

ANOVA F-Test

Sélection de variables en Machine Learning

K. Kadri

Plan

- 1 Intuition générale
- 2 Principe mathématique
- 3 Application en Machine Learning

Pourquoi un F-Test ?

- Mesurer si une variable (feature) **diffère vraiment** selon les classes.
- Analyse statistique de la **séparation** entre classes.
- Utilisé dans : `SelectKBest(f_classif)`.

Question fondamentale

Une feature sépare-t-elle suffisamment bien les classes ?

Deux types de variances

- **Variance inter-classe** : différences entre les moyennes des groupes.
- **Variance intra-classe** : dispersion à l'intérieur d'un même groupe.

Statistique F

$$F = \frac{\text{Variance entre classes}}{\text{Variance intra-classes}}$$

Interprétation

- F élevé \rightarrow bonne séparation.
- F faible \rightarrow pas de différence \rightarrow feature inutile.

SelectKBest(f_classif)

- Calcule un F-score pour chaque feature.
- Classe les features du plus discriminant au moins utile.
- Garde les k meilleurs.

Avantages

- Ultra rapide.
- Simple, efficace, non paramétrique.

- Uniquement pour variables **numériques**.
- Ne capture ****pas**** les interactions entre features.
- Suppose variances similaires entre groupes (ANOVA).

Règle pratique

Utiliser ANOVA comme **premier filtre**, puis affiner avec RFE ou SelectFromModel.