Université de Strasbourg

07/05/2020

Projet – Apprentissage bout-en-bout sur des données

Analyse de données synthétiques par Réseau de Neurones Artificiel

Projet réalisé par

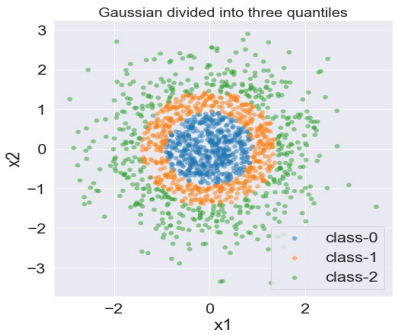
Mike Duran
Ilias Deligiannis

I. Introduction:

Dans le cadre de l'UE d'introduction à l'IA, durant la dernière année de formation en licence d'informatique, il nous est demandé de mettre en place un modèle prédictif basé sur une analyse de données par réseau de neurones. Dans un premier temps nous vous présenterons les données d'entraînement de notre modèle, puis nous expliquerons sont fonctionnement avant de finir par son analyse.

Note: Nous répondons aux questions posées dans le sujet au travers du rapport.

II. Données:



Visualisation des données.

Nous avons entraîné notre modèle sur des données générées par Sklearn et qui sont répertoriées dans un fichier csv. Ces données comportent 2 attributs distincts et sont linéairement non séparables en 3 classes. Ces données ne représentent aucun objet concret mais permettent une simulation de modèle pour pouvoir analyser son efficacité et ses performances.

III. Le modèle:

Au départ le model initialise les poids et les bias. Pour chaque époque il répète la même procédure. Pour chaque donnée d'entraînement il fait une propagation vers l'avant, obtient le cas d'erreur de celle ci et retro-propage cette erreur pour ajuster les poids et les bias. Une fois toutes les données traitées, il les propage dans le réseau et stoque les cas d'erreur. Une fois cette étape réalisée, les données d'entraînement sont mélangées pour empêcher l'overfitting et une époque ce termine. les erreurs stoquées seront affichées dans ANN.png pour chaque époque.

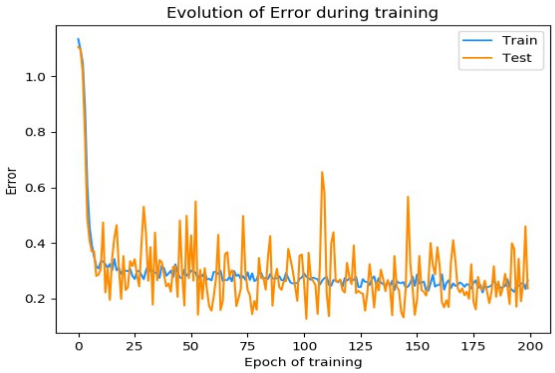
IV. Analyse:

1) Outils d'analyse:

Pour l'analyse du comportement de notre modèle nous nous sommes basés sur trois outils. *L'évolution de l'erreur au cours des époques,* le ratio *prédictions correctes / total*

prédictions et sur l'utilisation *d'une matrice de confusion* qui respectivement vérifie la cohérence de l'apprentissage, évalue la justesse, la précision puis la qualité de notre modèle de classification.

2) L'évolution de l'erreur au cours des époques:



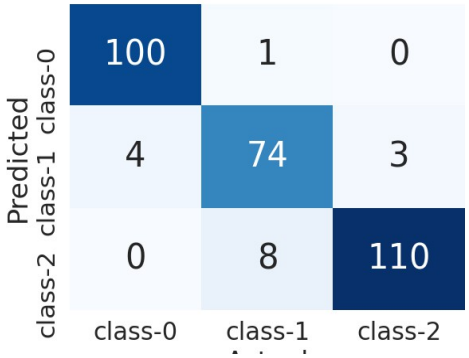
En observant la courbe générée durant l'entraînement de notre modèle on peut observer que la courbe d'erreur sur les données de test et celle d'entraînement semble évoluer dans la même direction. L'erreur diminue au fur et à mesure des époques. De part cette simple observation on peut se dire que l'entraînement de notre modèle se passe correctement et que celui-ci ne se retrouve pas bloquée. L'erreur semble acceptable. Si nous augmentons le taux d'apprentissage à 0.1, il se peut qu'on ais besoin de moins d'époque ou il se peut qu'on ais un ensemble sous-optimal de poids trop vite et donc un processus d'entraînement instable.

3) Ratio prédictions correctes / total prédictions:

En entraînant plusieurs fois le modèle avec ces même données on est arrivé à un ration moyen de 89%. Il s'agit là d'un score très satisfaisant. En effet un résultat trop proche d'un score parfait signifierait très probablement un modèle qui serait tombé dans de l'overfitting. Or ici nous avons un taux d'erreurs acceptable et qui sera confirmé par l'étude de la matrice de confusion.

4) Matrice de confusion:

La matrice de confusion est une métrique intéressante car elle nous permet de visualiser les performances d'un modèle puisque les bonnes prédictions sont sur la diagonale et les fausses hors de la diagonale.



Avec cette matrice de confusion on se rend compte que le modèle classifie la majeur partie du temps correctement. Le plus intéressant est qu'on se rend compte que les fausses classifications restent cohérentes avec la visualisation des données qui nous ont été fournies.
La classe-0 est mal estimée en tant que class-1. La classe-1 est mal estimée en étant considérée indistinctement comme classe-0 ou classe-2. Enfin la classe-2 est mal estimée en étant considérée comme classe-1.

V. Conclusion:

En effet en observant la visualisation des données on se rends compte que les 3 classes de données ont une disposition spatiale particulière. Elles forment une sorte de disque à 3 couches. La couche classe-0 est entourée par la classe-1 qui elle est entourée par la classe-2. Ainsi le fait que certaines instances de classe-0 peuvent être confondues avec des instances de classe-1 est envisageable car les deux classes sont voisines. C'est cette répartition spatiale particulière qui nous permet de dire que la matrice de confusion montre que le modèle est cohérent et qu'il a gardé une certaine souplesse.