Robustesse de l'apprentissage machine KNN face aux données incomplètes

Noé Aubin-Cadot

1er mars 2020

But

But : Déterminer à quel point l'apprentissage machine KNN est robuste face à des données incomplètes.

Plan

Plan :

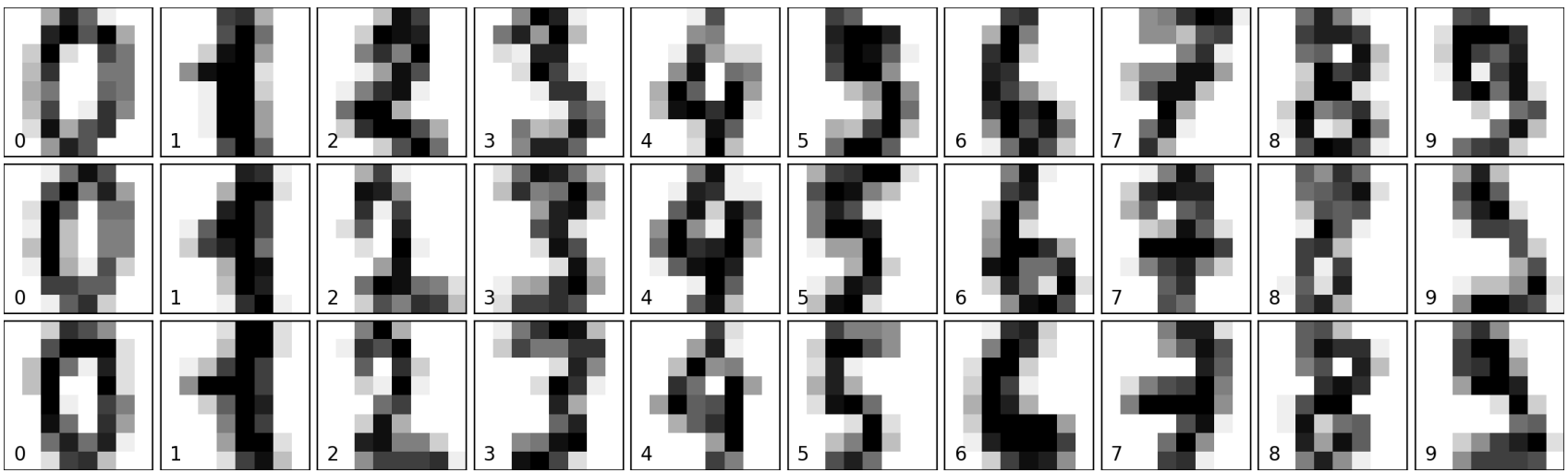
1. Trouver des données.
2. Perforer les données.
3. Remplir ou reconstruire les données perforées.
4. Apprentissage machine sur les données reconstruites.

Trouver des données

On considère les données `digits` de `Scikit-learn` :

- *source* $\mathbf{X}$ = images.
- *but* $\mathbf{y}$ = chiffres $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$.

Ces données contiennent 1797 images de 8×8 pixels monochromes de 4 bits, i.e. à valeurs en $\{0, 1, 2, ..., 15\}$.



Trouver des données

Les données $(\mathbf{X}, \mathbf{y})$ sont scindées en deux sous-ensembles :

