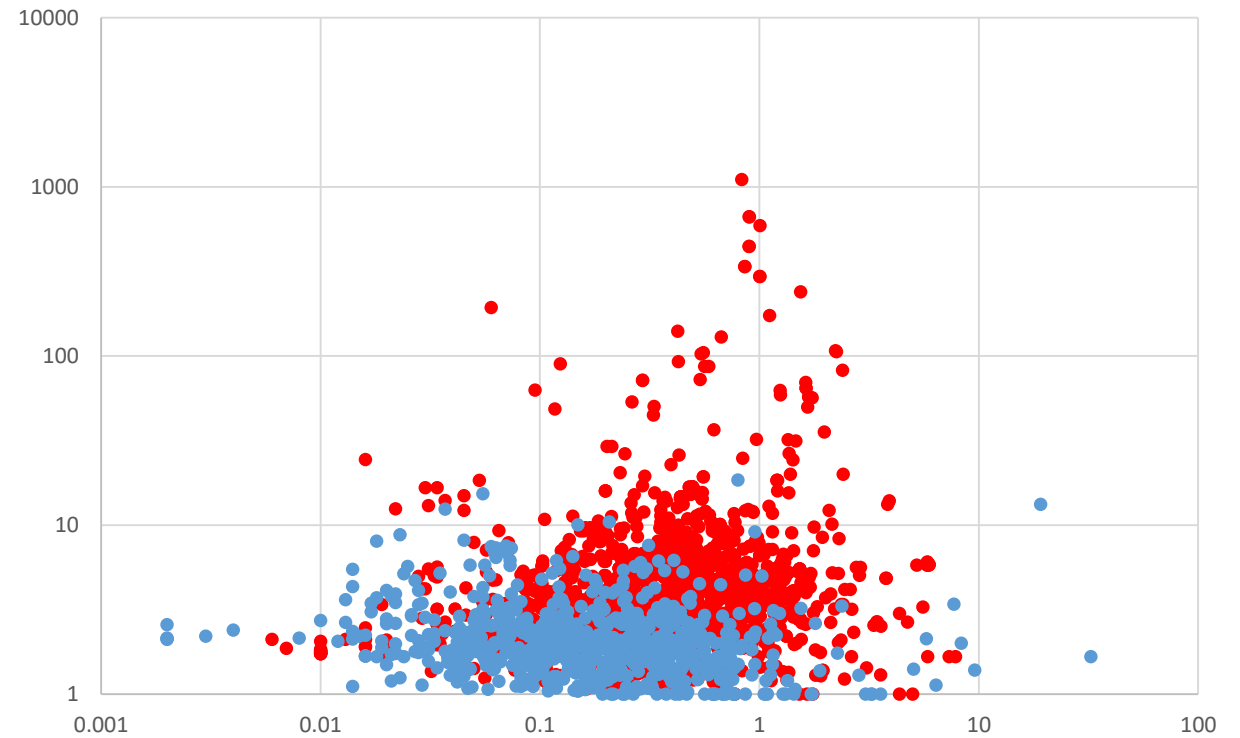


Задача классификации

- Ввод (Input): \mathbf{X}
- Вывод (Output): \mathbf{Y}
- Целевая зависимость (Target function): $f: \mathbf{X} \rightarrow \mathbf{Y}$
- Данные (Data): $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_N, y_N)$
- Гипотеза (Hypothesis): $h: \mathbf{X} \rightarrow \mathbf{Y}$

\mathbf{Y} – конечное множество, множество классов



Метрические классификаторы

Гипотеза компактности

«Гипотеза компактности — в задачах классификации предположение о том, что схожие объекты гораздо чаще лежат в одном классе, чем в разных; или, другими словами, что классы образуют компактно локализованные подмножества в пространстве объектов.»

«В математическом анализе *компактными* называются ограниченные замкнутые множества. *Гипотеза компактности* не имеет ничего общего с этим понятием и должна пониматься в «более бытовом» смысле этого слова.»

