k-近邻（k-NN）：理论与实践

2025 年 9 月 10 日

1 引言

k-近邻（k-Nearest Neighbors, k-NN）是一种非参数、基于实例的“惰性学习”方法：预测时在训练集中查找与查询样本最近的 k 个邻居。其思想直观、实现简单，在低维、良好缩放的数据上常有竞争力，但对特征缩放敏感，并在高维下性能下降。

2 原理与公式

给定查询点 $\mathbf{x}$，在指定距离度量 $d(\cdot,\cdot)$（如欧氏、曼哈顿）下找到其最近的 k 个邻居。分类任务采用多数表决（可选距离加权 `weights=distance` 使近邻权重更高）；回归任务取邻居目标值的平均（或距离加权平均）。

计算上，朴素查找每次预测代价为 $\mathcal{O}(nd)$（n 为样本数，d 为维度）。中等维度时可使用 KDTree/BallTree 提速。k-NN 受“维度灾难”影响，合适的特征缩放与度量选择至关重要。

3 应用与技巧

- **选择 k**：通过交叉验证调参；二分类常取奇数 k 以减少平票。
- **缩放**：对特征做标准化/归一化，或使用 Pipeline；距离对量纲非常敏感。
- **度量**：尝试欧氏与曼哈顿；必要时考虑领域特定距离。
- **权重**： `uniform` 与 `distance` 可对类间重叠区产生不同效果。
- **复杂度**：预测代价随数据量增长；大规模可考虑近似近邻检索。