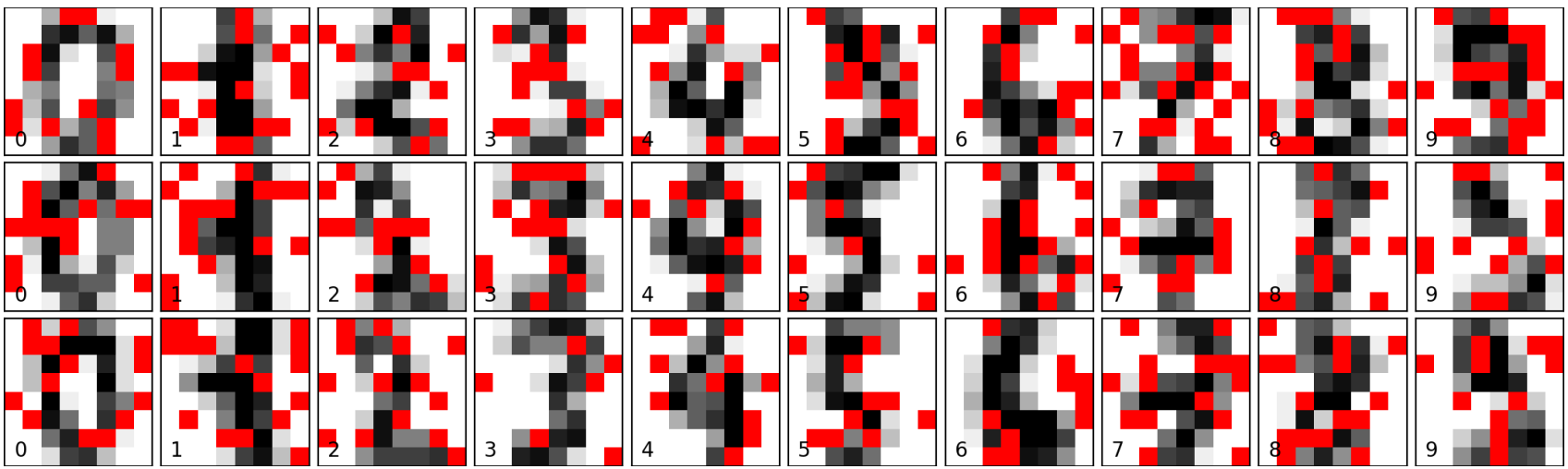
- 75% : *entraînement* $(\mathbf{X}_{\text{train}}, \mathbf{y}_{\text{train}})$, 1347 images.
- 25% : *test* $(\mathbf{X}_{\text{test}}, \mathbf{y}_{\text{test}})$, 450 images.

Les 1347 images de $\mathbf{X}_{\text{train}}$ sont constituées de 139 chiffres `0`, 145 chiffres `1`, 134 chiffres `2`, 137 chiffres `3`, 131 chiffres `4`, 133 chiffres `5`, 129 chiffres `6`, 141 chiffres `7`, 129 chiffres `8`, 129 chiffres `9`.