Метрический классификатор

Lazy learning

$$h(x; D) = \arg \max_{y \in Y} \underbrace{\sum_{x_i \in D} [y_i = y] w(x_i, x)}_{\Gamma_y(x)}$$

$w(x_i, x)$ — вес соседа x_i

$\Gamma_y(x)$ — близость x к классу y

kNN – k ближайших соседей

$w(x_i, x) = 1$, если x_i – один из k ближайших соседей

$w(x_i, x) = 1$, если $\rho(x_i, x) < R$ (Radius Neighbors)

Оценка качества:

$$LOO(k, D) = \sum_{x_i \in D} h(x_i; D \setminus x_i; k) \neq y_i$$

WkNN – k взвешенных ближайших соседей

Варианты w :

$$w_i = \frac{k - \rho(x, x_i)}{k}$$

$$w_i = q^{-\rho(x, x_i)}$$

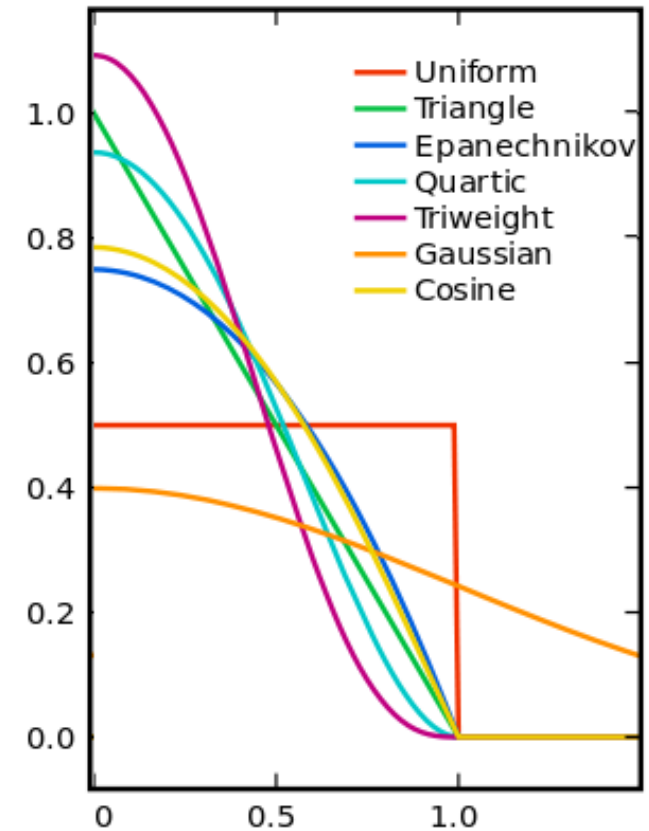
Метод окна Парзена (Parzen window):

$$w(x, x_i) = K \left(\frac{\rho(x, x_i)}{h} \right)$$

Фиксированная ширина

$$w(x, x_i) = K \left(\frac{\rho(x, x_i)}{\rho(x, x_j)} \right), \text{ где } x_j - (k+1)\text{-й сосед}$$

Переменная ширина



Метод потенциальных функций

$$h(x, D) = \arg \max_{y \in Y} \sum_{x_i \in D} [y_i = y] \gamma_i K \left(\frac{\rho(x, x_i)}{r_i} \right)$$

γ_i — веса объектов (заряд) Инициализация: $\gamma_i = 0$

r_i - радиус действия Если $h(x_i) \neq y_i \rightarrow \gamma_i = \gamma_i + 1$

В самом начале можно выбрать случайно или по наибольшему классу.

Prototype selection (Отбор эталонов)

$$f(\mathbf{x}): X \rightarrow Y$$

$$h(\mathbf{x}) = \arg \max_{y \in Y} \Gamma_y(\mathbf{x})$$

$$\text{Margin (отступ): } M(\mathbf{x}_i) = \Gamma_{y_i}(\mathbf{x}_i) - \max_{y \in Y \setminus y_i} \Gamma_y(\mathbf{x}_i)$$

Prototype selection (Отбор эталонов)

$$h(\mathbf{x}; \Omega) = \arg \max_{y \in Y} \sum_{\mathbf{x}_i \in \Omega} [y_i = y] w(\mathbf{x}_i, \mathbf{x})$$

