Méthodes de filtrage (Feature Selection par filtres)
Formules, intuition, (mini-)démonstrations et interprétation des résultats

Objectif du document

Ce document résume les principales **méthodes de filtrage** (feature selection par *filtres*) utilisées en apprentissage automatique et en statistique :

— **Corrélation de Pearson**
— **Test du χ^2 (Khi-deux)**
— **Information mutuelle**
— **ANOVA (F-test)**
— **Seuil de variance**
— **Score de Fisher**
— **Différence absolue moyenne (MAD au sens "mean absolute difference") et MAD robuste (median absolute deviation)**
— **Rapport de dispersion**
— **Test de Kruskal–Wallis**
— **V de Cramér**

Idée générale. Un filtre évalue une variable explicative X **sans entraîner** (ou presque) un modèle complexe. On calcule un score/statistique mesurant l'association entre X et la cible Y (ou parfois uniquement la variabilité de X), puis on conserve les variables les plus informatives.

1 Notations communes

On observe n exemples (x_i, y_i) pour $i = 1, \dots, n$.

— $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$: moyenne ; $\mathrm{Var}(X)$: variance ; σ_X : écart-type.
— Pour une classification k-classes : groupes $g = 1, \dots, k$, tailles n_g, moyenne du groupe $\bar{x}_g$.
— En catégoriel : table de contingence O_{ij}, effectifs attendus E_{ij}.

2 Corrélation de Pearson

2.1 Formule

Pour X et Y quantitatives,

$$r = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}, \qquad r \in [-1, 1].$$

2.2 Intuition

r mesure l'intensité de la **relation linéaire** :

— $r = 1$: relation parfaitement linéaire croissante ; $r = -1$: décroissante.
— $r \approx 0$: pas de relation *linéaire* détectable (une relation non linéaire peut exister).

2.3 Mini-démonstration / justification

En régression linéaire simple $Y \approx aX + b$ (moindres carrés), le meilleur coefficient vaut

$$a = \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(X)}.$$

La corrélation r est une covariance *normalisée* par les échelles (σ_X et σ_Y), ce qui la rend comparable entre variables.

2.4 Interprétation des résultats

- Le signe indique le sens (croissant/décroissant).
- La magnitude $|r|$ indique la force (repères usuels) :

$$|r| < 0.1 \text{ négligeable}, \quad 0.1-0.3 \text{ faible}, \quad 0.3-0.5 \text{ modérée}, \quad > 0.5 \text{ forte}.$$

On peut aussi tester $H_0 : \rho = 0$ avec une p-value, mais **en filtrage on privilégie $|r|$ (taille d'effet) plutôt que la seule p-value**, car avec grand n presque tout devient "significatif".

2.5 Quand l'utiliser

- X quantitatif, Y quantitatif, relation attendue **à peu près linéaire**.
- Attention aux **outliers** et aux relations non linéaires.

Exemple

Données : $(X, Y) = \{(1, 2), (2, 4), (3, 6)\}$. On a $Y = 2X$ donc $r = 1$: relation linéaire parfaite. La variable X est **très informative** pour prédire Y. $\Rightarrow$ On la conserve presque toujours.

3 Test du χ^2 (Khi-deux) d'indépendance

3.1 Cadre

X et Y sont **catégorielles** (ou X a été discrétisée). On construit une table de contingence.

3.2 Formules

Effectifs observés O_{ij} (catégorie i de X, catégorie j de Y). Sous indépendance,

$$E_{ij} = \frac{(\text{total ligne } i)\,(\text{total colonne } j)}{n}.$$

Statistique :

$$\chi^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \geq 0.$$

Sous H_0 (indépendance) et sous des conditions d'effectifs, $\chi^2 \approx \chi^2_{(r-1)(c-1)}$.

3.3 Intuition

On compare **observé** et **attendu** si X et Y étaient indépendantes. Grand $\chi^2 \Rightarrow$ grands écarts $\Rightarrow$ association.

3.4 Justification (idée)

En grand échantillon, les écarts $O_{ij} - E_{ij}$ se comportent comme des fluctuations approximativement normales ; la somme de carrés normalisés mène à une loi χ^2.

3.5 Interprétation

- p-value petite (ex. < 0.05) : on rejette l'indépendance.
- χ^2 augmente avec n : **la p-value ne donne pas une force d'association comparable**. On préfère compléter par **V de Cramér**.

3.6 Quand l'utiliser

- Classification avec X catégorielles (ou discrétisées) et Y catégorielle.
- Prudence si beaucoup de cases ont des E_{ij} petits (règle pratique : éviter $E_{ij} < 5$ trop souvent).

