Детальное математическое решение задачи классификации Iris с помощью случайного леса

Анализ машинного обучения

1 Постановка задачи

1.1 Исходные данные

- **Датасет**: Iris, $n = 150$ наблюдений
- **Признаки**: $x = (x_1, x_2, x_3, x_4)$
  - x_1: длина чашелистика (sepal length)
  - x_2: ширина чашелистика (sepal width)
  - x_3: длина лепестка (petal length)
  - x_4: ширина лепестка (petal width)
- **Целевые классы**: $y \in \{0, 1, 2\}$ (Setosa, Versicolor, Virginica)

1.2 Разделение данных

$X_{\text{train}} = \{(x_i, y_i)\}_{i=1}^{105} \quad (70\%)$
$X_{\text{test}} = \{(x_i, y_i)\}_{i=106}^{150} \quad (30\%)$
Stratified split: Соотношение классов сохраняется

2 Базовый алгоритм: Дерево решений

2.1 Функция предсказания одного дерева

Для дерева с параметрами θ (глубина, критерий разделения):

$$f(x; \theta) = \sum_{j=1}^{J} c_j \cdot \mathbb{I}(x \in R_j)$$

где:

- J - количество листьев (терминальных узлов)
- R_j - область пространства признаков, соответствующая j-му листу
- c_j - предсказанный класс для листа j

2.2 Критерий Джини - подробное решение

Для узла S с n_S наблюдениями:

$$G(S) = 1 - \sum_{k=1}^{3} (p_k)^2$$

Пример расчета для узла с 30 наблюдениями:

- Setosa: 15 наблюдений $\Rightarrow p_1 = \frac{15}{30} = 0.5$
- Versicolor: 10 наблюдений $\Rightarrow p_2 = \frac{10}{30} = 0.333$
- Virginica: 5 наблюдений $\Rightarrow p_3 = \frac{5}{30} = 0.167$

$$\begin{aligned} G(S) &= 1 - (0.5^2 + 0.333^2 + 0.167^2) \\ &= 1 - (0.25 + 0.111 + 0.028) \\ &= 1 - 0.389 = 0.611 \end{aligned}$$

2.3 Прирост информации (Information Gain)

Для разделения узла S на S_{left} и S_{right} по признаку j с порогом t:

$$\Delta G = G(S) - \left(\frac{n_{\text{left}}}{n_S} G(S_{\text{left}}) + \frac{n_{\text{right}}}{n_S} G(S_{\text{right}}) \right)$$

Пример расчета:

- S: $n_S = 30$, $G(S) = 0.611$
- S_{left}: $n_{\text{left}} = 18$, $G(S_{\text{left}}) = 0.2$
- S_{right}: $n_{\text{right}} = 12$, $G(S_{\text{right}}) = 0.1$

$$\begin{aligned} \Delta G &= 0.611 - \left(\frac{18}{30} \times 0.2 + \frac{12}{30} \times 0.1 \right) \\ &= 0.611 - (0.6 \times 0.2 + 0.4 \times 0.1) \\ &= 0.611 - (0.12 + 0.04) = 0.611 - 0.16 = 0.451 \end{aligned}$$

3 Случайный лес: математическая модель

3.1 Ансамбль деревьев

Случайный лес состоит из $B = 200$ деревьев:

$$F(x) = \{f^{(1)}(x), f^{(2)}(x), \ldots, f^{(B)}(x)\}$$

3.2 Bagging (Bootstrap Aggregating)

Для каждого дерева $b \in \{1, 2, \ldots, B\}$:

Формирование bootstrap выборки:

$$D^{(b)} = \{(x_i^{(b)}, y_i^{(b)})\}_{i=1}^{105} \sim \text{Uniform}(D_{\text{train}})$$

Вероятность включения наблюдения в bootstrap выборку:

$$P(\text{наблюдение } i \in D^{(b)}) = 1 - \left(1 - \frac{1}{105}\right)^{105} \approx 1 - e^{-1} \approx 0.632$$

3.3 Случайный выбор признаков

На каждом узле выбираем m случайных признаков из $p = 4$:

$$m = \lfloor \sqrt{p} \rfloor = \lfloor \sqrt{4} \rfloor = 2$$

Вероятность выбора конкретного признака:

$$P(\text{признак } j \text{ выбран}) = \frac{m}{p} = \frac{2}{4} = 0.5$$

4 Процесс предсказания для тестового примера

4.1 Входные данные

Рассмотрим тестовый объект:

$$x_{\text{test}} = (5.1, 3.5, 1.4, 0.2)$$

4.2 Голосование деревьев

Для каждого дерева $f^{(b)}$ получаем предсказание:

$$f^{(b)}(x_{\text{test}}) \in \{0, 1, 2\}$$

Подсчитываем голоса:

$$\text{Count}(0) = \sum_{b=1}^{200} \mathbb{I}(f^{(b)}(x_{\text{test}}) = 0) = 195$$

$$\text{Count}(1) = \sum_{b=1}^{200} \mathbb{I}(f^{(b)}(x_{\text{test}}) = 1) = 5$$

$$\text{Count}(2) = \sum_{b=1}^{200} \mathbb{I}(f^{(b)}(x_{\text{test}}) = 2) = 0$$

4.3 Вероятности классов

$$P(y = 0|x) = \frac{195}{200} = 0.975$$

$$P(y = 1|x) = \frac{5}{200} = 0.025$$

$$P(y = 2|x) = \frac{0}{200} = 0$$

4.4 Финальное предсказание

$$\hat{y} = \arg\max_{k\in\{0,1,2\}} \text{Count}(k) = \arg\max\{195, 5, 0\} = 0$$

5 Теоретическое обоснование

5.1 Разложение ошибки ансамбля

Общая ошибка случайного леса:

$$\text{Error}(F) = \text{Bias}^2 + \text{Var}(F) + \epsilon$$

где ϵ - неустранимая ошибка.