4 Python 实战

在本章节目录运行下述命令，图片将保存到 `figures/`：

Listing 1: 生成 k-NN 配图

```
python gen_knn_figures.py
```

Listing 2: gen_knn_figures.py 源码

```
"""
Figure generator for the k-NN chapter.

Generates illustrative figures and saves them into the chapter's 'figures/'
folder next to this script, regardless of current working directory.

Requirements:
- Python 3.8+
- numpy, matplotlib, scikit-learn

Install (if needed):
  pip install numpy matplotlib scikit-learn

This script avoids newer or experimental APIs for broader compatibility.
"""
from __future__ import annotations

import os
import numpy as np
import matplotlib.pyplot as plt
from matplotlib.colors import ListedColormap

try:
    from sklearn.datasets import make_moons, make_regression, make_classification
    from sklearn.neighbors import KNeighborsClassifier, KNeighborsRegressor
    from sklearn.preprocessing import StandardScaler
    from sklearn.pipeline import make_pipeline
    from sklearn.model_selection import cross_val_score
except Exception:
    raise SystemExit(
```

```python
        "Missing scikit-learn. Please install with: pip install scikit-learn"
    )


def _ensure_figures_dir(path: str | None = None) -> str:
    """Create figures directory under this chapter regardless of CWD."""
    if path is None:
        base = os.path.dirname(os.path.abspath(__file__))
        path = os.path.join(base, "figures")
    os.makedirs(path, exist_ok=True)
    return path


def _plot_decision_boundary(ax, clf, X, y, title: str):
    x_min, x_max = X[:, 0].min() - 0.5, X[:, 0].max() + 0.5
    y_min, y_max = X[:, 1].min() - 0.5, X[:, 1].max() + 0.5
    xx, yy = np.meshgrid(
        np.linspace(x_min, x_max, 400), np.linspace(y_min, y_max, 400)
    )
    Z = clf.predict(np.c_[xx.ravel(), yy.ravel()]).reshape(xx.shape)
    cmap_light = ListedColormap(["#FFEEEE", "#EEEEFF"])
    cmap_bold = ListedColormap(["#E74C3C", "#3498DB"])
    ax.contourf(xx, yy, Z, cmap=cmap_light, alpha=0.8, levels=np.unique(Z).size)
    ax.scatter(X[:, 0], X[:, 1], c=y, cmap=cmap_bold, edgecolors="k", s=20)
    ax.set_title(title)
    ax.set_xlabel("Feature 1")
    ax.set_ylabel("Feature 2")


def fig_knn_k_compare(out_dir: str) -> str:
    np.random.seed(0)
    X, y = make_moons(n_samples=500, noise=0.3, random_state=0)
    models = [
        (KNeighborsClassifier(n_neighbors=1), "k=1 (high variance)"),
        (KNeighborsClassifier(n_neighbors=15), "k=15 (smoother)")
    ]
    fig, axes = plt.subplots(1, 2, figsize=(9.5, 4.2), dpi=150, sharex=True, sharey=True)
    for ax, (m, title) in zip(axes, models):
```

```python
        m.fit(X, y)
        _plot_decision_boundary(ax, m, X, y, f"k-NN: {title}")
    fig.suptitle("Effect of k on decision boundary")
    out_path = os.path.join(out_dir, "knn_k_compare.png")
    fig.tight_layout(rect=[0, 0.03, 1, 0.95])
    fig.savefig(out_path)
    plt.close(fig)
    return out_path


def fig_knn_metric_compare(out_dir: str) -> str:
    np.random.seed(1)
    X, y = make_moons(n_samples=500, noise=0.28, random_state=1)
    models = [
        (KNeighborsClassifier(n_neighbors=11, metric="euclidean"), "metric=euclidean"),
        (KNeighborsClassifier(n_neighbors=11, metric="manhattan"), "metric=manhattan"),
    ]
    fig, axes = plt.subplots(1, 2, figsize=(9.5, 4.2), dpi=150, sharex=True, sharey=True)
    for ax, (m, title) in zip(axes, models):
        m.fit(X, y)
        _plot_decision_boundary(ax, m, X, y, f"k-NN: {title}")
    fig.suptitle("Effect of distance metric")
    out_path = os.path.join(out_dir, "knn_metric_compare.png")
    fig.tight_layout(rect=[0, 0.03, 1, 0.95])
    fig.savefig(out_path)
    plt.close(fig)