Perforer les données

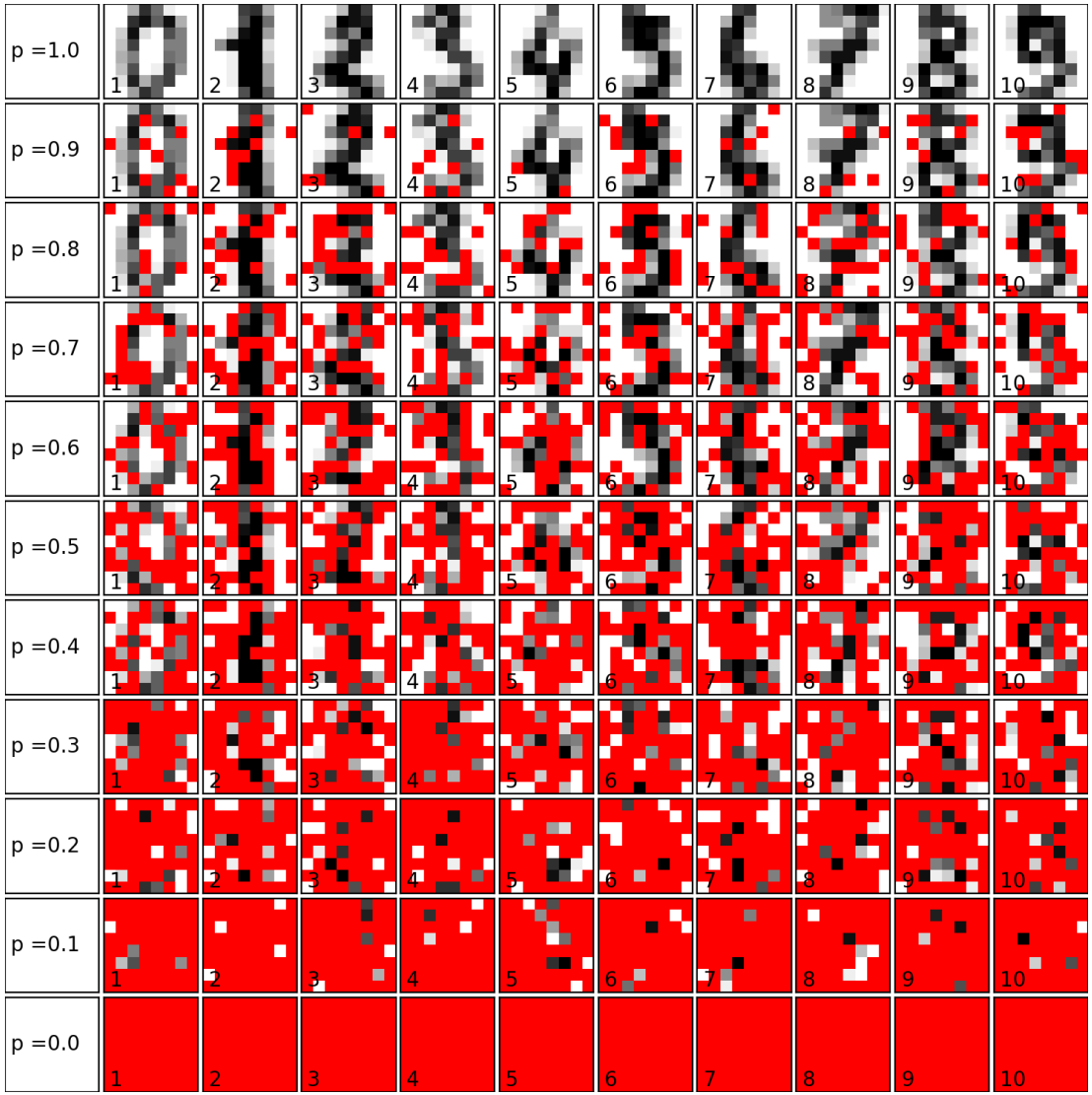
On perfore les images comme suit. La probabilité qu'un pixel soit présent est $p \in [0, 1]$. Le nombre de pixels présents dans une image est une variable aléatoire qui suit une loi binomiale $B(n = 64, p)$.

Images perforées (pixels absents en rouge) :



Perforer les données

En faisant varier p on a plus ou moins de pixels présents :



Perforer les données

Maintenant que les données perforées sont entre nos mains, deux options s'ouvrent à nous :

1. Faire de l'apprentissage machine sur les données perforées.
2. Remplir ou reconstruire les données perforées puis faire de l'apprentissage machine sur ces dernières données.

KNN sur les données perforées

Pour faire de l'apprentissage machine KNN ($k = 1$) sur les images perforées, on se donne une notion de distance entre images perforées comme suit.

On prend deux images perforées. On regarde les pixels communs aux deux images. On prend la distance euclidienne (au carré) d^2 sur ces pixels communs. On multiplie d^2 par $n = 64$ et on divise par le nombre de pixels communs pour faire comme si on avait 64 pixels communs. Puis on divise par le nombre de pixels communs pour donner plus de poids aux paires d'images ayant beaucoup de pixels en communs.

Ceci permet de faire du KNN sur des images perforées.

Remplissage et reconstruction d'images

Voici plusieurs manières de remplir les pixels manquants des images perforées :

1. Remplissage par des pixels blancs.
2. Remplissage par des pixels aléatoires.
3. Remplissage par des pixels moyens.
4. Autre méthode de reconstruction.

Remplissage et reconstruction d'images

Le remplissage par des pixels blancs consiste à mettre des pixels blancs là où il n'y a pas de pixel.

Le remplissage par des pixels aléatoires consiste à mettre des pixels de valeurs aléatoires là où il n'y a pas de pixel.