5.2 Дисперсия ансамбля

$$\text{Var}(F(x)) = \rho\sigma^2 + \frac{1-\rho}{B}\sigma^2$$

Подробный расчет для наших параметров:

- $\sigma^2 = 0.1$ (дисперсия одного дерева)
- $\rho = 0.2$ (средняя корреляция между деревьями)
- $B = 200$

$$\begin{aligned}\text{Var}(F) &= 0.2 \times 0.1 + \frac{1-0.2}{200} \times 0.1 \\ &= 0.02 + \frac{0.8}{200} \times 0.1 \\ &= 0.02 + 0.004 \times 0.1 = 0.02 + 0.0004 = 0.0204\end{aligned}$$

Сравнение с одним деревом: $\frac{0.0204}{0.1} = 0.204$ - дисперсия уменьшена в 5 раз.

6 Оценка качества модели

6.1 Кросс-валидация

Используем RepeatedStratifiedKFold:

- $K = 10$ фолдов
- $R = 3$ повторения
- Всего: $10 \times 3 = 30$ оценок

6.2 Формула accuracy

$$\text{Accuracy} = \frac{1}{K \cdot R} \sum_{r=1}^{R} \sum_{k=1}^{K} \left(\frac{1}{n_k} \sum_{j=1}^{n_k} \mathbb{I}(\hat{y}_j^{(k,r)} = y_j^{(k,r)}) \right)$$

Пример расчета для одного фолда:

- $n_k = 15$ наблюдений в тестовом фолде
- Правильные предсказания: 14

$$\text{Accuracy}_{\text{fold}} = \frac{14}{15} = 0.933$$

6.3 Итоговая оценка

После 30 запусков:

$$\begin{aligned} \text{Accuracy}_{\text{mean}} &= 0.967 \\ \text{Accuracy}_{\text{std}} &= 0.025 \\ 95\% \text{ доверительный интервал} &= 0.967 \pm 1.96 \times \frac{0.025}{\sqrt{30}} = 0.967 \pm 0.009 \end{aligned}$$

7 Важность признаков

7.1 Формула важности

Для признака j:

$$\text{Importance}(j) = \frac{1}{B} \sum_{b=1}^{B} \sum_{t \in T_b} \Delta\text{Gini}(t, j) \cdot \frac{n_t}{n}$$

где:

- T_b - множество узлов дерева b
- $\Delta\text{Gini}(t, j)$ - уменьшение неопределенности в узле t при использовании признака j
- n_t - количество наблюдений в узле t
- n - общее количество наблюдений

7.2 Пример расчета для petal length

Для одного дерева:

- Узел 1: $\Delta G = 0.3$, $n_t = 105$
- Узел 2: $\Delta G = 0.2$, $n_t = 60$
- Узел 3: $\Delta G = 0.1$, $n_t = 30$

$$\begin{aligned} \text{Imp}_{\text{single}} &= (0.3 \times 105 + 0.2 \times 60 + 0.1 \times 30)/105 \\ &= (31.5 + 12 + 3)/105 = 46.5/105 = 0.443 \end{aligned}$$

Усреднение по 200 деревьям даёт итоговую важность 0.452.

8 Сравнение алгоритмов

8.1 Математическое сравнение

Параметр	Дерево решений	Случайный лес
Bias^2	0.02	0.03
Variance	0.10	0.0204
$\text{Error}_{\text{total}}$	0.12	0.0504
Accuracy	0.933	0.967
Стабильность	Низкая	Высокая

8.2 Обучение vs тестирование

Для дерева решений с увеличением глубины:

$$\text{Accuracy}_{\text{train}} \to 1.0$$
$$\text{Accuracy}_{\text{test}} \to \max \text{ в точке } d = 4, \text{ затем } \downarrow$$

Для случайного леса:

$$\text{Accuracy}_{\text{train}} \approx 0.98 - 0.99$$
$$\text{Accuracy}_{\text{test}} \approx 0.96 - 0.97 \text{ (стабильно)}$$

9 Заключение

Случайный лес демонстрирует превосходство над одним деревом за счет:

1. **Уменьшения дисперсии** в 5 раз: $0.100 \to 0.0204$
2. **Стабильности предсказаний**: $\sigma_{\text{accuracy}} = 0.025$
3. **Устойчивости к переобучению**: test accuracy стабилен при увеличении глубины
4. **Робастности**: работа с шумными данными и пропусками

Математическое решение подтверждает эффективность ансамблирования для задачи классификации ирисов.