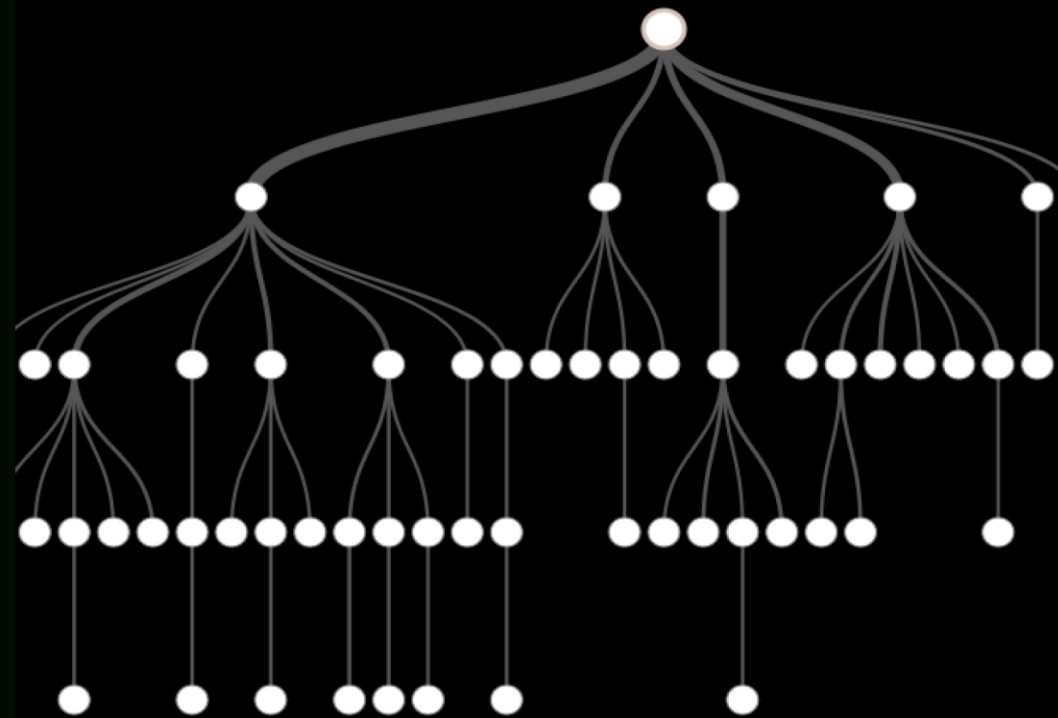
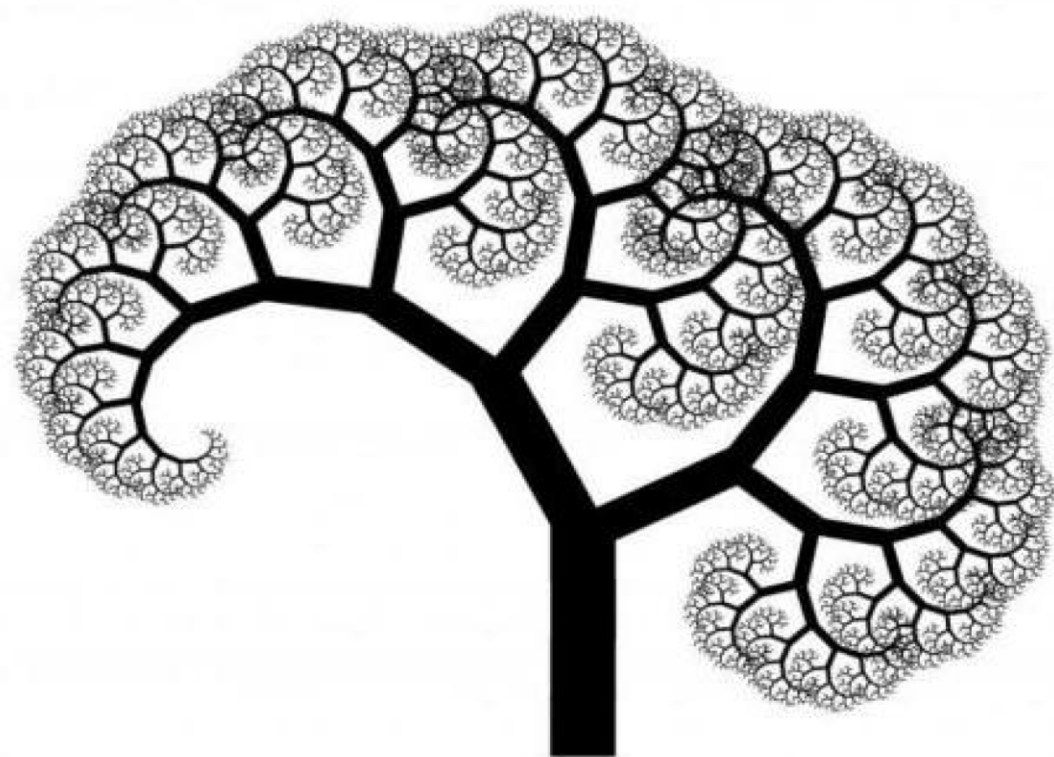


