Journée Jeunes Chercheuses/Chercheurs de FARE

Ateliers (?)

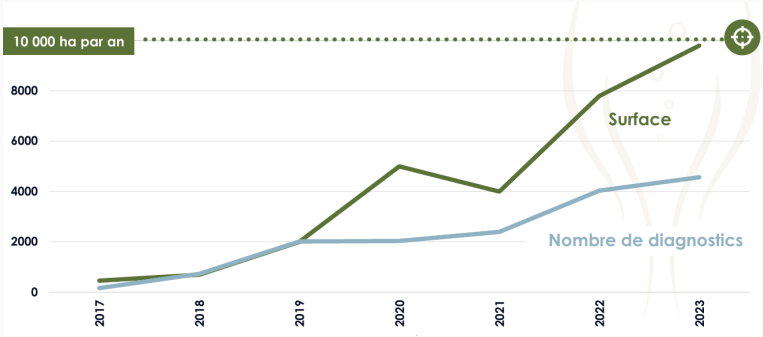
Alban Goupil

20 janvier 2025

Projet DASY

Problématique

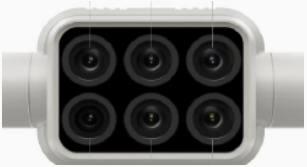
- Surveillance à grande échelle de l'évolution des jaunisses
- Flavescence Dorée
  - Propagation rapide
  - Traitement = arrachage
- Détection manuelle = coût / temps



DASY : Détection Automatisée des SYmptômes de jaunisse

Objectifs

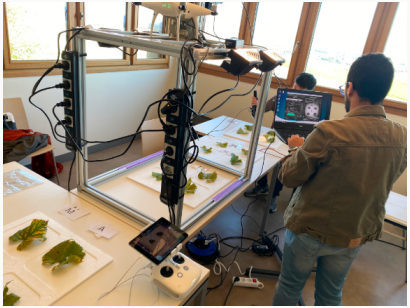
- Extraction de bandes spectrales discriminantes
  - Méthodes à développer
  - Spécification de caméras multispectrales
  - Acquisition à large échelle
- Systèmes embarqués
  - Transport sur nacelle / drone
  - Mécatronique / robotique

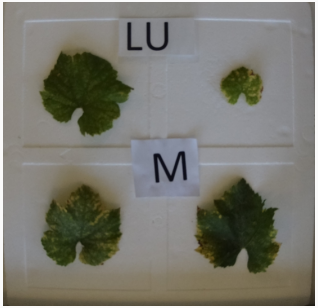


Consortium

- Comité Champagne / CIVC
- SEGULA Technologies
- CReSTIC

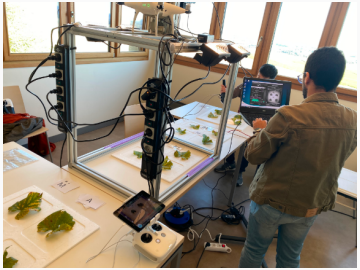
Acquisition des spectres





- Spectres Vis + NIR allant de 350nm à 1350nm
- 2 spectres / feuilles
- Acquisitions de 2019 à 2024
- ≈ 150 ceps sur 5 zones avec 5 classes en 2022

Acquisition des images



- Conditions contrôlées + hauteur nacelle + hauteur drone
- Données ci-dessous selon campagne de mesures 2022
- ≈ 2400 images de feuilles Vis-NIR + SWIR soit 31 000 images
- ≈ 150 images de ceps Vis-NIR sur terrain $\times 2$ prises $\times 3$ luminosités soit 3 600 images
- Acquisition drone artisanale Vis-NIR (5 bandes)

Prise en main des données

Étapes de la prise en main

- Base de données : spectres de 2020
- Lecture des données en format CSV compressé
- Taille de la base # observations / # variables
- Visualisation / réduction de dimension
- Prétraitements

- Site : `https://github.com/alban-goupil/jc-fare-2025`
- Notebook pour la prise en main

Premiers tests avec ML sur étagère

Classifications

Classificateurs utilisés

- Support Vector Classifier (SVC)
- Random Forest (RF)
- Linear Discriminant Analysis (LDA)

Notebook sur la classification

Qu’est-ce qu’un modèle ?

