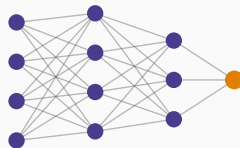


Apprentissage profond

Chapitre 6 : Régularisation : décrochage, pénalisation, arrêt précoce



Le problème, posé par les chiffres

L'arrêt précoce

La pénalisation des poids

Le décrochage

Les autres leviers

Ce qu'il faut retenir

Le problème, posé par les chiffres

Trop de paramètres pour trop peu de données

Ce que le chapitre 2 a montré

Notre réseau modeste compte **1 217 paramètres** pour vingt variables d'entrée.

Avec mille observations, il y a davantage de paramètres que de données. Le chapitre 5 d'*Apprentissage statistique* l'a établi : dans ce régime, l'ajustement parfait est **automatique** et sans valeur.

Le diagnostic

Tracer les deux courbes de perte, en apprentissage et en validation, à chaque époque.

Configuration	Diagnostic
Les deux baissent ensemble	apprentissage sain
Apprentissage baisse, validation remonte	sur-ajustement
Les deux stagnent haut	sous-ajustement, ou taux inadapté

Le point où la validation remonte est le point d'arrêt. Tout ce chapitre découle de cette observation.

L'arrêt précoce

Arrêt précoce

Surveiller la perte de validation à chaque époque, conserver les poids du **meilleur** point, et cesser l'entraînement lorsqu'elle ne s'améliore plus pendant un nombre fixé d'époques — la **patience**.

Pourquoi cela régularise

Les poids partent de valeurs petites et grandissent au fil de l'entraînement.

Arrêter tôt, c'est **limiter leur amplitude** — donc faire, indirectement, ce que fait une pénalité L_2 . La différence est qu'aucun coefficient n'est à régler.

Le réflexe à installer

L'arrêt précoce est **gratuit** : il n'exige aucun réglage, réduit le temps de calcul, et évite le sur-ajustement.

Il doit être activé par défaut sur tout entraînement, avant même d'envisager les autres techniques.

La pénalisation des poids

Le chapitre 4 d'Apprentissage statistique, transposé

Décroissance des poids

On ajoute à la perte une pénalité sur l'amplitude des paramètres :

$$L_{\text{total}} = L + \lambda \sum_j w_j^2$$

C'est exactement la régression **ridge**, appliquée à un réseau. On l'appelle ici *weight decay*.

Ce que cela produit

Des poids **petits** signifient des sorties peu sensibles à de petites variations des entrées — donc une fonction plus lisse, moins capable d'épouser le bruit.

λ se choisit par validation, comme toujours. Les valeurs usuelles vont de 10^{-5} à 10^{-3} .

Une précaution

On ne pénalise **pas les biais**, pour la même raison qu'on ne pénalisait pas la constante en régression : ils ne mesurent aucune sensibilité, et les contraindre déplace inutilement le modèle.

Le décrochage

Décrochage

À chaque passage, **désactiver au hasard** une fraction p des neurones d'une couche — typiquement $p = 0,2$ à $0,5$.

À l'inférence, tous les neurones sont actifs, et les sorties sont mises à l'échelle pour compenser.

Deux façons de comprendre pourquoi cela marche

Interdire la co-adaptation — un neurone ne peut plus compter sur la présence d'un voisin particulier, puisque celui-ci peut disparaître. Chacun doit donc porter une information utile par lui-même.

Un ensemble implicite — chaque passage entraîne un sous-réseau différent, et l'inférence les moyenne tous.

Cette seconde lecture relie directement le décrochage au **bagging** du chapitre 9 d'*Apprentissage statistique* : moyenner des modèles différents réduit la variance. Ici, l'ensemble est obtenu sans coût.

Application en gestion

Le décrochage se place sur les couches **entièrement connectées**, jamais sur la couche de sortie.

Il est peu employé dans les réseaux convolutifs modernes, où la normalisation par lot du chapitre 7 remplit un rôle voisin, et où les deux ensemble se gênent parfois.

Les autres leviers

Trois compléments

L'augmentation des données

Créer artificiellement des observations nouvelles par transformations qui **préservent l'étiquette** : rotations et recadrages pour une image, décalages temporels pour une série.

C'est la régularisation la plus efficace de toutes, parce qu'elle attaque la cause — le manque de données — et non le symptôme.

Réduire le modèle

Un réseau plus petit ne peut pas mémoriser. C'est une régularisation directe, souvent négligée par réflexe d'ambition.

Devant un sur-ajustement, **essayer un réseau plus petit avant d'empiler les techniques** est une démarche saine.

Le bruit

Ajouter un léger bruit aux entrées ou aux poids agit comme une régularisation, en interdisant au modèle de s'appuyer sur des détails précis.

En quel ordre procéder

1. **Arrêt précoce** — toujours, gratuitement.

Ce qu'il faut retenir

Cinq points

1. Le **diagnostic** se lit sur les deux courbes de perte : la validation qui remonte signale le sur-ajustement.
2. L'**arrêt précoce** est gratuit et doit être systématique.
3. Le **weight decay** est la régression ridge transposée; on ne pénalise pas les biais.
4. Le **décrochage** empêche la co-adaptation et constitue un ensemble implicite.
5. L'**augmentation des données** est la plus efficace, car elle traite la cause.

La suite

Ces techniques supposent que l'entraînement démarre correctement. Le chapitre 7 traite des deux conditions qui le permettent : initialisation et normalisation.