Le remplissage par des pixels moyens se fait comme suit. On choisit un chiffre. On prend les images perforées de $\mathbf{X}_{\text{train}}$ qui correspondent au chiffre. On prend la moyenne sur toutes ces images des pixels existants. On remplace chaque pixel manquant par un pixel moyenné du même lieu.

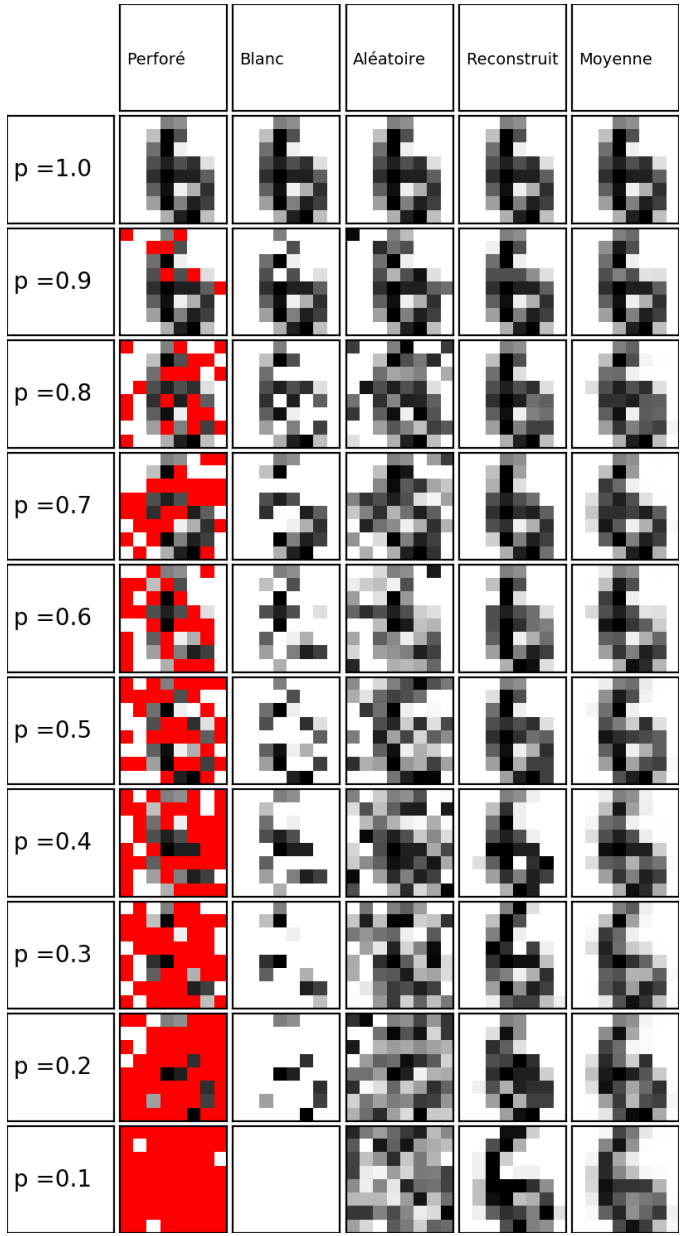
Remplissage et reconstruction d'images

Enfin, voici une méthode de reconstruction d'images. On prend une première image de $\mathbf{X}_{\text{train}}$. On trouve une seconde image de $\mathbf{X}_{\text{train}}$ du même chiffre qui est la plus près de la première image (pour la métrique entre images perforées définie plus haut) puis on transfert les pixels connus de la seconde image dans les pixels inconnus de la première image. On recommence le processus.

Visualisons maintenant chaque méthode de remplissage et de reconstruction.

Remplissage et reconstruction d'images

Voici une image du chiffre 6 subissant respectivement une perforation puis soit un remplissage blanc, soit un remplissage aléatoire, soit une reconstruction, soit un remplissage moyen, pour diverses probabilités p qu'un pixel soit correct.

