

## Частина 1. Лабораторна робота №2, Алгоритми класифікації

### 1. ЗАГАЛЬНІ ВІДОМОСТІ

Даний документ містить інформацію щодо лабораторної роботи №2 з алгоритмів машинного навчання. Всього передбачено 12 варіантів. Ваш варіант відповідає Вашому номеру у списку з журналу (якщо Ваш номер більше 12, то відніміть від Вашого номера 12).

Разом з цим документом надіслано архів що містить, файли з даними, які слід аналізувати, кожному варіанту відповідає два файли - один з розширенням .data - містить дані, другий з розширенням .names, містить інформацію про колонки. У кожному варіанті буде вказана змінна - відгук, в якості регресорів слід узяти всі числові змінні. Самі дані взяті з сайту <https://archive.ics.uci.edu/ml/index.php> (це досить відомий сайт, що містить багато інформації та датасетів для машинного навчання).

Для проведення **класифікації** потрібно зробити наступне:

1) Опрацювати дані на предмет пропущених значень. Якщо можливо то видалити рядки із пропущеними значеннями, якщо ж їх занадто багато то замінити пропущені значення середніми по регресору.

2) Побудувати наївний байєсівський класифікатор.

3) Провести класифікацію методом  $k$ -nn, спробуйте різні значення  $n$ .

4) Провести класифікацію за допомогою логістичної регресії.

5) Додайте до моделі факторні дані. Для цього можна скористатися різними техніками: присвоїти кожному фактору деяке число і проводити звичайний аналіз (для яких методів цей підхід може спрацювати краще?), інший підхід полягає у тому щоб розбити вибірку на підвибірки по кожному значенню фактора і аналізувати підвибірки окремо.

6) Напишіть висновок у якому Ви порівняєте різні алгоритми класифікації та опишіть отримані результати

Також пропонується додаткове завдання. Завдання є опціональним і виконується за бажанням студента.

7) Скачайте датасет

<https://www.kaggle.com/team-ai/spam-text-message-classification>

що містить набір повідомлень (спамових або ні) та побудуйте спам-фільтр на основі наївного байєсівського класифікатора.

[https://en.wikipedia.org/wiki/Naive\\_Bayes\\_spam\\_filtering](https://en.wikipedia.org/wiki/Naive_Bayes_spam_filtering).

### 2. ВИМОГИ ДО ВИКОНАННЯ, ОФОРМЛЕННЯ ТА ЗДАЧІ

Роботу можна виконувати в R або Python. Вибірку слід розбити на тестову та тренувальну у пропорціях 20% та 80%.

Робота має бути оформлена у вигляді .pdf (можливо .doc/.docx) файлу який містить всю необхідну інформацію. Роботи потрібно здавати на парах або надсилати на email.

Кожна робота буде розглядатися на відповідність критеріям описаним вище, та на обґрунтованість прийнятих рішень. Кожен студент, повинен виконати свою роботу самостійно. Ідентичні, або майже ідентичні роботи прийматися до уваги не будуть.

### 3. ВАРІАНТИ

#### Варіант 1

URL: <https://archive.ics.uci.edu/ml/datasets/Abalone>

Відгук: Rings, розбийте на три групи  $< 8$ ,  $[8, 12]$ ,  $> 12$

Факторна змінна: Sex

#### Варіант 2

URL: <https://archive.ics.uci.edu/ml/datasets/Adult>

Відгук: Остання колонка

Факторна змінна: Education

#### Варіант 3

URL: <https://archive.ics.uci.edu/ml/datasets/Arrhythmia>

Відгук: Остання колонка

Регресори: Колонки 1, 4, 5-15

Факторна змінна: Sex

#### Варіант 4

URL: <https://archive.ics.uci.edu/ml/datasets/Credit+Approval>

Дані: `crx.data`, `crx.names`

Відгук: Остання колонка

Факторна змінна: A13

#### Варіант 5

URL: <https://archive.ics.uci.edu/ml/datasets/Contraceptive+Method+Choice>

Дані: `cmc.data`, `cmc.names`

Відгук: Остання колонка

Факторна змінна: Колонка 8

#### Варіант 6

URL: <https://archive.ics.uci.edu/ml/datasets/Cylinder+Bands>

Відгук: Остання колонка

Факторна змінна: Колонка 2

#### Варіант 7

URL: <https://archive.ics.uci.edu/ml/datasets/Echocardiogram>

Відгук: Остання колонка

Факторна змінна: Колонка 4

Коментар: Зверніть увагу на опис даних, деякі рядки мають бути видалені!

#### Варіант 8

URL: <https://archive.ics.uci.edu/ml/datasets/Flags>

Відгук: religion, колонка 7

Факторна змінна: language

#### Варіант 9

URL: <https://archive.ics.uci.edu/ml/datasets/Glass+Identification>

Відгук: Остання колонка

Факторна змінна: Fe, колонка 10, перетворіть її на фактор з двома значеннями

- менше середнього та більше середнього

**Варіант 10**

**URL:** <https://archive.ics.uci.edu/ml/datasets/Haberman%27s+Survival>

**Відгук:** Остання колонка

**Факторна змінна:** Перетворіть на фактор колонку Age - до 30 років, 30-60, більше 60

**Варіант 11**

**URL:** <https://archive.ics.uci.edu/ml/datasets/Heart+Disease>

**Дані:** `processed.cliveland.data`, `heart-disease.names`

**Відгук:** Остання колонка

**Факторна змінна:** Колонка 2

**Варіант 12**

**URL:** <https://archive.ics.uci.edu/ml/datasets/Hepatitis>

**Відгук:** Перша колонка

**Факторна змінна:** Комбінація 9 та 10, 4 можливих значення