Point de vue bayésien

- L'équation de base

$$\text{Modèle} = \text{Données} + \text{Préjugés}$$

- Format mathématique = Bayes

$$\underbrace{p(\text{Modèle} \mid \text{Données})}_{\text{a posteriori}} = \underbrace{p(\text{Données} \mid \text{Modèle})}_{\text{vraisemblance}} \times \underbrace{p(\text{Modèle})}_{\text{a priori}}$$

- Modèle : architecture et paramètres
  - Modèle linéaire : $y = ax + b$; architecture : équation droite, paramètres : a, b
  - Réseaux de neurones : $y = f(x, w)$; architecture : réseaux de n couches avec m entrées, etc ; paramètres : w
  - GPT-3 : token $= g(\text{token}, W)$; architecture : transformers et MLP ; paramètres : W =175 milliards de nombres

Modèle $\approx$ fonction indicatrice

Approche probabiliste / énergétique

- $p(x, y)$: probabilité de compatibilité entre x et y
- $p(x, y)$ est une fonction à trouver en fonction du discours
- $p(x, y) \approx 1 \Longrightarrow x$ et y compatibles / accord
- $p(x, y) \approx 0 \Longrightarrow x$ et y incompatibles / désaccord
- Objectivisme (Fisher) / subjectivisme (Jaynes)

Exemples

- $p($🦊, renard$) = 90\%$
- $p($🐼, renard$) = 20\%$
- $p($"Il était une", "fois"$) = 80\%$
- $p($"Le plus grand groupe de rock est", "Rolling Stones"$) = 0.1\%$

Utilisation d'un modèle

$$p(x, y) = p(x \mid y) \times p(y) = p(y \mid x) \times p(x)$$

Modèle prédictif, régression

- $\arg\max_y p(y \mid x)$
- $\arg\max_y p(y \mid$ 🐼$) =$ panda
- $p($renard$|$🦊$) = 99\%$
- $p($panda$|$🦊$) = 1\%$

Modèle génératif

- $\arg\max_x p(x \mid y)$
- $p($"fois" $\mid$ "Il était une"$) = 60\%$
- $p($"Rolling Stones" $\mid$ "Le meilleur groupe est"$) = 49\%$
- $p($"Beattles" $\mid$ "Le meilleur groupe est"$) = 49\%$

Apprentissage

$$\begin{aligned}
\text{Modèle}^\star &= \underset{\text{Modèle}}{\arg\min} \quad p(\text{Modèle} \mid \text{Données}) \\
&= \underset{\text{Modèle}}{\arg\min} \quad \underbrace{p(\text{Données} \mid \text{Modèle})}_{\text{Attachement aux données}} \times \underbrace{p(\text{Modèle})}_{\text{Architecture + régularisation}}
\end{aligned}$$

Exemple

Générateur de texte

- Basé sur fréquences des n-grammes
- Apprentissage sur corpus de 3 livres français
- Architecture très simple
- Exemples digrammes $p(\text{es}) \approx 3.0\%$, $p(\text{le}) \approx 2.2\%$, $p(\text{qu}) \approx 1.1\%$, $p(\text{lz}) \approx 0\%$,

Liens externes

- Notebook sur un générateur de texte
- Dasher : un modèle prédictif pour l'écriture rapide

Impact de la dimension

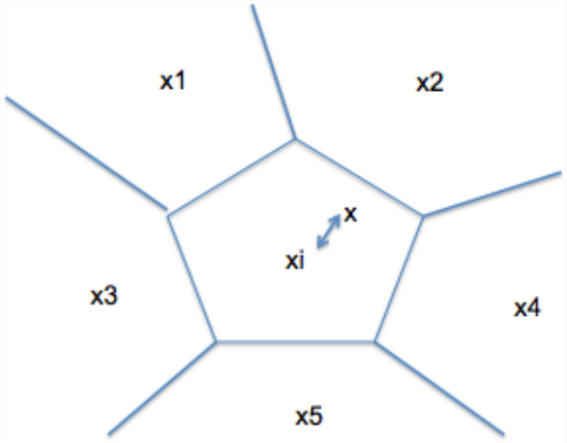
Généralisation et régularité

Problématique

- Données d'entraînement
  $\Longrightarrow$ toutes les données / entrées
- Mesure de généralisation ?
  $\Longrightarrow$ données test