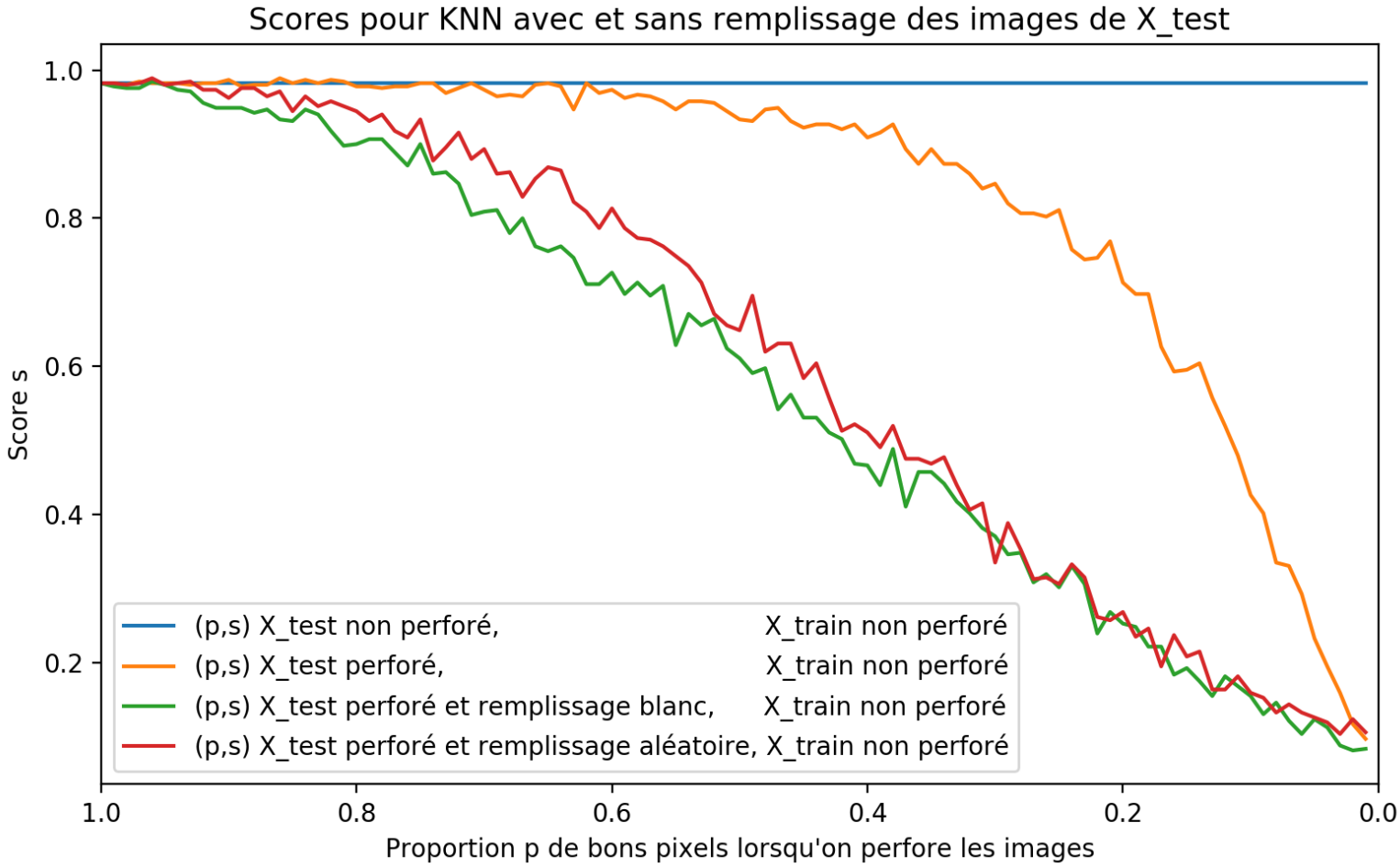


Scores d'apprentissage machine

On peut maintenant regarder les scores d'apprentissage machine KNN ($k = 1$) en fonction de p pour les diverses méthodes de remplissage et de reconstruction d'images perforées ci-haut.