Марія Талах

АНСАМБЛЕВІ АРХІТЕКТУРИ ТА ТЕХНОЛОГІЇ АНАЛІЗУ ДАНИХ



АНСАМБЛЕВІ АРХІТЕКТУРИ ТА ТЕХНОЛОГІЇ АНАЛІЗУ ДАНИХ



КАФЕДРА КОМП'ЮТЕРНИХ НАУК

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ЧЕРНІВЕЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ ЮРІЯ ФЕДЬКОВИЧА

М.В. Талах

АНСАМБЛЕВІ АРХІТЕКТУРИ ТА ТЕХНОЛОГІЇ АНАЛІЗУ ДАНИХ

Навчальний посібник

Механізм

2023

УДК 004.896

T-16

Рекомендовано до друку вченою радою навчально-наукового інституту фізико-технічних та комп'ютерних наук Чернівецького національного університету імені Юрія Федьковича (протокол №11 від 30 грудня 2022 р.)

Рецензент:

Арсирій О.О., д.т.н., завідувач кафедри Інформаційних систем Національного університету «Одеська політехніка»

Талах М.В.

T-16 Ансамблевій архітектурі та технології аналізу даних. /
М.В. Талах – Чернівці: Технодрук, 2023. – 246 с.

Навчальний посібник призначений для студентів, що навчаються за спеціальністю 122 «Комп'ютерні науки», освітньо-професійна програма «Інтелектуальний аналіз даних в комп'ютерних інформаційних системах» за першим (бакалаврським) освітнім рівнем та всіх бажаючих.

У навчальному посібнику викладено дослідженням інноваційних підходів до аналізу даних, що базуються на використанні ансамблевих моделей та алгоритмів.

УДК 004.896

©Чернів. нац. ун-т, 2023
©ПВКФ «Технодрук», 2023
©Талах М.В., 2023

ЗМІСТ

ВСТУП	7
<u>1. АНСАМБЛЕВІ МЕТОДИ: ВИЗНАЧЕННЯ ТА ФОРМУЛЮВАННЯ ПРОБЛЕМИ АНСАМБЛЕВОГО НАВЧАННЯ</u>	8
1.1 АНСАМБЛЕВІ МЕТОДИ: МУДРІСТЬ НАТОВПУ	9
1.2 ЧОМУ НАМ МАЄ БУТИ ЦІКАВИМ АНСАМБЛЕВЕ НАВЧАННЯ	12
1.3 НАВЧАННЯ ТА СКЛАДНІСТЬ В ОДИНОЧНИХ МОДЕЛЯХ	14
1.3.1 НА ОСНОВІ ДЕРЕВ РІШЕНЬ	14
1.3.2. НАВЧАННЯ ПРОТИ СКЛАДНОСТІ В ДЕРЕВАХ РІШЕНЬ	15
1.3.3. РЕГРЕСІЯ ЗА ДОПОМОГОЮ МЕТОДУ ОПОРНИХ ВЕКТОРІВ	20
1.3.4. ПІДГОНКА ПРОТИ СКЛАДНОСТІ В ОПОРНИХ ВЕКТОРНИХ МАШИНАХ	22
1.4. КОМПРОМІС: ЗМІЩЕННЯ-ДИСПЕРСІЯ	23
1.5 НАШ ПЕРШИЙ АНСАМБЛЬ	25
1.6 ТЕРМІНОЛОГІЯ ТА ТАКСОНОМІЯ ДЛЯ АНСАМБЛЕВИХ МЕТОДІВ	30
<u>2 ОДНОРІДНІ ПАРАЛЕЛЬНІ АНСАМБЛІ: БЕГІНГ ТА ВИПАДКОВИЙ ЛІС</u>	33
2.1 ПАРАЛЕЛЬНІ АНСАМБЛІ	34
2.2 БЕГІНГ: БУТСТРАП АГРЕГАЦІЯ	35
2.2.1 ПОВТОРНА ВИБІРКА ТА АГРЕГАЦІЯ МОДЕЛЕЙ	37
2.2.2 РЕАЛІЗАЦІЯ БЕГІНГУ	40
2.2.3 БЕГІНГ НА ОСНОВІ SCIKIT-LEARN	43
2.2.4 ШВИДШЕ НАВЧАННЯ З РОЗПАРАЛЕЛЮВАННЯМ	45
2.3 ВИПАДКОВИЙ ЛІС	46
2.3.1 РАНДОМІЗОВАНІ ДЕРЕВА РІШЕНЬ	46
2.3.2 ВИПАДКОВІ ЛІСИ З SCIKIT-LEARN	48
2.3.3 ВАЖЛИВІСТЬ ОЗНАК	49
2.4 БІЛЬШ ОДНОРІДНІ ПАРАЛЕЛЬНІ АНСАМБЛІ	51
2.4.1 «ВСТАВКА» (PASTING)	51
2.4.2 ВИПАДКОВІ ПІДМНОЖИНИ ТА ВИПАДКОВІ ПАТЧІ	52
2.4.3. EXTRA TREES (ЕКСТРИМАЛЬНО РАНДОМІЗОВАНІ ДЕРЕВА)	54
2.5 НАВЧАЛЬНИЙ ПРИКЛАД: ДІАГНОСТИКА РАКУ ГРУДЕЙ	55
2.5.1. РОЗМІР АНСАМБЛЮ ПРОТИ ШВИДКОСТІ АНСАМБЛЮ	57
2.5.2. СПІВВІДНОШЕННЯ БАЗОВОЇ СКЛАДНОСТІ УЧНЯ ТА ПРОДУКТИВНОСТІ АНСАМБЛІВ	59