    return out_path


def fig_knn_scaling_effect(out_dir: str) -> str:
    np.random.seed(2)
    X, y = make_classification(
        n_samples=600,
        n_features=2,
        n_informative=2,
        n_redundant=0,
        n_clusters_per_class=1,
        class_sep=1.0,
        random_state=2,
    )
```

```python
    # Impose different scales on features
    X_scaled_variance = X.copy()
    X_scaled_variance[:, 0] *= 8.0  # make feature 0 dominate distances

    knn_raw = KNeighborsClassifier(n_neighbors=11)
    knn_std = make_pipeline(StandardScaler(), KNeighborsClassifier(n_neighbors=11))

    fig, axes = plt.subplots(1, 2, figsize=(9.5, 4.2), dpi=150, sharex=True, sharey=True)
    knn_raw.fit(X_scaled_variance, y)
    _plot_decision_boundary(axes[0], knn_raw, X_scaled_variance, y, "Without scaling")
    knn_std.fit(X_scaled_variance, y)
    _plot_decision_boundary(axes[1], knn_std, X_scaled_variance, y, "With StandardScaler")
    fig.suptitle("Feature scaling impact on k-NN")
    out_path = os.path.join(out_dir, "knn_scaling_effect.png")
    fig.tight_layout(rect=[0, 0.03, 1, 0.95])
    fig.savefig(out_path)
    plt.close(fig)
    return out_path


def fig_knn_weight_compare(out_dir: str) -> str:
    np.random.seed(3)
    X, y = make_moons(n_samples=500, noise=0.32, random_state=3)
    models = [
        (KNeighborsClassifier(n_neighbors=11, weights="uniform"), "weights=uniform"),
        (KNeighborsClassifier(n_neighbors=11, weights="distance"), "weights=distance"),
    ]
    fig, axes = plt.subplots(1, 2, figsize=(9.5, 4.2), dpi=150, sharex=True, sharey=True)
    for ax, (m, title) in zip(axes, models):
        m.fit(X, y)
        _plot_decision_boundary(ax, m, X, y, f"k-NN: {title}")
    fig.suptitle("Uniform vs distance weighting")
    out_path = os.path.join(out_dir, "knn_weight_compare.png")
    fig.tight_layout(rect=[0, 0.03, 1, 0.95])
    fig.savefig(out_path)
    plt.close(fig)
```

```python
    return out_path


def fig_knn_regression_curve(out_dir: str) -> str:
    rng = np.random.RandomState(4)
    # 1D regression: y = sin(x) + noise
    X = np.sort(rng.uniform(-3.0, 3.0, size=150)).reshape(-1, 1)
    y = np.sin(X).ravel() + rng.normal(scale=0.25, size=X.shape[0])

    grid = np.linspace(-3.5, 3.5, 600).reshape(-1, 1)
    models = [
        (KNeighborsRegressor(n_neighbors=1), "k=1"),
        (KNeighborsRegressor(n_neighbors=15), "k=15"),
        (KNeighborsRegressor(n_neighbors=45), "k=45"),
    ]
    fig, ax = plt.subplots(figsize=(7.5, 4.2), dpi=160)
    ax.scatter(X[:, 0], y, s=18, c="#555", alpha=0.7, label="data")
    colors = ["#E74C3C", "#3498DB", "#2ECC71"]
    for (m, title), col in zip(models, colors):
        m.fit(X, y)
        y_pred = m.predict(grid)
        ax.plot(grid[:, 0], y_pred, color=col, lw=2, label=title)
    ax.set_title("k-NN regression: smoothing vs k")
    ax.set_xlabel("x")
    ax.set_ylabel("y")
    ax.legend()
    ax.grid(True, linestyle=":", alpha=0.4)
    out_path = os.path.join(out_dir, "knn_regression_curve.png")
    fig.tight_layout()
    fig.savefig(out_path)
    plt.close(fig)
    return out_path


def main():
    out_dir = _ensure_figures_dir(None)
    generators = [
        fig_knn_k_compare,
        fig_knn_metric_compare,
        fig_knn_scaling_effect,
        fig_knn_weight_compare,
        fig_knn_regression_curve,
    ]
```

```python
    print("Generating figures into:", os.path.abspath(out_dir))
    for gen in generators:
        try:
            p = gen(out_dir)
            print("Saved:", p)
        except Exception as e:
            print("Failed generating", gen.__name__, ":", e)


if __name__ == "__main__":
    main()
```

