Mini-projet “scatter plot matrix” sur Altair

Première partie.

Vous devez construire une “scatter plot matrix” à partir de données sur des voitures. Nous avons préparé un notebook où le chargement des données est déjà spécifié (“TP1_MLAI_init”). Vous allez le compléter avec le code qui génère ce diagramme.

Une matrice de “scatter plot” est un diagramme qui contient plusieurs sous-diagrammes représentant des nouages de points, les diagrammes sont arrangés en forme de matrice.

Pour un ensemble de variables X_1, X_2, ... , X_k, la « matrice de nuages de points » montre tous les nuages de points par paires des variables sur une seule vue. Pour *k* variables, la matrice de nuages de points contiendra *k* lignes et *k* colonnes, à l'intersection de la i-ème ligne et de la j-ème colonne nous trouverons un diagramme qui trace les valeurs de X_i en fonction de X_j (un nuage de points à deux dimensions).

La diagonale d’une “scatter plot matrix”, tel que nous venons de la définir, contient une information triviale. En effet, la i-ème ligne et i-ème colonne trace la variable X_i vers elle même et présente donc, dans tous les cas, une ligne de pente 1. Nous vous proposons de construire un deuxième diagramme avec des histogrammes sur les variables X_i.

Nous vous proposons aussi d’ajuster une ligne à certains nouages de points 2D pour mieux comprendre la corrélation entre variables.

Pas à suivre :

1 – Télécharger (depuis https://github.com/uwdata/visualization-curriculum) ou copier sur votre ordinateur les deux notebooks suivants :

altair_marks_encoding
altair_data_transformation

Ces deux notebooks font partie du « Data Visualization Curriculum » proposé par Jeffrey Heer à l’Université de Washington (Seattle).

2 – Familiarisez-vous avec les concept de « encoding » et, en utilisant une marque circulaire, générez un « scatter plot » ou diagramme de nouage de points. Vous trouverez comment faire cela dans le notebook « altair_marks_encoding ».

3 – Utilisez un encodage couleur sur des groups de voitures, par exemple qui montre leur origine géographique.

4 - Familiarisez-vous avec le concept de « facet » qui vous sera utile pour générer la matrice de diagrammes. Vous pouvez télécharger le notebook « Multi-View Composition » si vous le souhaitez mais le notebook « altair_marks_encoding » contient une section « Column and Row Facets » qui devrait vous suffire pour cet exercice.

5 – Rendez le diagramme interactif et consultable avec le pointeur de la souris (« tooltip »).

6 – Un histogramme nécessite un type particulier de transformation de données. Consultez la section « Histograms » du notebook « altair_data_transformation ». Générez un diagramme qui contient les histogrammes de toutes les variables.

7 – Finalement, dessinez la ligne de régression pour certains nouages de points 2D de votre choix, elle correspond simplement à une transformation des diagrammes de points (transform_regression) qui doit être superposé au diagramme originale.

Deuxième partie.

Vous devez prendre plusieurs diagrammes d'Altair et montrer qu'ils ne se comportent pas correctement par rapport à un volume croissant de données.

Vous pouvez choisir les diagramme et la méthode. Nous vous conseillons de commencer par un simple diagramme type « nouage de points 2D » et le remplir en utilisant un mélange de gaussiennes (deux ou trois). Vous pouvez générer systématiquement de plus en plus échantillons à partir du mélange de gaussiennes et regarder son effet sur le(s) diagramme(s).

Finalement, vous allez proposer une solution basée sur l'estimation d'une densité de probabilité. Vous pouvez utiliser pour cela Scikit-Learn :

https://scikit-learn.org/stable/modules/density.html

Vous pouvez vous inspirer du exemple suivant :

https://towardsdatascience.com/simple-example-of-2d-density-plots-in-python-83b83b934f67

Remarquez que cet exemple utilise MatplotLib. Vous pouvez simplement le reproduire ou vous inspirer (à vous de choisir).

TP "BigData : projections et visualisation"

Suivre les instructions du Notebook « TP-module1-large-data-projections.ipynb ». Réaliser ses 3 exercices (à la fin du Notebook).

Deadline : 9 octobre 2023 (alejandro.ribes@edf.fr)