Exemple

Variable $X = \{\text{Homme,Femme}\}$, $Y = \{\text{Oui,Non}\}$. Si 90% des femmes disent Oui contre 10% des hommes, χ^2 est grand $\Rightarrow$ dépendance. La distribution observée est très différente de l'indépendance. La variable catégorielle explique la cible. $\Rightarrow$ On la garde pour la classification.

4 V de Cramér

4.1 Formule

À partir du χ^2 d'une table $r \times c$:

$$V = \sqrt{\frac{\chi^2}{n\,(m-1)}}, \qquad m = \min(r, c), \qquad V \in [0, 1].$$

4.2 Intuition

V est une **taille d'effet** : une version normalisée de χ^2, donc plus comparable entre jeux de données et tailles d'échantillon.

4.3 Interprétation

Repères usuels (indicatifs) :

$$V < 0.1 \text{ négligeable}, \quad 0.1-0.3 \text{ faible}, \quad 0.3-0.5 \text{ modérée}, \quad \geq 0.5 \text{ forte}.$$

En filtrage, on garde les variables avec les plus grands V.

4.4 Quand l'utiliser

Association **catégoriel–catégoriel** (souvent avec le test χ^2).

Exemple

Si $\chi^2 = 40$, $n = 200$, $m = 2$:

$$V = \sqrt{\frac{40}{200}} = 0.45 \Rightarrow \text{association modérée à forte.}$$

Association modérée à forte. Variable catégorielle **utile**. $\Rightarrow$ À conserver en priorité.

5 Information mutuelle (Mutual Information, MI)

5.1 Définition et formules

Pour variables discrètes :

$$I(X;Y) = \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \geq 0.$$

Formules équivalentes :

$$I(X;Y) = H(X) + H(Y) - H(X,Y) = H(Y) - H(Y \mid X),$$

où H est l'entropie (incertitude).

5.2 Intuition

MI mesure la **réduction d'incertitude** sur Y quand on connaît X.

— $I(X;Y) = 0 \iff X$ et Y indépendantes.

— Capte des dépendances **non linéaires** et plus générales que Pearson.

5.3 Justification (idée : divergence KL)

$$I(X;Y) = D_{\mathrm{KL}}(p(x,y) \,\|\, p(x)p(y)) \geq 0,$$

puisque une divergence KL est toujours ≥ 0.

5.4 Interprétation

— MI est toujours ≥ 0.

— Il n'y a pas de maximum universel en continu : on interprète **surtout de façon relative** (classement des variables).

5.5 Quand l'utiliser

— Quand la relation peut être **non linéaire**.

— Pour X et/ou Y continus, il faut une estimation (bins, kNN, etc.) : le résultat dépend de l'estimateur.

Exemple

Si $Y = X^2$ (relation non linéaire) :

— Pearson ≈ 0

— Information mutuelle > 0 (dépendance détectée)

Dépendance non linéaire détectée. Variable utile même si les méthodes linéaires échouent. $\Rightarrow$ À conserver pour modèles non linéaires.

6 ANOVA (Analyse de la variance) : F-test

6.1 Cadre

X est **quantitatif**, Y est **catégorielle** (k classes). On teste l'égalité des moyennes de X entre groupes.

6.2 Formules

Moyenne globale $\bar{x}$, moyenne de groupe $\bar{x}_g$.

$$SSB = \sum_{g=1}^{k} n_g(\bar{x}_g - \bar{x})^2 \quad \text{(inter-groupes)}$$

$$SSW = \sum_{g=1}^{k} \sum_{i \in g} (x_i - \bar{x}_g)^2 \quad \text{(intra-groupes)}$$

$$MSB = \frac{SSB}{k-1}, \qquad MSW = \frac{SSW}{n-k}, \qquad F = \frac{MSB}{MSW} \geq 0.$$

Sous H_0 (mêmes moyennes) et hypothèses classiques (normalité approx + variances égales), $F \sim F_{k-1,n-k}$.

6.3 Intuition

- SSB mesure la séparation des **moyennes** entre classes (signal).
- SSW mesure la dispersion **dans** chaque classe (bruit).
- F grand $\Rightarrow$ bonne séparation.

6.4 Justification (idée)

Sous normalité et homoscédasticité, SSB et SSW (normalisés) se comportent comme des χ^2 indépendants, et leur ratio donne une loi F.

6.5 Interprétation

- $F \approx 1$: pas de séparation claire ; $F \gg 1$: séparation forte.
- La p-value indique si les différences sont détectables, mais **en filtrage on trie plutôt par F (taille d'effet)**.

