

Test préliminaire aux entretiens data science

Merci de répondre aux questions suivantes en justifiant votre solution. Au besoin, vous expliciterez clairement vos hypothèses et vous les discuterez.

**Question**  J’ai réussi à estimer un taux de transformation (nombre achats/nombre de visites) avec une précision de ±1% à l’aide de 15,000 visites. En d’autres termes l’intervalle de confiance à 95% du taux de transformation est :

De combien de visites ai-je besoin pour avoir une précision dix fois supérieure, à savoir ±0,1% ?

**Question**  Un test AB consiste à séparer aléatoirement les visiteurs en deux groupes (A et B), à présenter deux versions différentes du site aux groupes et à estimer l’incrément de divers indicateurs (par exemple le taux de transformation, le panier moyen, le CA etc.) entre les groupes. Un test AB donne les résultats suivants :

|  |  |  |
| --- | --- | --- |
|  | Groupe A | Groupe B |
| Nombre de visites | 117,415 | 117,284 |
| Taux de transformation observé | 7.07% | 9.36% |

Calculez l’intervalle de confiance à 95% du gain incrémental du taux de transformation observé,

**Question** Reprenons la question précédente. On a codé une procédure de tirage aléatoire dans laquelle un visiteur a la même probabilité de se retrouver dans le groupe A ou B. On a donc tiré échantillons, dont se sont retrouvés dans le groupe B. Calculez la probabilité d’observer cette valeur ou moins (expression analytique et application numérique). Quel est l’intérêt de ce calcul ?

**Question**  Une base de données SQL contient une table sessions qui, pour chaque session de navigation réalisée par un internaute, donne la séquence des (fiches) produits consultées. On va donc trouver les champs :

* session\_id : identifiant de la session
* product\_id : identifiant du produit consulté
* timestamp : date/heure de consultation du produit

Ecrivez une requête SQL qui, pour chaque paire de produits, donne le nombre de sessions au cours desquelles les deux produits ont été consultés : les champs de sortie de la requête seront product\_1, product\_2, nb\_sessions\_12. D’après vous, à quelle(s) application(s) peuvent servir les données issues de cette requête ?

**Question**  On dispose de 50 Go de données sur lesquelles on souhaite réaliser une régression linéaire, mais on ne dispose que de 8 Go de RAM sur notre machine. Citez plusieurs stratégies possibles pour parvenir à réaliser la régression.

**Question**  Choisissez une fonction qui fitte correctement les données du fichier regression\_data.csv. Donnez les paramètres de la fonction qui minimisent l’erreur de la régression, et l’erreur correspondante. Vous fournirez le code que vous avez développé pour les besoins de cette question, et des illustrations graphiques.

**Question**  Le fichier train.csv contient un échantillon de produits appartenant aux 20 plus grandes catégories du catalogue de Cdiscount. Pour chaque produit on a les champs :

* category\_id : identifiant de la catégorie du produit (entier entre 0 et 19)
* category : nom de la catégorie du produit
* title : titre du produit
* description : description du produit

Ecrivez un algorithme de classification supervisée qui prévoit la catégorie d’un produit à partir de son titre et de sa description (quelque chose de simple : pas la peine de construire un algorithme ultra optimisé). Vous mettrez de côté un ensemble de validation et donnerez le taux de classifications correctes sur ledit ensemble. Vous appliquerez l’algorithme sur les données de test contenues dans le fichier test.csv, et vous nous soumettrez un fichier predictions.csv avec les champs :

* id : identifiant du produit, tel que donné dans le fichier test.csv
* predicted\_category\_id : catégorie prédite par votre algorithme

N’oubliez pas de nous fournir le code développé pour les besoins de cette question.

**Question**  Le fichier purchases.csv contient la liste des achats réalisés le 1er septembre 2017. Les champs sont les suivants :

* time : date et heure d’achat du produit
* amount : prix en euros du produit acheté

On note le nombre d’heures écoulées de la journée et le prix d’achat, considérés comme des variables aléatoires dont le fichier ci-dessus donne un échantillon. Ecrivez un programme pour estimer la fonction de densité de probabilité conjointe, par la méthode d’estimation par noyau. Illustrez la fonction sous forme de carte de chaleur.