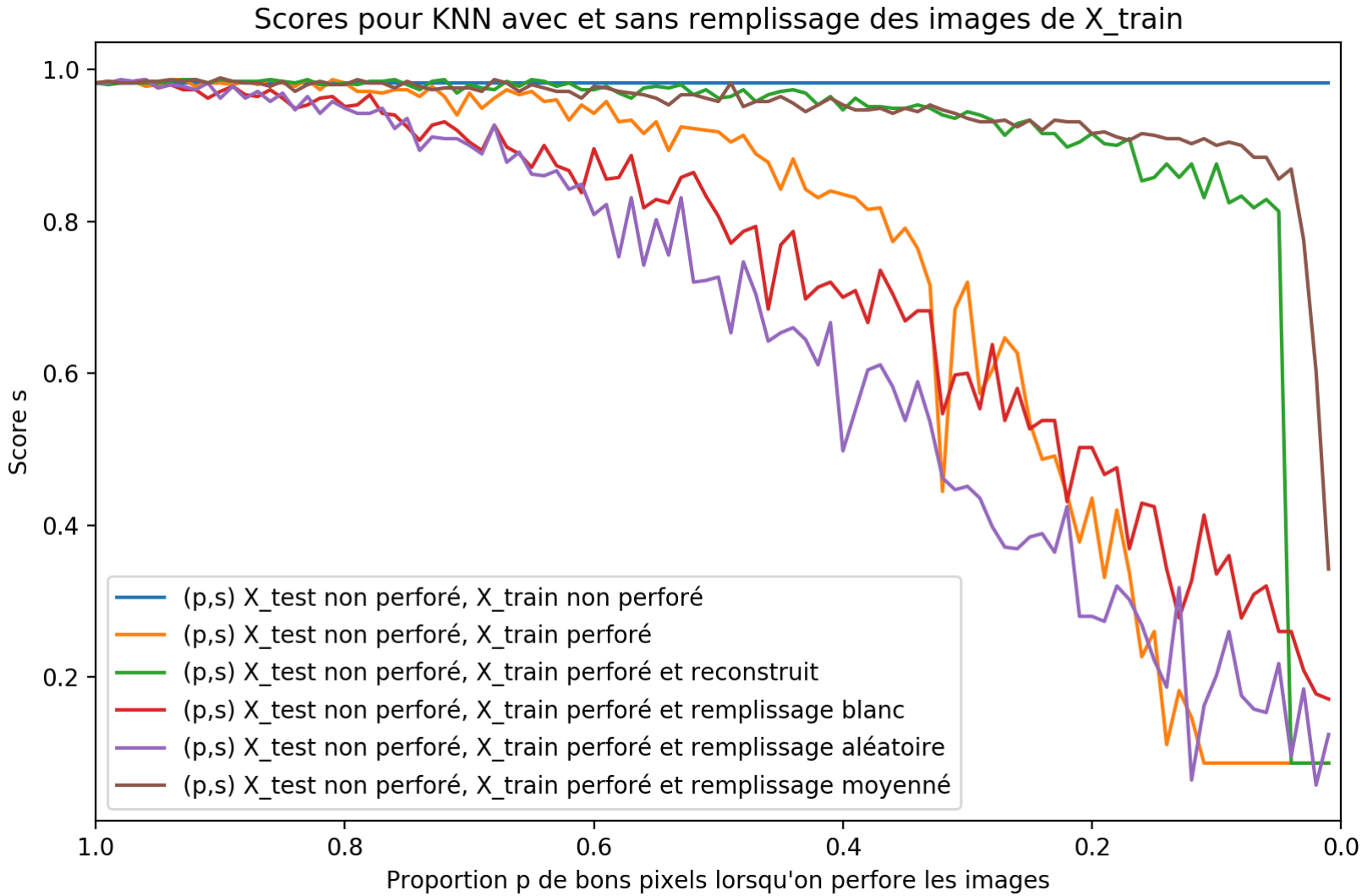
Scores d'apprentissage machine

L'apprentissage KNN sur les données perforées pour la métrique ci-haut est plus performante que l'apprentissage KNN sur les données remplies par du blanc ou de l'aléatoire.