6.6 Quand l'utiliser

- Filtrage supervisé en classification avec features quantitatives.
- Si hypothèses douteuses (outliers, non-normalité, variances très différentes), préférer **Kruskal–Wallis**.

Exemple

Notes moyennes :

- Classe A : 10
- Classe B : 18

Variances internes faibles $\Rightarrow$ F grand $\Rightarrow$ variable discriminante. Moyennes très différentes entre classes. Variable très discriminante. $\Rightarrow$ Améliore fortement la séparation des classes.

7 Seuil de variance (Variance Threshold)

7.1 Formule

$$\mathrm{Var}(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \geq 0.$$

Pour une variable binaire ($X \in \{0,1\}$, $p = P(X = 1)$) :

$$\mathrm{Var}(X) = p(1-p),$$

maximale à $p = 0.5$, minimale près de 0 ou 1.

7.2 Intuition

Une variable quasi constante contient peu (voire pas) d'information utile.

7.3 Interprétation

— Si $\mathrm{Var}(X) \approx 0$, X est presque constante $\Rightarrow$ candidate naturelle à supprimer.

— Choix du seuil τ : dépend de l'échelle ; souvent on standardise avant ou on choisit un seuil adapté au domaine.

7.4 Quand l'utiliser

— **Pré-nettoyage** rapide (surtout en haute dimension, one-hot, texte).

— Méthode **non supervisée** (ne regarde pas Y).

Exemple

$X = \{1,1,1,1,1\}$:

$$\mathrm{Var}(X) = 0 \Rightarrow \text{variable inutile.}$$

Aucune variabilité. Variable inutile pour tout modèle. $\Rightarrow$ À supprimer sans risque.

8 Score de Fisher

8.1 Formule (binaire)

Pour deux classes 0 et 1 (moyennes μ_0, μ_1, variances σ_0^2, σ_1^2) :

$$\mathrm{Fisher}(X) = \frac{(\mu_1 - \mu_0)^2}{\sigma_1^2 + \sigma_0^2} \geq 0.$$

Version multi-classes : ratio *dispersion inter-classes / intra-classes* (même esprit que l'ANOVA).

8.2 Intuition

— Numérateur : séparation des moyennes.

— Dénominateur : bruit intra-classe.

— Score élevé $\Rightarrow$ bonne séparabilité.

8.3 Justification (idée)

En classification linéaire (ex. LDA), un principe central est de maximiser la séparation inter-classes tout en minimisant la dispersion intra-classe. Fisher score est la version 1D de ce principe.

8.4 Interprétation

Pas de p-value standard : on interprète **relativement** (classement) :

plus Fisher est grand, plus la variable est utile.

8.5 Quand l'utiliser

Filtrage supervisé très utilisé en classification ; attention aux outliers (moyenne/variance).

Exemple

$\mu_0 = 2$, $\mu_1 = 8$, variances faibles $\Rightarrow$ Fisher très grand. Très bonne séparabilité. Variable clé pour la classification. $\Rightarrow$ Forte amélioration de la précision.

9 Différence absolue moyenne (MAD) et MAD robuste

9.1 (A) "Mean Absolute Difference" (différence absolue des moyennes, binaire)

$$\mathrm{MAD}_{\text{mean-diff}}(X) = |\mu_1 - \mu_0|.$$

Interprétation : 0 = pas de séparation ; plus c'est grand, plus la séparation des moyennes est forte.

Important. Sans standardisation, ce score dépend fortement de l'unité (euros vs pourcent). On standardise souvent X avant comparaison.

9.2 (B) MAD robuste = Median Absolute Deviation (dispersion robuste)

$$\mathrm{MAD}_{\text{median}}(X) = \mathrm{median}\Big(|x_i - \mathrm{median}(X)|\Big).$$

Intuition : mesure de dispersion robuste aux outliers.

Usage en filtrage. Comme la variance, la MAD robuste peut servir à supprimer les variables quasi constantes de façon robuste.

Exemple

Classe 0 : moyenne 5, Classe 1 : moyenne 15 :

$$|\mu_1 - \mu_0| = 10 \Rightarrow \text{fort pouvoir discriminant.}$$

Séparation importante. Variable informative mais sensible à l'échelle. $\Rightarrow$ Standardiser avant usage.