3. НЕОДНОРІДНІ ПАРАЛЕЛЬНІ АНСАМБЛІ: ОБ'ЄДНАННЯ СИЛЬНИХ УЧНІВ 64

3.1	БАЗОВІ ОЦІНЮВАЧІ ТА НЕОДНОРІДНІ АНСАМБЛІ	66
3.1.1	ТРЕНУВАННЯ БАЗОВИХ ОЦІНЮВАЧІВ	67
3.1.2	ІНДИВІДУАЛЬНІ ПЕРЕДБАЧЕННЯ БАЗОВИХ ОЦІНЮВАЧІВ	69
3.2	ОБ'ЄДНАННЯ ПРОГНОЗІВ ШЛЯХОМ ЗВАЖУВАННЯ	73
3.2.1	ГОЛОСУВАННЯ БІЛЬШІСТЮ	75
3.2.2	ТОЧНІСТЬ ЗВАЖУВАННЯ	76
3.2.3	ЕНТРОПІЙНЕ ЗВАЖУВАННЯ	79
3.2.4	КОМБІНАЦІЯ ДЕМПСТЕРА-ШЕЙФЕРА	82
3.3	КОМБІНУВАННЯ ПРОГНОЗІВ ЗА ДОПОМОГОЮ МЕТА-НАВЧАННЯ	85
3.3.1	СТЕКІНГ	86
3.3.2	СТЕКІНГ З ПЕРЕХРЕСНОЮ ВАЛІДАЦІЄЮ	92
3.4	НАВЧАЛЬНИЙ ПРИКЛАД: АНАЛІЗ НАСТРОЇВ	96
3.4.1	ПОПЕРЕДНЯ ПІДГОТОВКА	97
3.4.2	ЗМЕНШЕННЯ РОЗМІРНОСТІ	101
3.4.3	СТЕКІНГОВИЙ КЛАСИФІКАТОР	103

4. ПОСЛІДОВНІ АНСАМБЛІ: БУСТИНГ 107

4.1	ПОСЛІДОВНІ АНСАМБЛІ З СЛАБКИМИ УЧНЯМИ	109
4.2	ADABoOST: АДАПТИВНИЙ БУСТИНГ	111
4.2.1	ІНТУІЦІЯ: НАВЧАННЯ З ЗВАЖЕНИМИ ЗРАЗКАМИ	111
4.2.2	РЕАЛІЗАЦІЯ ADABoOST	115
4.2.3	ADABoOST ІЗ SCIKIT-LEARN	122
4.3	ADABoOST НА ПРАКТИЦІ	124
4.3.1	ШВИДКІСТЬ НАВЧАННЯ	126
4.3.2	РАННЯ ЗУПИНКА ТА ОБРІЗКА	128
4.4	НАВЧАЛЬНИЙ ПРИКЛАД: РОЗПІЗНАВАННЯ РУКОПИСНИХ ЦИФР	131
4.4.1	ЗМЕНШЕННЯ РОЗМІРНОСТІ ЗА ДОПОМОГОЮ T-SNE	132
4.4.2	БУСТИНГ	135
4.5	LOGITBoOST: БУСТИНГ З ЛОГІСТИЧНИМИ ВТРАТАМИ	138

