Anonymisation des données et production de données synthétiques

problème soumis par Desjardins

Contexte

Les enjeux liés à la protection des renseignements personnels sont de plus en plus importants pour la société. L'apprentissage automatique requiert des données massives et granulaires et suscite donc des défis, en particulier la protection des renseignements personnels et le défi d'éviter les liens entre données et individus bien identifiés. Les applications d'apprentissage automatique posent des risques lors de la valorisation des données et peuvent entraver des travaux collaboratifs nécessitant le partage des données entre les collaborateurs (entreprises, fournisseurs, centres de recherche, chercheurs universitaires, etc.).

Objectif

Nous désirons étudier les meilleures pratiques d'anonymisation de données et de production de données synthétiques. Naturellement ces techniques doivent permettre de conserver un maximum de caractéristiques des données « originelles », afin de permettre le développement de bons modèles de prévision. Notre objectif est de permettre des collaborations de recherche (dans le cadre d'activités de valorisation des données) où les données ne seront pas exploitées à des fins indésirables.

Plus précisément, les données anonymisées ou synthétiques

(a) doivent être telles qu'on ne puisse les attribuer à des individus bien identifiés et

(b) permettront à des modèles explicatifs ou de prévision d'« apprendre » pour fournir des résultats aussi proches que possible de ceux des modèles équivalents apprenant à partir des données originelles.

Un autre aspect de ces méthodes est très important pour nous.

(c) Nous devrions être en mesure d'utiliser le travail d'un collaborateur nous soumettant le résultat d'une analyse descriptive basée sur certaines transformations de données. Ceci implique que nous puissions déterminer à quelles variables « originelles » font référence les variables transformées.

Données

Les premières expériences concerneront des tables de données structurées, par exemple des données socio-démographiques ou transactionnelles.

Applications

Les récipiendaires des jeux de données sécurisés sont susceptibles de les utiliser aussi bien en apprentissage supervisé que non supervisé. Un travail descriptif devrait pouvoir être effectué sur les jeux transformés.

Approches possibles

Pour anonymiser les données, on peut retirer les éléments permettant d'identifier les individus.

La production des données synthétiques peut utiliser les encodeurs automatiques variationnels (VAE en anglais) ou les réseaux contradictoires génératifs (GAN en anglais).

Cette liste d'approches n'est pas exhaustive.