Cours

- # Détection d'anomalies
  - ## Approches
    - Détéction d'outliers
      - Apprendre à détecter des anomalies dans le jeu de donnée initial en cherchant des régions denses tout en ignorant les anomalies
    - Détection de nouveauté
      - Ici le jeu de donnée pas pollué par les annomalies
      - Il faut detecter des anomalies dans les données futures non observées
  - ## Méthodes
    - ### Non supervisées
      - Pas de labels fournis
      - Base d'apprentissage = Données normales + anomalies
      - Les anomalies sont très rares
      - #### Types
        - Approches basées sur le voisinages
          - Local Outlier Factor
            - Paramètres
              - $D_k(x)$ = Distance par rapport à son k^{ie} plus proche voisin
              - $N_k(x)$ = L'ensemble de ses k plus proche voisin
              - $R_k(x, y$ = Distance d'accessibilité de x par rapport à y comme étant le $max(d(x, y) et D_x(y))$
              - $AR_k(x)$ = Distance d'accessibilité moyenne de x comme étant égale à la moyenne des distances d'accessibilité de x avec tous les points de son voisinage ($N_k(x)$)
              - $f_k(x)$ = Densité d'accessibilité = Inverse de $AR_k(x)$
                - Une instance normale est sensée avoir une densité locale similaires à ses voisins, alors q'une instance anormale est sensée avoir une beaucoup plus petite densité locale
              - $LOF(x)$ = Moyenne du rapport $f_k(y)/f_k(x)$ pour tous les y dans $N_k(x)$
                - Mesure l'écart local d'un point par rapport à ses k voisins les plus proches
                - Si ce score est porche de 1, nous pouvons en conclure que l'observation est comparable à ses voisins
                - Si ce score est < 1, nous pouvons dire que l'observation se trouve dans une région dense
                - Dans les 2 cas, l'observation n'est pas considérée comme un outlier

- Un score est largement supérieur à 1 indique qu'on a à faire à un outlier
- Utilisée pour la détection de nouveauté ou d'outliers
- Méthode très puissante
- One class SVM : Pour la détection de nouveauté
  - Objectifs
- Isolation forest
  - Principe
    - Les anomalies sont rares et différentes ==> Elles sont donc susceptibles au mécanismes d'isolation
    - Construction d'ensemble d'arbres complement aléatoires : isolation tree
    - Chaque arbres est construit sur échantillon aléatoire des instances
    - Divisions opéré dans chaque nœud via un filtrage aléatoire d'une variable et
      - 1 seul
      - .....

Supervisées

- Labels à la fois pour les instances normales et anomalies
- Anomalies appartiennent à la classe rare
- Données déséquilibrées
- Adaptation des approches supervisées existantes

Types

- Random Under-sampling et Random Oversampling
  - Under-sampling
    - Sous échantillonnage
    - On diminue le nombre d'individus pour que les effectifs soient égaux
    - ==> Le classifieur risque d'apprendre dans un espace qui ne reflète pas la réalité ==> Il faut corriger les probabilité
  - Oversampling
    - Sur échantillonnage
    - On va dupliquer aléatoirement certains individus
    - ==> Le classifieur risque d'apprendre dans un espace qui ne reflète pas la réalité ==> Il faut corriger les probabilité
  - Balancing
    - Pondération des classes
    - On va utiliser ici un LogLoss pondéré pour calculer l'erreur de notre classifieur ==> On la veut à 0 ou proche
- SMOTE (Synthetic Minority Oversampling Technique)
  - Approche d'oversampling
  - Étapes
    - .....

- **Évaluations**
  - Accuracy déconseillé
  - Si on est intéressé par la classe en sortie
    - Balanced accuracy
      - Si on veut les classes positives et négatives
    - F1-score
      - Si on est intéressé par la classe positive
  - Si on est intéressé par les probabilités des classes en sortie
    - AUC-ROC
      - Si on est autant intéressé par les classe + que -
    - AUC-PR (Average Precision Score)
      - Calcul l'aire sous la courbe formée par les points .....
      - Si on est plus intéressé par la classe +
  - ==> Ne pas évaluer les modèles sur un échantillon équilibré
  - ==> Les anomalies sont souvent complètement nouvelle ==> Le modèle ne pourra pas détecter les nouvelles anomalies sur lesquelles il n'a pas été entraîné

Fouilles des données textuelles

- # Introduction
  - Processus d'extraction non triviale d'info utiles inconnues a priori à partir de grands volumes de textes
- # Préparation des données
  - ## Pré-traitement des données textuelles
    - Uniformisation du codage, élimination éventuelle de ceraines caractèe spéciaux
  - ## Extraction d'informations
    - Suppression des mot ignorés
  - ## Extraction d'entités primaires
  - ## Étiquetage grammaticale
  - ## Extraction d'entités nommées
    - Nom de personnes, lieux, organisation, dates qui ont un role important
- # Exploitation
  - ## Représentation vectorielle des textes
    - Elle prend les mot qui apparaisse le plus dans me texte mais ne prend pas en compte le contexte la grammaire et syntaxe ==> On perds beaucoup d'info
    - Comparaison des vecteurs avec la distance cosinus :
      - Le norme du vecteur étant proportionnelle à la longueur du texte

- On utilise parfois la similarité cosinus
- .....
- Évolutions de la représentation vectorielle de base
  - Pondération des termes : TF-IDF
    - Pondération les termes selon leur importance déterminé dans le texte
    - On utilise le TF (Term Frequency) dans le document
    - On multiplie le TD par IDF (Inverse Document Frequency)
      - C'est l'importance d'un terme pour tous les documents .....
      - Concept du LSA (=Latent Semantic Analysis)
        - .....
  - Sélection des termes :
  -

- **Développement de modèles**
- **Utilisation des modèles**

- **Challenges**
  - **Résolution référentielle**
  - **Analyse syntaxique (générale ou spécifique)**