Détermination des seuils pour classer le type d’activité physique

Introduction à l’IA, Aude Sportisse / Meggy Hayotte

Comment déterminer le type d'activité physique ?

- But: **déterminer des *seuils* sur des données d'accéléromètre (ou de capteurs VO2 etc) pour conclure sur le type de l'activité physique, par exemple en trois catégories: faible, modérée, intense.**

- **2 méthodes mentionnées ici:**

  - Méthode utilisant des **outils statistiques** (dont graphique: courbe ROC): c'est ce qui est utilisé dans le logiciel du TP.

Source: Eslinger, D. W., Rowlands, A. V., Hurst, T. L., Catt, M., Murray, P., & Eston, R. G. (2011). Validation of the GENEA Accelerometer.

  - Méthode utilisant des **algorithmes d'apprentissage statistique** (dont forêts aléatoires)

Source: Ahmadi, M. N., & Trost, S. G. (2022). Device-based measurement of physical activity in pre-schoolers: Comparison of machine learning and cut point methods. *Plos one*, *17*(4)

Contexte

GENEA: accéléromètre

- Avec l'accéléromètre, les données collectées sont du type:

(Temps, Accélération sur l'axe x, Accélération sur l'axe y, Accélération sur l'axe z)

Le temps est mesuré en secondes ($= s$), l'accélération en unité gravitationnelle ($g := m/s^{-2}$).

Déroulement de l'étude:

- 60 participants à l'étude
- Chaque participant a 3 accéléromètres GENEA: un sur chacun des poignets et un sur la taille.
- Chaque participant réalise une activité physique faible, moyenne et intense.
- Les données sont donc:

(Temps, Accélération sur l'axe x, Accélération sur l'axe y, Accélération sur l'axe z, Type d'activité)

Analyse des données (1): synthétiser l'information

- Données de l'accéléromètre recueillies sur des intervalles d'une minute et résumées en SVM_{gs} comme suit:

- Calcul de l'**amplitude** du vecteur de signal d'accélération

$$SVM_{gs} = \sum \left| \sqrt{x^2 + y^2 + z^2} - g \right|$$

On peut voir ce calcul comme une synthèse de l'information donnée par les accélérations des trois axes (x,y,z). Au lieu d'avoir trois nombres, on n'en a plus qu'un seul.

On retranche g pour soustraire le terme d'accélération à l'arrêt dû à la gravité.

- On a synthétiser les données comme:

(SVM_{gs}, Type d'activité)

Analyse des données (2)

- But: **déterminer des seuils** sur les données pour conclure sur le type de l'activité physique: faible, modérée, intense.
- Sur le jeu de données d'entrainement, nous avons accès au type d'activité physique: les **données sont donc étiquetées** et c'est de l'**apprentissage supervisé**.

Typiquement, on a envie de déterminer a et b tels que

$$SVM_{gs} < a \Rightarrow \text{activité faible}$$

$$a < SVM_{gs} < b \Rightarrow \text{activité modérée}$$

$$b < SVM_{gs} \Rightarrow \text{activité intense}$$

Analyse des données (3): courbe ROC

- Pour **déterminer des seuils**, on va utiliser la courbe ROC qui permet de dire si la méthode est discriminante ou non.

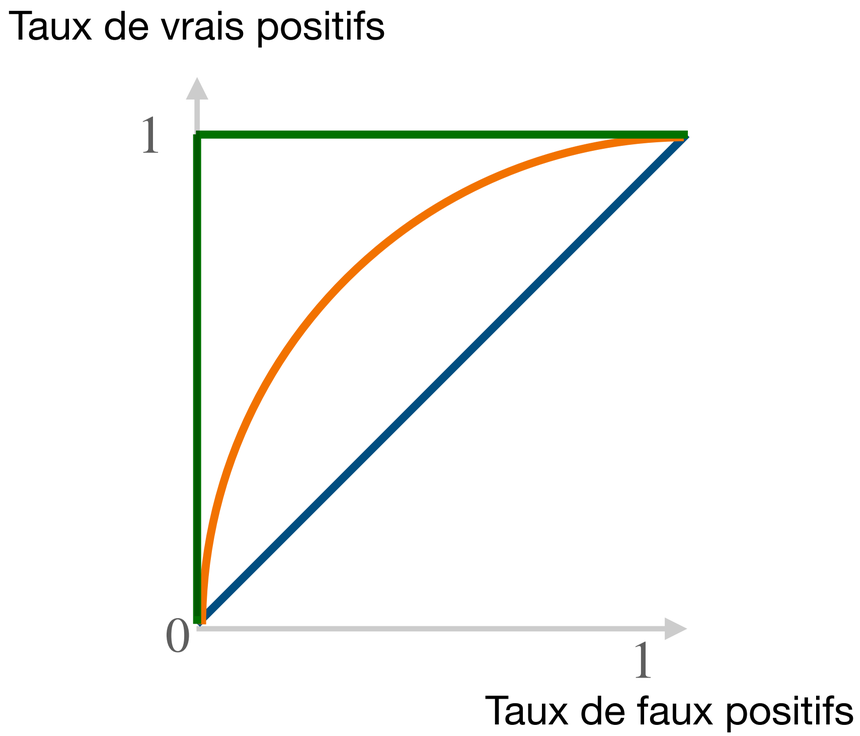
- On veut optimiser le choix des seuils pour classer le type d'activité physique le mieux possible. On veut **maximiser le taux de vrais positifs** et **minimiser le taux de faux positifs**.

  ✳ Vrai positif: l'activité physique a été classée faible alors qu'elle était vraiment faible.

  ✳ Faux positif: l'activité physique a été classée faible alors qu'elle ne l'était pas.

Analyse des données (3): courbe ROC

- But: **déterminer des seuils de discrimination** sur les données pour conclure sur le type de l'activité physique: faible, modérée, intense:
- Pour chaque type d'activité, on choisit les seuils pour **optimiser la courbe ROC**



Performance de la méthode:

- **Parfait: la méthode classe toujours bien l'activité physique**
- **Moyenne: la méthode fait mieux que l'aléatoire, mais n'est pas parfait.**
- **Aléatoire: la méthode fait pareil que classer de manière aléatoire l'activité physique**

Deuxième méthode: classification du type d'activité physique avec une forêt aléatoire

- Rappel de cours: un arbre de décision est un type d'algorithme d'apprentissage capable d'apprendre des **règles de décision** pour classifier les données. Une forêt aléatoire va combiner les résultats de plusieurs arbres de décision.

- L'algorithme va **apprendre** les seuils a et b, en passant plusieurs fois sur les données et calculant son erreur de classification à chaque fois, il va chercher à minimiser son erreur.