Solution

- Nécessité de régularité
- Échantillon représentatif
  $\Longrightarrow$ recouvrement / pavage de l'espace des entrées possibles
- Le hasard fait bien les choses



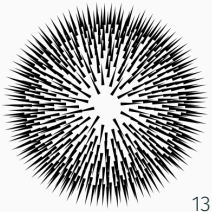
Malédiction de la dimension

Pavage

- Un point “couvre” un rayon de $\epsilon = 0.1$
- Il faut ≈ 10 points pour couvrir le segment de longueur 1
- Il faut ≈ 100 points pour couvrir le carré de côté 1
- Il faut ≈ 1000 points pour couvrir le cube de côté 1
- Il faut plus de $\epsilon^{-d} \left[\frac{d}{2\pi e}\right]^{d/2}$ points dans un hypercube de côté 1
- Réponse : concentration de la mesure

Autres phénomènes de la dimension (notebook illustratif)

- Tout l’espace est loin des données
  - “Dans l’(hyper)-espace personne ne vous entend crier.”
- Toutes les directions sont orthogonales
- Durcissement des sphères / toute la masse est à l’équateur



Impact du nombre d’observations

Bayes pour expliquer l'impact

Bayes cas dans le cas des données i.i.d

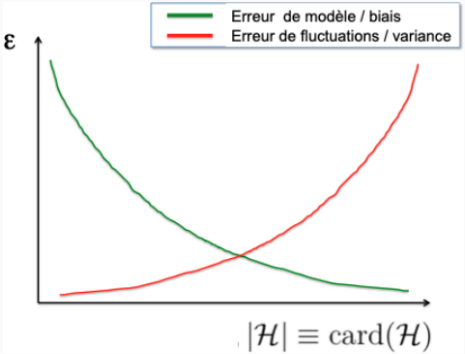
$$\log p(M \mid D) = \log p(M) + \sum_{i=1}^{n} \log p(D_i|M)$$

- M : modèle, D_i : i-ème données sur n
- Si $n \approx 0$ alors l'*a priori* est prédominant
- Si $n \to \infty$ alors l'*a priori* devient négligeable
  - Succi, Coveney, 2019, Big data : the end of the scientific method ?
- Notebook illustratif

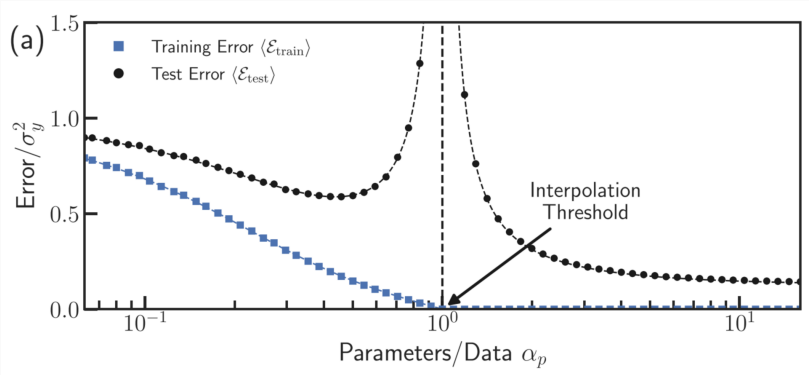
Dilemme Biais-Variance

$$R(f_I) \leq \tilde{R}(\tilde{f}) \leq R(f_I) + 2 \max_{h \in \mathcal{H}} |R(h) - R(\tilde{h})|$$

- $R(.)$: Erreur de généralisation
- $\mathcal{H}$: ensemble des estimateurs possibles
- f_I : meilleur estimateur avec toutes les données possibles et inimaginables
- $\tilde{f}$: meilleur estimateur avec les données