Методы отбора эталонов

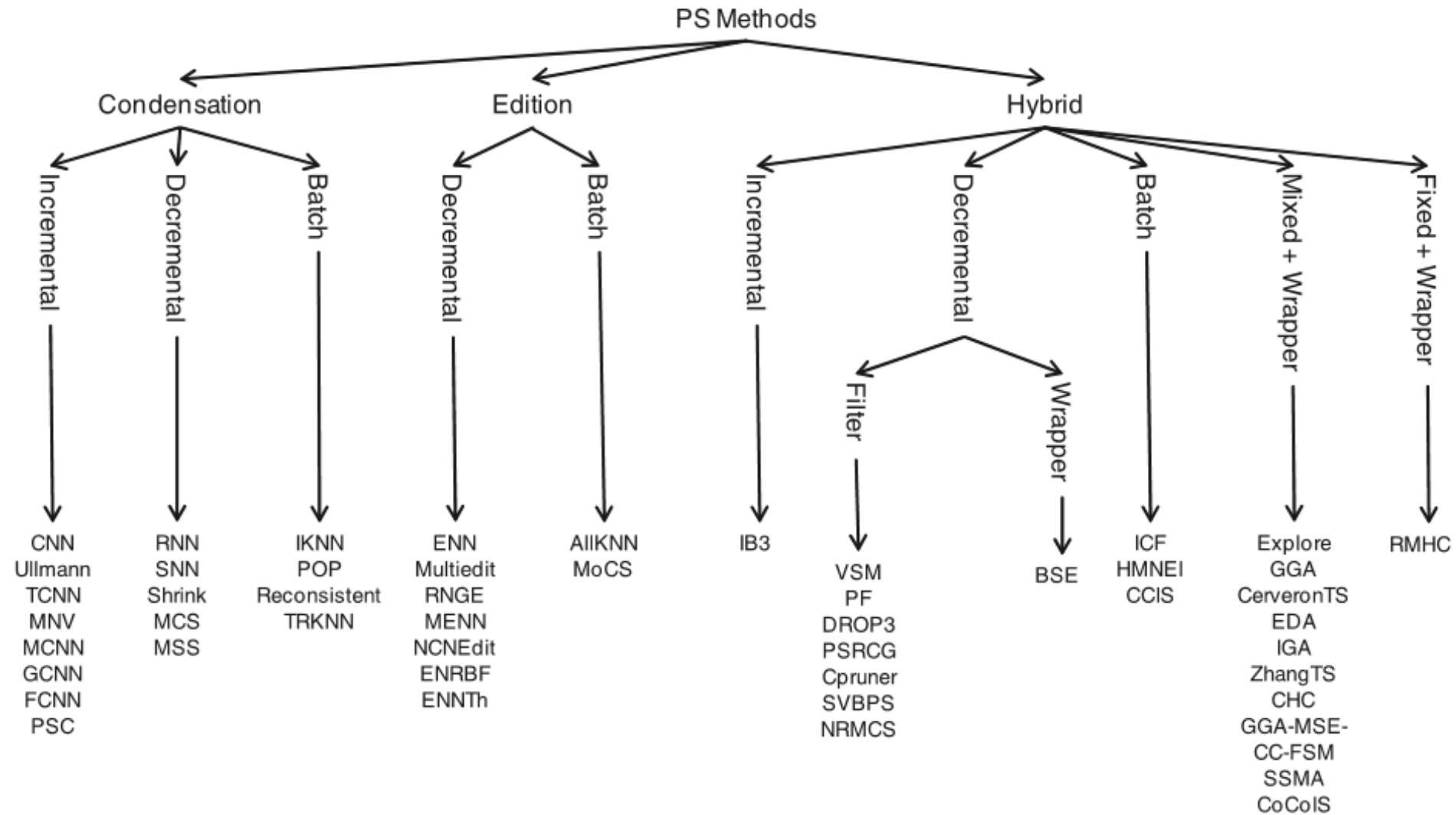
Направление поиска:

- Incremental
- Decremental
- Batch (Decremental)
- Mixed
- Fixed (Mixed)
- Replacement

Тип выбора:

- Condensation
- Edition
- Hybrid

Методы отбора эталонов



DROP5

(Decremental Reduction Optimization Procedure)

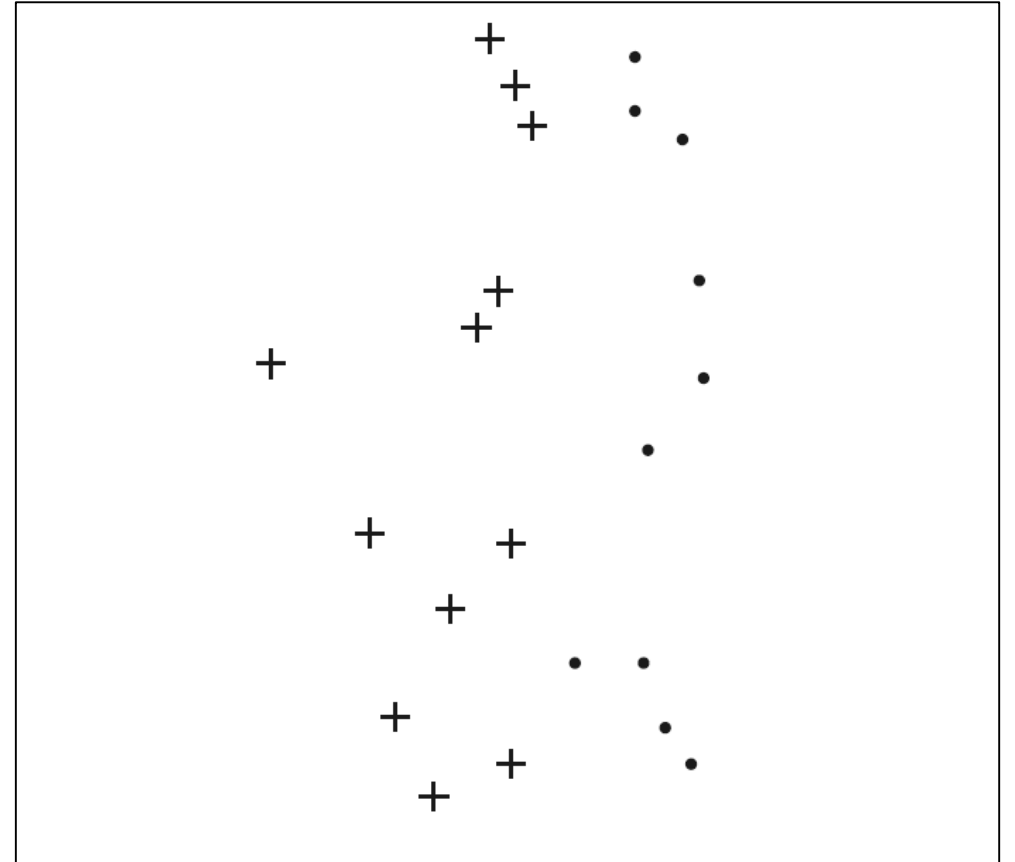
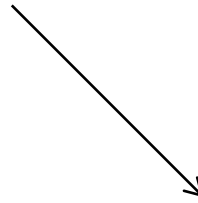
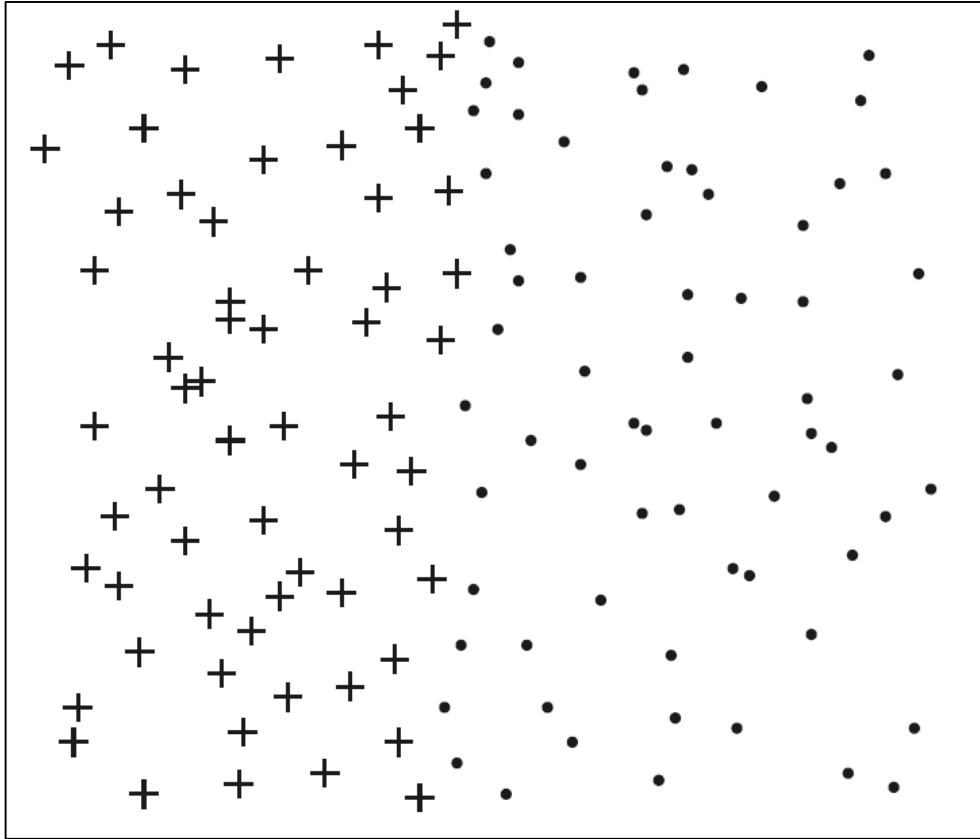
Начинаем с полного набора обучающих точек.