5. ПОСЛІДОВНІ АНСАМБЛІ: ГРАДІЄНТНИЙ БУСТИНГ 144

5.1 Градієнтний спуск для мінімізації	145
5.1.1 Градієнтний спуск із наочним прикладом	147
5.1.2 Імплементация та пояснення градієнтного спуску	148
5.1.3 Властивості градієнтного спуску	152
5.1.4 Градієнтний спуск через функцію втрат для навчання	153
5.2 Градієнтний бустинг: градієнтний спуск + бустинг	157
5.2.1 Інтуїція: навчання із залишками	158
5.2.2 AdaBoost проти градієнтного бустингу	158
5.2.3 Від слабких учнів до приблизних градієнтів	160
5.2.4 Градієнтний бустинг це градієнтний спуск + бустинг	161
5.2.5 Імплементация градієнтного бустингу	163
5.2.6 Візуалізація ітерацій градієнтного бустингу.	165
5.2.7 Градієнтний бустинг з scikit-learn	169
5.2.8 Градієнтний бустинг на основі гістограм	171
5.3 LightGBM: фреймворк для градієнтного бустингу	173
5.3.1 Що робить LightGBM «легким» ?	174
5.3.2 Одностороння вибірка на основі градієнта (GOSS)	175
5.3.3 Exclusive Feature Bundling (EFB)	176
5.3.4 Градієнтний бустинг з LightGBM	177
5.4 LightGBM на практиці	178
5.4.1 Швидкість навчання	179
5.4.2 Швидкість навчання через перехресну валідацію	179
5.4.3 Перехресна валідація з LightGBM	181
5.4.4 Рання зупинка	183
5.4.5 Спеціальні функції втрати	185
5.4.6 Фокусні втрати	186
5.4.7 Градієнтний бустинг з фокусними втратами	188
5.5 Навчальний приклад: пошук документів	190
5.5.1 Набір даних Letor	190
5.5.2 Пошук документів з LightGBM	192

6. ПОСЛІДОВНІ АНСАМБЛІВ: БУСТИНГ НЬЮТОНА 197

6.1 Ньютонівський метод для мінімізації	198
6.1.1 Друга похідна та матриця Гессі	200
6.1.2 Метод Ньютона з наочним прикладом	201
6.1.3 Властивості Ньютонівського спуску	206
6.1.4 Ньютонівський спуск через функцію втрат для навчання	207
6.2 Ньютонівський бустинг: метод Ньютона + бустинг	211

6.2.1 Інтуїція: навчання зі зваженими залишками	211
6.2.2 Ньютонівський бустинг це ньютонівський спад + бустинг	211
6.2.3 Як працюють вставки ГЕССА?	213
6.2.4 Інтуїція: навчання з регуляризованими функціями втрат	216
6.2.5 Реалізація ньютонівського бустингу	219
6.2.6 Візуалізація ітерацій градієнтного бустингу	222
6.3 XGBoost: Фреймворк для бустингу Ньютона	224
6.3.1 Що робить XGBoost «екстримальним»?	226
6.3.2 Регуляризовані функції втрат для навчання	226
6.3.3 (Зважений) квантильний ньютонівський бустинг	227
6.3.4 Ньютонівський бустинг з XGBoost	229
6.4 XGBoost на практиці	231
6.4.1 Швидкість навчання	231
6.4.2. Швидкість навчання через перехресну перевірку	232
6.4.3. Перехресна перевірка з XGBoost	234
6.4.4 Рання зупинка	235
6.5 Навчальний приклад: Пошук документів	236
6.5.1 Набір даних LETOR	236
6.5.2 Пошук документів з XGBoost	237

<u>ВИСНОВОК</u>	<u>241</u>
<u>ЛІТЕРАТУРА</u>	<u>242</u>

Навчальне видання

Талах Марія Віталіївна

Ансамблеві архітектури та технології аналізу даних

Навчально-методичний посібник

Літературний редактор: О. В. Лупул

Папір офсетний. Формат 60x84/16.

Умов. друк. арк.. 14,36. Обл.- вид. арк. 15,44. Тираж – 50.

Видавець та виготівник: ПБКФ «Технодрук»

Свідоцтво суб'єкта видавничої справи ДК №1841 від 10.06.2004 р.
58000, м. Чернівці, вул. І. Франка, 20, оф.18, тел. (0372) 55-05-85