Sur-apprentissage et double descente

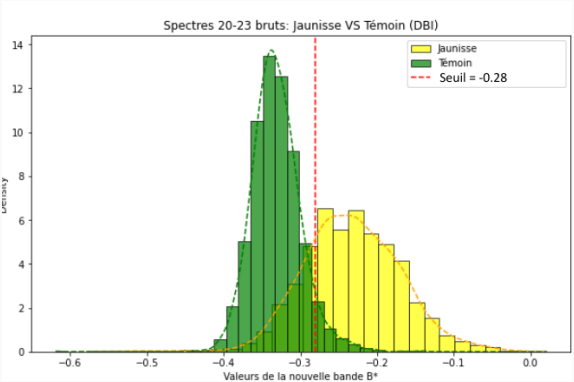


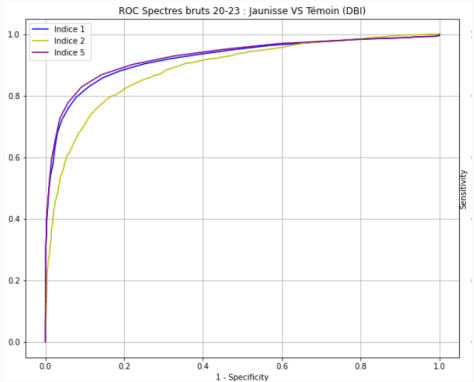
Sur-apprentissage

- # paramètres $\gg$ # données
- Mémorisation des données d'entraînement
- Généralisation remise en cause
- Nécessité de régularisation

Sélection de variables

Création d'un indice de jaunissité

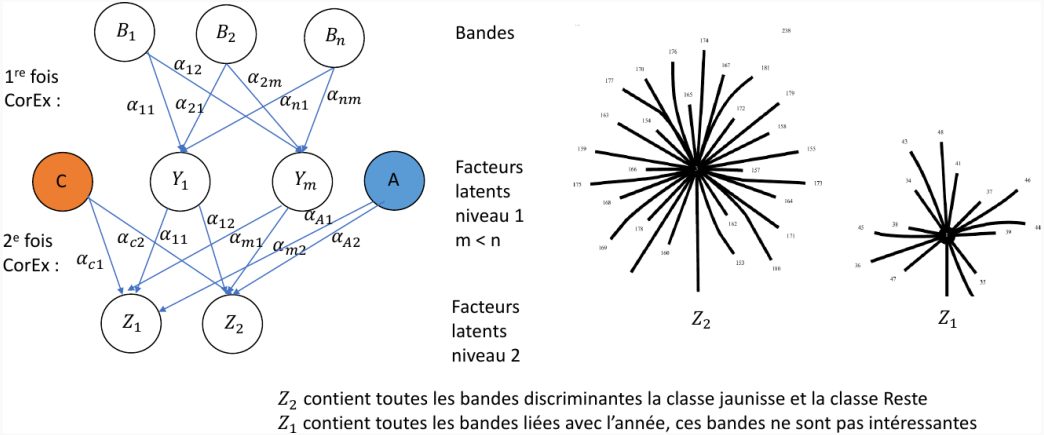




- Indice 1 = 3 bandes (2 Vis + 1 NIR)
- Indice 2 = 3 bandes (2 Vis + 1 NIR)
- Indice 5 = 4 bandes (2 Vis + 2 NIR)

Exemple de méthode de sélection

- Utilisation du CorEx pour regrouper les variables
- Trouver les liaisons classes / bandes
- Mais bandes agnostiques aux années



Z_2 contient toutes les bandes discriminantes la classe jaunisse et la classe Reste
Z_1 contient toutes les bandes liées avec l'année, ces bandes ne sont pas intéressantes

Pour finir

Conclusion

- Outils de l'apprentissage automatique = puissants + simples à manipuler
- Véritables points critiques
  1. Expertise des opérateurs / a priori / préjugé
  2. Qualité des données
  3. Critique des résultats
  4. Choix des modèles et de la méthode d'apprentissage
  5. Connaissances des limites des outils
- Aide d'expertes en science des données = points 4 et 5

  $\Longrightarrow$ Bouclage entre experts du domaine et du ML