

Clustering Parameter Tuning Guidelines

Recommended Ranges, Methods, and Best Practices

K-Means

Parameter	Range	Default	Tuning Method
n_clusters (K)	2-10	3-5	Elbow/Silhouette
init	['k-means++', 'random']	k-means++	Always k-means++
n_init	10-100	10	More for stability
max_iter	100-1000	300	Increase if no convergence
tol	1e-6 to 1e-2	1e-4	Smaller for precision

DBSCAN

Parameter	Range	Default	Tuning Method
eps	0.01-2.0	0.5	k-distance plot
min_samples	3-20	2*dims	Domain knowledge
metric	['euclidean', 'manhattan']	euclidean	Data dependent
algorithm	['auto', 'ball_tree']	auto	Auto is fine
leaf_size	10-50	30	Memory vs speed

GMM

Parameter	Range	Default	Tuning Method
n_components	2-10	3-5	BIC/AIC
covariance_type	['full', 'diag', 'full_spherical']	full	Start full, simplify
max_iter	50-500	100	Monitor convergence
n_init	1-10	1	More for stability
init_params	['kmeans', 'random']	kmeans	kmeans faster

Tuning Strategies

Grid Search

Pros: Exhaustive, Reproducible, Simple

Cons: Slow, Curse of dimensionality

Use when: Small parameter space

Random Search

Pros: Faster, Better for many params, Parallelizable

Cons: May miss optimum, Not reproducible

Use when: Large parameter space

Bayesian Opt

Pros: Efficient, Learns from history, Fewer iterations

Cons: Complex, Overhead for simple problems

Use when: Expensive evaluations

Validation Metrics

Metric	Range	Interpretation	Use For
Silhouette	[-1, 1]	Higher is better	General quality
Davies-Bouldin	[0, ∞)	Lower is better	Cluster separation
Calinski-Harabasz	[0, ∞)	Higher is better	Dense clusters
Inertia	[0, ∞)	Lower is better	K-means only
BIC/AIC	(-∞, ∞)	Lower is better	GMM selection

Tuning Best Practices

- Start with defaults, then tune
- Use cross-validation when possible
- Consider computational budget
- Log all experiments
- Visualize parameter effects
- Use domain knowledge
- Check stability across runs
- Don't overfit to metrics

IMPORTANT:
No metric is perfect!
Always validate with:

- Visual inspection
- Domain expertise
- Business goals

TUNING WORKFLOW: 1. Default Parameters → 2. Coarse Grid Search → 3. Fine-tune Best Region → 4. Validate Stability → 5. Production Settings