Отсортируем точки по расстоянию до ближайшей точки другого класса.

Идем от наименьшего расстояния к наибольшему (DROP5).

Удалить точку P , если среди точек из полного набора, у которых P была ближайшим соседом, количество правильно классифицированных не уменьшится.

DROP5



Неравномерные признаки:

$$\rho(x, x_i) = \left(\sum_{j=1}^n w_j |x_j - x_{ij}|^p \right)^{\frac{1}{p}}$$

Метрика Минковского

Роль весов w :

1. Нормировка
2. Степень важности
3. Отбор

Добавление признаков

1. Выбираем один лучший признак k : $\rho_k(x_i, x_j) = |x_{jk} - x_{ik}|$
2. Есть расстояние ρ .
3. Добавляем признак k' :

$$\rho(x_i, x_j) = \rho(x_i, x_j) + w_{k'} \rho_{k'}(x_i, x_j)$$

4. Можно заменять признаки:

$$\rho(x_i, x_j) = \rho(x_i, x_j) - w_{k''} \rho_{k''}(x_i, x_j) + w_{k'} \rho_{k'}(x_i, x_j)$$

Будем добавлять, пока LOO уменьшается.

Быстрый поиск соседей k-d tree

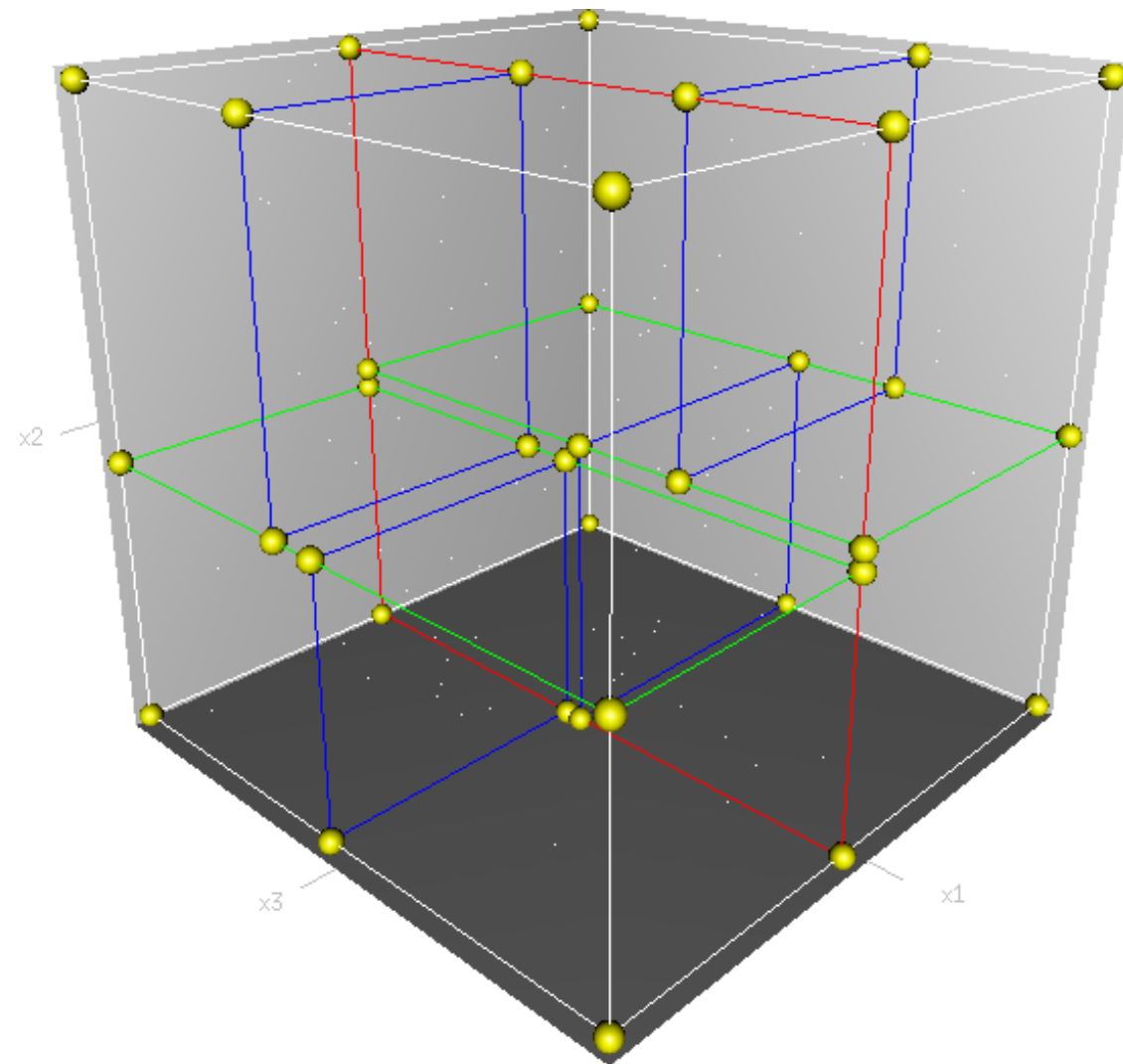
Разбиваем пространство гиперплоскостями, ортогональными одной из координатных осей, последовательно по медианным точкам.

Получаем участки с близким количеством точек в них (листья дерева).

Поиск соседей начинаем с точек листа, в котором находится точка, если соседей недостаточно – поднимаемся на узел выше.

Возможны ошибки, но их можно избежать, если следить, насколько близко точка лежит к границе листа.

На картинке слева – дерево глубины 3, делящее пространство на 8 листов (сначала по красной гиперплоскости, потом по зеленым гиперплоскостям, потом по синим).



Домашнее задание

Будет появляться в папке курса в субботу после лекции.

Мягкий дедлайн (на полный балл) – в следующую субботу,

Жесткий дедлайн (на половинный балл) – еще через неделю.

На паре после лекции можно будет сдавать домашнее задание и задавать вопросы, если что-то не получается.

Ближайшее домашнее задание появится в папке 11 марта, мягкий дедлайн – 18 марта, жесткий – 25 марта.

9 марта занятия не будет.