

QCM - Méthodes d'Optimisation

Questions

1. Quelle est la principale différence entre la descente de gradient (GD) et la descente de gradient stochastique (SGD) ?
 - a) GD utilise un seul échantillon à chaque itération, tandis que SGD utilise l'ensemble des données.
 - b) GD utilise l'ensemble des données à chaque itération, tandis que SGD utilise un échantillon unique.
 - c) GD converge toujours plus rapidement que SGD.
 - d) SGD est toujours plus précis que GD.
2. Pourquoi la descente de gradient stochastique est-elle souvent plus rapide que la descente de gradient standard ?
 - a) Elle utilise la totalité des données pour chaque itération.
 - b) Elle met à jour les paramètres plus fréquemment en utilisant des échantillons individuels.
 - c) Elle utilise des approximations de la fonction de coût.
 - d) Elle calcule les dérivées seconde de la fonction de coût.
3. La descente de gradient par mini-lots (BGD) consiste à :
 - a) Utiliser un seul échantillon pour la mise à jour des paramètres.
 - b) Utiliser l'ensemble du jeu de données pour chaque mise à jour.
 - c) Utiliser un sous-ensemble des données (mini-lot) pour chaque mise à jour.
 - d) Calculer la dérivée seconde de la fonction de coût.
4. Quelle est la formule de mise à jour dans la descente de gradient standard ?
 - a) $\theta_{t+1} = \theta_t + \eta \nabla f(\theta_t)$
 - b) $\theta_{t+1} = \theta_t - \eta \nabla f(\theta_t)$
 - c) $\theta_{t+1} = \theta_t - \eta f(\theta_t)$

$$d) \theta_{t+1} = \theta_t + \eta f(\theta_t)$$

5. La méthode de Newton utilise :
 - a) Le gradient uniquement.
 - b) Le gradient et la matrice Hessienne.
 - c) La dérivée troisième.
 - d) Les différences finies.
6. Qu'est-ce que l'étape de mise à jour adaptative dans Adam ?
 - a) Utilisation d'un taux d'apprentissage fixe.
 - b) Adaptation du taux d'apprentissage pour chaque paramètre.
 - c) Calcul de la dérivée seconde.
 - d) Mise à jour de tous les paramètres simultanément.
7. Pourquoi la méthode de Newton peut-elle être plus rapide que la descente de gradient ?
 - a) Elle utilise des gradients stochastiques.
 - b) Elle converge toujours en une seule itération.
 - c) Elle utilise des informations de la courbure pour mieux ajuster les pas de mise à jour.
 - d) Elle ne nécessite pas de calcul du gradient.
8. Quelle est la principale limitation de la méthode de Newton ?
 - a) Le calcul du gradient est très coûteux.
 - b) Elle ne fonctionne que pour les fonctions linéaires.
 - c) Elle converge lentement pour les grandes dimensions.
 - d) Le calcul de la matrice Hessienne peut être coûteux et nécessiter une inversion.
9. Qu'est-ce que le gradient de Nesterov ?
 - a) Un gradient calculé avec une approximation de la dérivée seconde.
 - b) Une méthode de descente utilisant une correction de la vitesse pour anticiper la mise à jour.
 - c) Une forme d'approximation de la méthode de Newton.
 - d) Une technique d'optimisation des fonctions convexes seulement.
10. La descente de gradient avec moment (momentum) aide à :
 - a) Réduire le taux d'apprentissage.

- b) Accélérer la convergence en ajoutant un terme de vitesse basé sur les gradients passés.
 - c) Augmenter le bruit dans les mises à jour.
 - d) Calculer la dérivée seconde.
11. Comment la descente de gradient stochastique (SGD) régularisée affecte-t-elle le modèle ?
- a) Elle aide à éviter le surapprentissage en ajoutant une pénalisation sur les poids.
 - b) Elle augmente la taille du lot pour chaque itération.
 - c) Elle utilise une dérivée seconde pour la mise à jour.
 - d) Elle ne change pas le modèle.
12. Quel est le rôle du taux d'apprentissage (learning rate) ?
- a) Il contrôle le nombre de paramètres à mettre à jour.
 - b) Il contrôle la taille des pas lors de la mise à jour des paramètres.
 - c) Il détermine le nombre de gradients à calculer.
 - d) Il fixe la direction de la mise à jour.
13. La méthode RMSprop est conçue pour :
- a) Utiliser un taux d'apprentissage constant.
 - b) Adapter le taux d'apprentissage basé sur la moyenne des gradients récents.
 - c) Calculer des approximations de la dérivée seconde.
 - d) Utiliser uniquement la descente de gradient standard.
14. L'accélération de Nesterov est particulièrement utile lorsque :
- a) La fonction objectif est convexe.
 - b) Les gradients changent rapidement de direction.
 - c) Le taux d'apprentissage est fixe.
 - d) Le modèle est très complexe.
15. La descente de gradient adaptative (Adagrad) ajuste le taux d'apprentissage :
- a) En utilisant un facteur constant.
 - b) Basé sur les gradients moyens globaux.
 - c) Basé sur l'inverse de la somme des gradients au carré.
 - d) En fonction de la dérivée seconde.

