertitude	• Aléatoire : Bruit d'observation dans les données. • Epistémique : Incertitude dans le modèle, les paramètres et la convergence.
Biais dans les données	• Biais humain dans la labellisation des données, par exemple, au sein d'un groupe, homogénéité et biais implicites. • Le biais de collecte qui se réfère aux données qui ne reflètent pas les distributions réelles, par exemple le biais de convergence, le biais de déclaration, le biais de participation et le biais d'échantillonnage.
Sécurité et vie privée	• Les technologies de protection de la vie privée et leurs applications dans le domaine de l'AA, par exemple le cryptage homomorphe, les calculs multipartites sécurisés, ... • Confidentialité des entrées et des sorties (confidentialité), par exemple, confidentialité différentielle
Reproductibilité	• Contrôle de la version du code (développement, test, mise en œuvre) et des données • Pipeline automatisé
Explicabilité	• Transparence • Documentation
Maintenance	• Surveiller la cohérence, l'exactitude, la couverture et la représentativité des données d'apprentissage dans le temps afin de détecter toute dérive ou tout dérapage de la qualité

Merci/Thank you!

saeid.molladavoudi@Canada.ca