

Problématique

Given anonymized information on thousands of photo albums, predict whether a human evaluator would mark them as 'good' <http://www.kaggle.com/c/PhotoQualityPrediction>

Description

Sur certains sites contenant des milliers de photos, les administrateurs souhaiteraient pouvoir faire ressortir quelles sont les photos ayant de fortes chances de plaire aux utilisateurs. Pour des raisons de rapidité et de bande passante, il est impossible d'analyser directement les photos, nous allons essayer de chercher un algorithme permettant de prévoir l'accueil des photos par les utilisateurs en fonction de ses méta-données.

Le résultat de notre algorithme doit être binaire, les utilisateurs vont aimer ou pas l'album.

Données

Voir donnees_photos.zip

Toutes ces photos contiennent 9 champs de méta-données :

- Latitude: latitude de l'album photo (sous forme entière)
- Longitude: longitude de l'album photo (sous forme entière)
- Width: largeur de l'image (pixels)
- Height: hauteur de l'image (pixels)
- Size: nombre d'images dans l'album
- Name: identifiant au nom de l'album
- Description: identifiant dans la description de l'album
- Caption: identifiant de toutes les photos de l'album
- Good: les utilisateurs aiment l'album (1) ou n'aiment pas (0)

Tous les mots importants des textes ont été codifié par un identifiant afin de garder l'anonymat. Cela permet d'avoir un algorithme dépendant pas d'une langue ou d'un encodage spécifique.

Évaluation

La pertinence des résultats que nous trouverons pourra être vérifié en passant par le site Kaggle.

Le site utilisera la méthode de déviance binomiale plafonnée (CBD).
(<https://www.kaggle.com/c/ChessRatings2/Details/Evaluation>)

Conseils

L'auteur indique qu'il y a une forte corrélation avec la localisation des photos. Par exemple, les photos prises dans la nature africaine sont souvent appréciées.

Ce sera à nous de déterminer d'autre corrélation pour avoir de meilleures prédictions