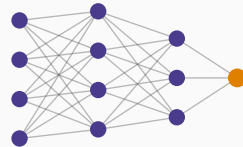


Apprentissage profond

Chapitre 12 : Ce que l'apprentissage profond ne résout pas



Les données d'abord

Trois limites de fond

Comment choisir

Bilan du cours

Les données d'abord

Le rapport paramètres / observations

Un réseau apprend ses **représentations**. C'est son avantage sur les méthodes qui exigent des variables construites à la main — et cela coûte beaucoup d'observations.

Contexte	Observations	Verdict
Vision, texte	millions	l'apprentissage profond domine
Tabulaire, quelques milliers de lignes	milliers	le boosting domine
Série financière quotidienne	quelques milliers	modèles économétriques compétitifs

Application en gestion

Sur données **tabulaires**, le boosting de gradient du chapitre 10 d'*Apprentissage statistique* bat régulièrement les réseaux de neurones.

Les raisons tenues pour principales : variables hétérogènes gérées nativement, aucune normalisation requise, interactions captées sans architecture à concevoir, et robustesse aux réglages approximatifs.

Ce cours ne conclut donc pas à la supériorité des réseaux. Il délimite le domaine où ils apportent quelque chose — et ce domaine n'est pas celui de la plupart des données d'un économiste.

Trois limites de fond

Ce qu'un réseau apprend

Des **associations** entre entrées et sortie, dans la distribution qui a produit les données d'apprentissage.

Rien de plus. Aucune architecture, aucune profondeur ne transforme une association en effet.

La conséquence pratique

Un réseau ne dit pas ce qui se passerait si l'on **intervenait**. Il ne prévoit correctement que tant que la distribution reste la même.

Un modèle de crédit entraîné avant un changement de politique d'octroi devient faux après, non parce qu'il a mal appris, mais parce que **le monde qu'il a appris n'existe plus**.

C'est la critique de Lucas, formulée en apprentissage automatique.

Trois formes

1. **Décalage de distribution** — les performances s'effondrent hors du domaine d'entraînement.
2. **Exemples adverses** — une perturbation imperceptible change la prédiction.
3. **Confiance mal calibrée** — le réseau se trompe avec assurance.

La calibration, souvent oubliée

Un réseau qui annonce 0,9 de probabilité se trompe plus souvent que dans un cas sur dix : il est **sur-confiant**.

En crédit ou en assurance, une probabilité mal calibrée est un problème sérieux : elle fausse tous les calculs de provisionnement en aval. Des méthodes de recalibration existent — elles doivent être appliquées avant tout usage décisionnel.

Ce qui reste possible

Les méthodes du chapitre 12 d'*Apprentissage statistique* — importance par permutation, Shapley — s'appliquent aux réseaux.

Elles expliquent le **modèle**, jamais le **monde**. Et un modèle à millions de paramètres reste, dans son fonctionnement, hors de portée d'une explication complète.

Quand cela devient rédhibitoire

Un refus de crédit doit être motivé auprès du client et du superviseur. Un modèle non explicable peut être **inutilisable**, quelle que soit sa performance.

C'est une contrainte réglementaire, non technique, et elle explique la persistance des scores logistiques dans les banques.

Comment choisir

Une règle de décision

Le tableau à retenir

Situation	Méthode
Tabulaire, milliers de lignes	Boosting
Image, texte, son	Réseaux profonds
Série unique, quotidienne	ARMA, GARCH
Beaucoup de séries, haute fréquence	Réseaux
Décision à motiver	Modèle linéaire ou arbre
Question causale	Ni l'un ni l'autre — voir <i>Inférence causale</i>

La démarche saine

1. Commencer par une **référence simple**.
2. Passer au **boosting** si le tabulaire domine.
3. N'aller aux réseaux que si la nature des données ou leur volume le justifient.
4. À chaque étape, vérifier que le gain **paie** sa complexité.

Bilan du cours

Ce que vous savez faire

Chapitres	Compétence
1-2	Comprendre un neurone, compter les paramètres
3-5	Choisir perte et activation, entraîner, diagnostiquer un gradient
6-7	Régulariser, initialiser, normaliser
8-10	Convolutif, récurrent, attention
11-12	Appliquer aux séries, et savoir quand ne pas le faire

Exemple

Chapitre	Élément	Valeur
2	Paramètres du réseau 20-32-16-1	1 217
3	Dérivée maximale de la sigmoïde	0,25
3	Gradient après 10 couches	$9,5 \times 10^{-7}$
4	Descente sur w^2 : $\eta = 0,1$ / $\eta = 1,1$	converge / diverge
8	Couche dense contre convolution	1 000 100 contre 320

Intuition

Les architectures changeront. Trois choses resteront :

1. **Un réseau est une composition de modèles linéaires et de non-linéarités**, entraînée par descente de gradient. Tout le reste en découle.
2. **Le volume de données commande le choix de méthode**, davantage que la nature du problème.
3. **Prévoir n'est pas expliquer**, et aucune architecture ne comblera cet écart.

Le troisième point est celui que le cours suivant prend en charge.