5 结果

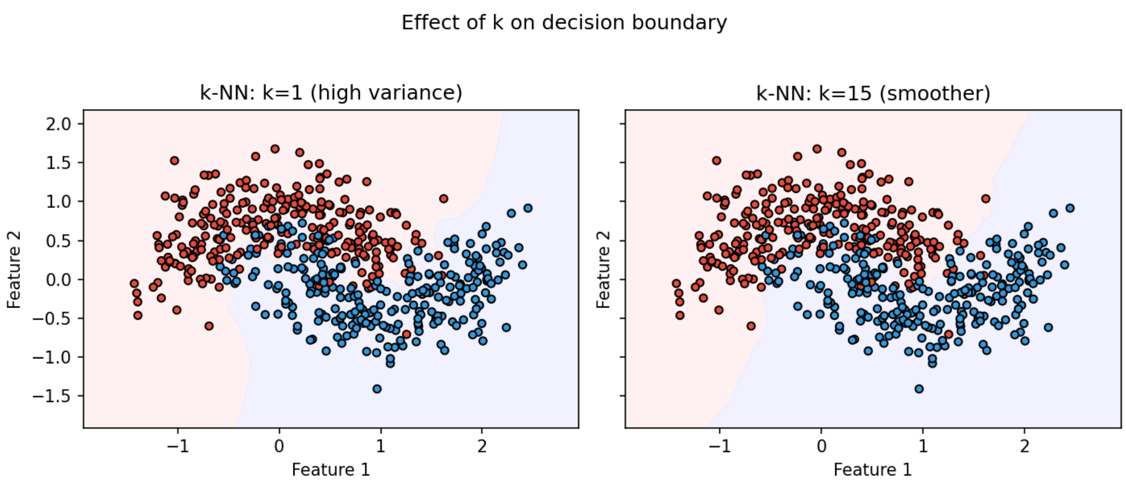

图 1: 不同 k（1 vs 15）的决策边界对比。

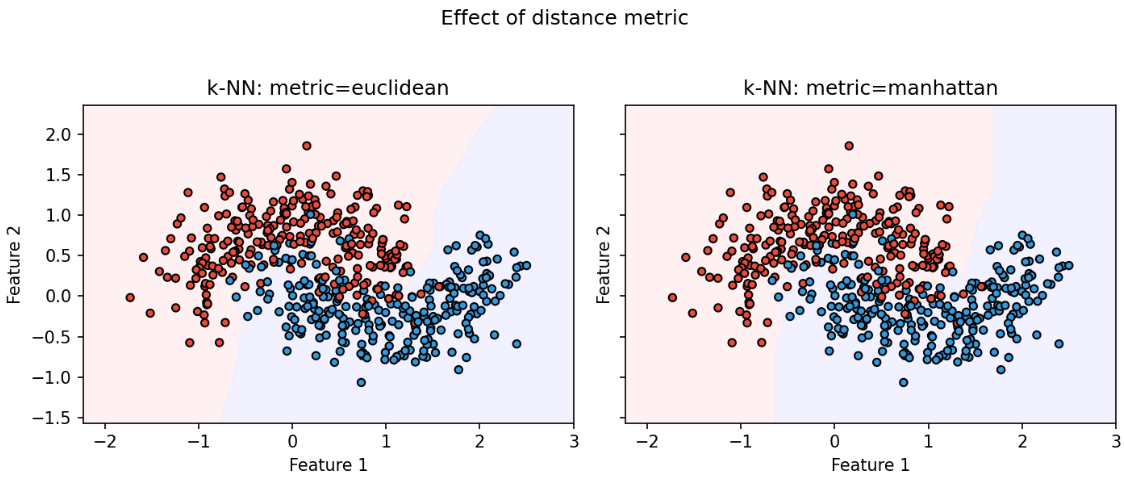

图 2: 不同距离度量：欧氏（Euclidean）vs 曼哈顿（Manhattan）。

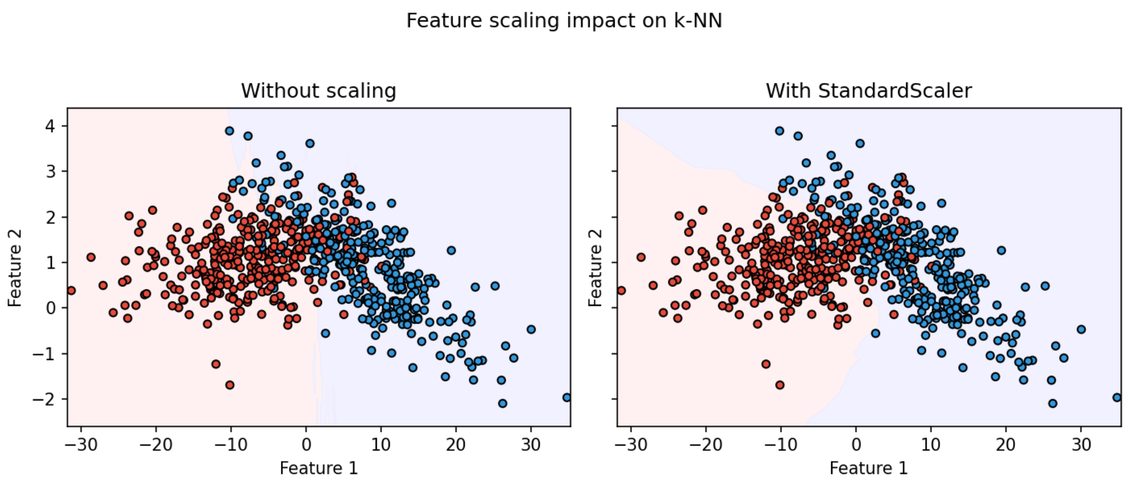

图 3: 特征缩放对决策边界的影响。