10 Rapport de dispersion (Dispersion Ratio)

10.1 Définition (supervisée, proche ANOVA)

Dans de nombreux cours, le "rapport de dispersion" désigne le ratio

$$DR(X) = \frac{\text{dispersion inter-classes}}{\text{dispersion intra-classes}} \approx \frac{MSB}{MSW} = F.$$

Donc, dans ce cadre, le rapport de dispersion est essentiellement **équivalent au F-test ANOVA** (même intuition et même interprétation : plus grand est meilleur).

10.2 Interprétation

$$DR \approx 1 \Rightarrow \text{peu de séparation}, \qquad DR \gg 1 \Rightarrow \text{bonne séparation}.$$

On l'utilise surtout comme **score de classement**.

Exemple

Dispersion inter = 50, intra = 5 :

$$DR = 10 \Rightarrow \text{variable très utile}.$$

Signal dominant sur le bruit. Très bon pouvoir discriminant. ⇒ À prioriser.

11 Test de Kruskal–Wallis (non paramétrique)

11.1 Cadre

X est ordinal/quantitatif, Y est catégorielle (k groupes). Test non paramétrique basé sur les **rangs** (alternative à ANOVA).

11.2 Formule

On remplace x_i par son rang R_i (moyenne des rangs en cas d'égalité). Soit R_g la somme des rangs dans le groupe g :

$$H = \frac{12}{n(n+1)} \sum_{g=1}^{k} \frac{R_g^2}{n_g} - 3(n+1).$$

Sous H_0 (mêmes distributions), $H \approx \chi^2_{k-1}$ (avec correction en cas de nombreux ex-aequo).

11.3 Intuition

Si un groupe a des valeurs systématiquement plus grandes, il reçoit des rangs plus grands, donc H augmente.

11.4 Interprétation

— p-value petite : au moins un groupe diffère (en position ou distribution).

— En filtrage, on trie par H (taille d'effet relative) plutôt que par la seule p-value.

11.5 Quand l'utiliser

— Quand ANOVA est risquée : non-normalité, outliers, variances inégales.

— Peut être moins puissant qu'ANOVA si les hypothèses d'ANOVA sont parfaitement satisfaites.

Exemple

Si toutes les valeurs du groupe A sont plus grandes que celles du groupe B, les rangs de A sont plus grands ⇒ H grand. Distributions différentes. Variable utile même sans normalité. ⇒ Préférable à ANOVA si outliers.

12 Guide d'interprétation (résumé décisionnel)

12.1 Que regarder : p-value vs taille d'effet

— **P-values** (tests χ^2, ANOVA, Kruskal) : indiquent si un effet est détectable, mais dépendent fortement de n.

— **Tailles d'effet / scores** ($|r|$, V, Fisher, F, MI, etc.) : plus utiles pour **classer** les variables.

Règle d'or : en filtrage, privilégier le **classement par score** (taille d'effet) et valider ensuite avec un modèle (validation croisée).

12.2 Tableau comparatif rapide

Méthode	Type de X	Type de Y	Interprétation principale
Pearson	quanti	quanti	$\lvert r\rvert$ grand = lien linéaire fort
χ^2	catégoriel	catégoriel	p-value, et χ^2 (mais dépend de n)
V de Cramér	catégoriel	catégoriel	V grand = association forte
MI	discret/continu	discret/continu	score $\uparrow$ = dépendance (souvent non linéaire)
ANOVA (F)	quanti	catégoriel	F grand = moyennes bien séparées
Fisher	quanti	catégoriel	score grand = séparabilité élevée
Variance threshold	quanti	—	var faible = quasi-constante
Kruskal–Wallis	ordinal/quanti	catégoriel	H grand = distributions différentes

12.3 Repères pratiques (indicatifs)

— Pearson : garder top-k selon $|r|$ (ou $|r| > \tau$).

— Cramér V : $V \geq 0.1$ commence à indiquer un lien ; $V \geq 0.3$ plutôt utile.

— ANOVA/Fisher/Kruskal : garder top-k selon F/Fisher/H.

— Variance : supprimer $\mathrm{Var}(X) < \tau$ (après normalisation si nécessaire).

— MI : garder top-k ; interprétation surtout relative.

Pièges et bonnes pratiques

— **Non-linéarité :** Pearson peut être proche de 0 malgré un lien non linéaire ; MI aide.

— **Outliers :** Pearson/ANOVA/Fisher sensibles ; Kruskal–Wallis et MAD robuste sont plus résistants.

— **Grand n :** p-values très petites même pour des effets faibles $\Rightarrow$ privilégier tailles d'effet.

— **Échelles :** standardiser si on compare des scores basés sur des moyennes/variances.

— **Validation finale :** après filtrage, toujours vérifier via validation croisée avec le modèle cible.