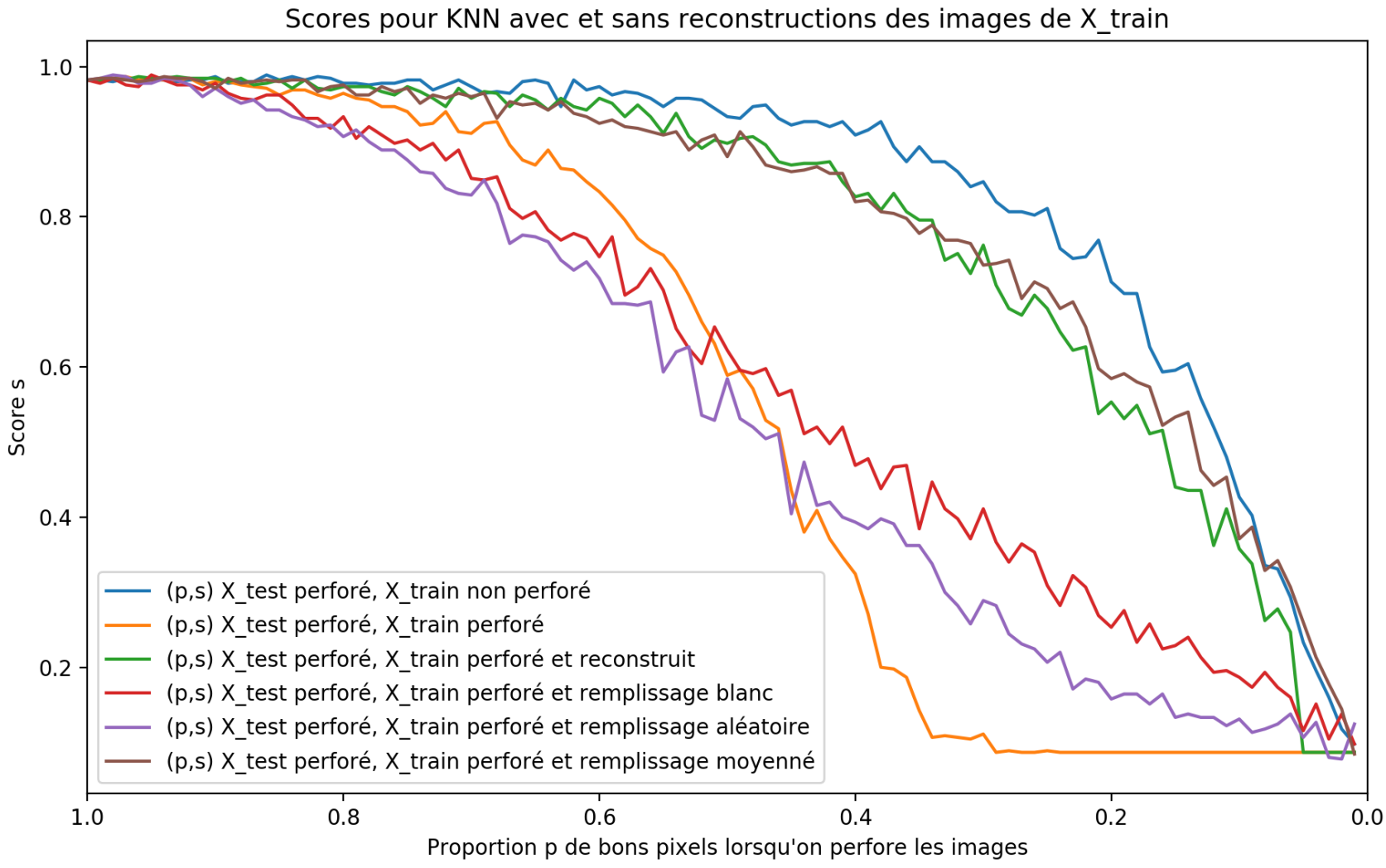
Scores d'apprentissage machine

Le remplissage moyenné est légèrement supérieur à la méthode de reconstruction d'images décrite ci-haut.



Scores d'apprentissage machine

Une dernière image.



Merci de votre attention ☺