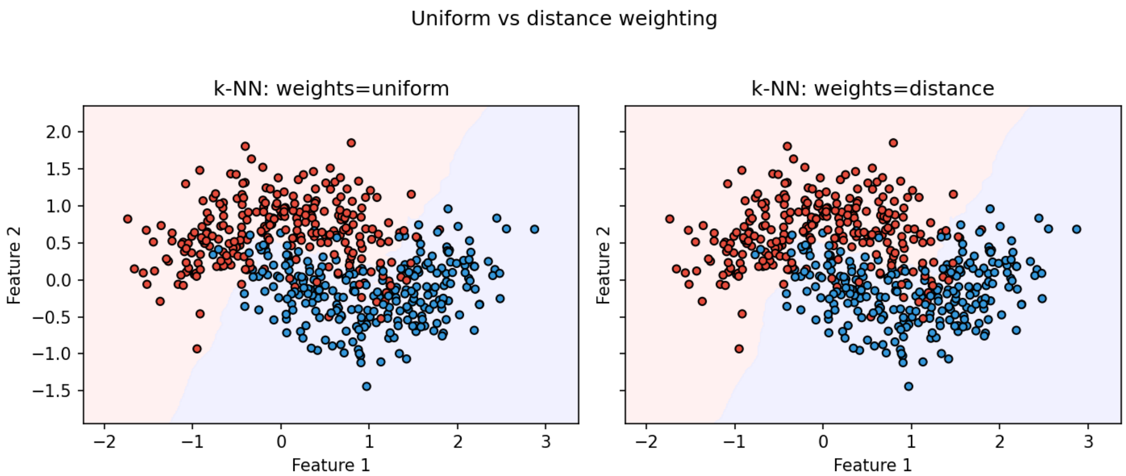

图 4: 均匀权重 vs 距离加权的对比。

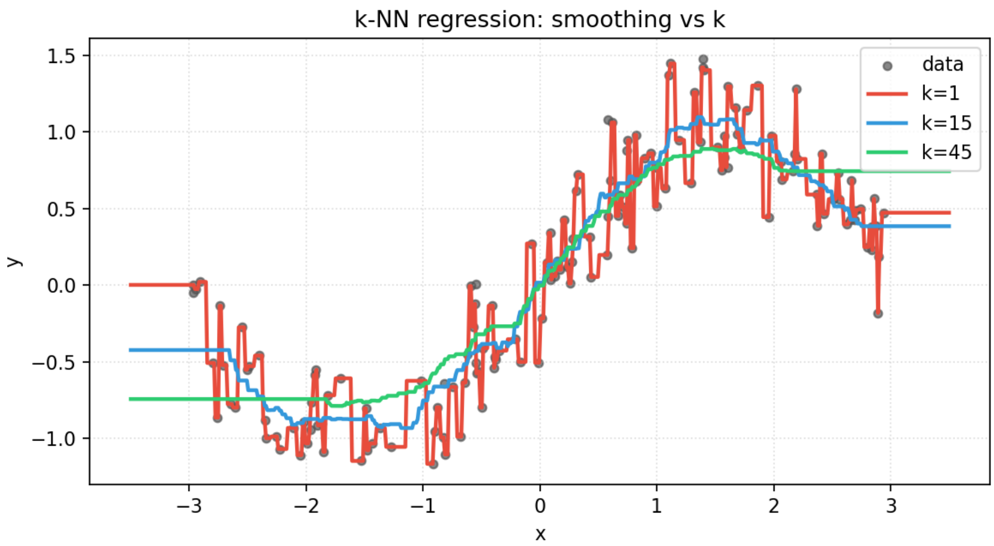

图 5: k-NN 回归：随 k 增大平滑程度的变化。

6 总结

k-NN 在特征良好缩放且维度适中的场景下是简洁有效的基线。通过验证选择合适的 k、距离度量与权重设置，并做好缩放预处理，可获得稳定表现。