16. Quelle est l'une des principales limitations de la descente de gradient standard (GD) ?
- a) Elle ne fonctionne que pour les fonctions convexes.
 - b) Elle converge toujours vers le minimum global.
 - c) Elle peut être lente pour les grandes quantités de données car elle utilise l'ensemble du jeu de données à chaque itération.
 - d) Elle ne nécessite pas de calcul du gradient.
17. Dans quelle situation l'utilisation de la descente de gradient par mini-lots (BGD) est-elle avantageuse ?
- a) Pour les petits jeux de données seulement.
 - b) Pour les grandes quantités de données afin d'équilibrer la fréquence de mise à jour et le bruit des gradients.
 - c) Lorsque les données sont uniformément distribuées.
 - d) Lorsque la fonction de coût est constante.
18. Qu'est-ce que la méthode de Newton requiert pour être applicable ?
- a) Que la fonction objectif soit deux fois différentiable.
 - b) Que la fonction soit linéaire.
 - c) Que le gradient soit toujours positif.
 - d) Que la dérivée seconde soit toujours nulle.
19. L'utilisation d'un terme de moment dans la descente de gradient permet :
- a) D'augmenter le taux d'apprentissage.
 - b) De lisser les oscillations en accumulant les gradients précédents.
 - c) De réduire la dimension du problème.
 - d) De calculer la matrice Hessienne.
20. Quel est le principal avantage d'Adam par rapport aux autres méthodes d'optimisation ?
- a) Il utilise un taux d'apprentissage fixe.
 - b) Il combine les avantages d'Adagrad et de RMSprop pour un ajustement adaptatif du taux d'apprentissage.
 - c) Il ne nécessite pas de gradient pour fonctionner.
 - d) Il utilise uniquement la descente de gradient standard.
21. Pourquoi est-il important de choisir un bon taux d'apprentissage dans les méthodes de descente de gradient ?

- a) Un taux d'apprentissage trop élevé peut entraîner des oscillations, tandis qu'un taux trop faible peut ralentir la convergence.
 - b) Un taux d'apprentissage élevé est toujours meilleur.
 - c) Un taux d'apprentissage faible garantit la convergence au minimum global.
 - d) Le taux d'apprentissage n'affecte pas la vitesse de convergence.
22. La descente de gradient avec moment (momentum) est similaire à :
- a) L'algorithme de Newton.
 - b) La méthode de Nesterov.
 - c) Un filtre passe-bas qui lisse les mises à jour en fonction des gradients passés.
 - d) Une technique de dérivée seconde.
23. Le préconditionnement dans la méthode de Newton consiste à :
- a) Modifier l'échelle des gradients pour accélérer la convergence.
 - b) Réduire le taux d'apprentissage.
 - c) Ajouter du bruit aux gradients.
 - d) Utiliser des gradients stochastiques.
24. Comment RMSprop diffère-t-il d'Adagrad ?
- a) Il utilise un taux d'apprentissage fixe.
 - b) Il intègre une moyenne mobile pour atténuer la somme des gradients au carré, ce qui permet d'éviter les problèmes d'Adagrad.
 - c) Il ne change pas le taux d'apprentissage.
 - d) Il est identique à Adagrad.
25. Que se passe-t-il si le taux d'apprentissage est trop élevé dans la descente de gradient ?
- a) La convergence sera plus rapide.
 - b) La fonction de coût peut osciller ou diverger.
 - c) Les gradients deviendront négatifs.
 - d) La méthode de Newton sera plus efficace.
26. Quelle méthode est généralement utilisée pour les problèmes d'optimisation de grande dimension ?
- a) La méthode de Newton.
 - b) La descente de gradient classique (GD).
 - c) La descente de gradient stochastique (SGD) ou par mini-lots (BGD).

- d) La méthode des moindres carrés.
27. Dans la méthode de Newton, que représente la matrice Hessienne ?
- a) La somme des gradients.
 - b) La dérivée première de la fonction de coût.
 - c) La matrice des dérivées secondes, représentant la courbure locale de la fonction.
 - d) Le gradient moyen.
28. Quelle méthode utilise un taux d'apprentissage qui diminue au fil des itérations ?
- a) Adagrad.
 - b) La descente de gradient classique (GD).
 - c) La méthode de Newton.
 - d) Adam avec un taux d'apprentissage fixe.
29. La méthode de Newton fonctionne bien pour les fonctions :
- a) Linéaires uniquement.
 - b) Différentiables avec des courbures significatives.
 - c) Sans gradient.
 - d) Où les gradients sont nuls.
30. La descente de gradient adaptative (Adagrad) est principalement utilisée pour :
- a) Les petits jeux de données.
 - b) Les fonctions sans courbure.
 - c) Les problèmes où les paramètres ont des fréquences de mise à jour très différentes.
 - d) Les fonctions linéaires seulement.