**Лабораторна робота 1.3: Логістична регресія для задач класифікації**

**Варіант 2: Розпізнавання рукописних цифр (MNIST Digits)**

**Опис**: Класифікація рукописних цифр від 0 до 9.

**Джерело даних**: [MNIST Dataset](https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_digits.html)

**Код для завантаження**:

from sklearn.datasets import load\_digits

data = load\_digits()

X = data.data

y = (data.target == 0).astype(int) # Класифікація "0" vs "не 0"

**Питання для самоперевірки**

**1.** Лінійна регресія передбачає неперервні значення і застосовується для задач регресії, де залежна змінна має числовий характер. **Логістична регресія** передбачає ймовірності для категоричних змінних і застосовується для задач класифікації, де залежна змінна має дискретний, зазвичай бінарний, характер.

**2.** Сигмоїдна функція в логістичній регресії виконує перетворення лінійної комбінації вхідних змінних в ймовірність, яка завжди лежить між 0 і 1. Це важливо, оскільки задача логістичної регресії полягає в класифікації, а результат, що виходить з моделі, має бути ймовірністю належності до одного з класів. Сигмоїдна функція забезпечує саме таке обмеження, оскільки її значення завжди знаходяться в межах від 0 до 1, що можна інтерпретувати як ймовірність того, що даний зразок належить до класу 1. Таким чином, вона дозволяє моделі приймати рішення, чи належить об'єкт до певного класу, залежно від порогу ймовірності.

**3.** Для логістичної регресії не використовується функція вартості, як у лінійній регресії, через різницю в природі задач. У лінійній регресії задача полягає в передбаченні неперервних значень, і функція вартості, яка використовується в цій моделі (середньоквадратична похибка), оптимізує різницю між реальними і передбаченими значеннями, що має смисл для числових даних. Однак у логістичній регресії ми маємо справу з класифікацією, де передбачуване значення є ймовірністю (між 0 і 1), і класичні функції вартості, що використовуються в регресії, не підходять, оскільки вони не враховують природу ймовірності. Тому для логістичної регресії застосовується інша функція вартості, яка називається логістичною функцією вартості, або лог-лікельхуд. Вона спеціально розроблена для оптимізації ймовірностей, що дає змогу краще відповідати вимогам класифікаційних задач.

**4.** Запобігання перенавчанню (overfitting) в логістичній регресії можна досягти за допомогою кількох методів:

1. **Регуляризація**. Один з найпоширеніших способів запобігання перенавчанню — це додавання регуляризаційного члена до функції вартості. Це допомагає обмежити величину коефіцієнтів моделі, що зменшує ймовірність надмірної підгонки моделі під тренувальні дані. Є два основних види регуляризації:
   * **L1-регуляризація (Lasso)**, яка намагається зменшити деякі коефіцієнти до нуля, таким чином автоматично відбираючи важливі ознаки.
   * **L2-регуляризація (Ridge)**, яка заохочує маленькі значення коефіцієнтів, але не робить їх рівними нулю.
2. **Розподіл даних на тренувальні та тестові вибірки**. Зазвичай дані ділять на тренувальну та тестову вибірки, що дозволяє перевірити, чи добре модель узагальнюється на нові дані. Важливо мати достатньо різноманітні дані в тренувальному наборі, щоб уникнути перенавчання на одному конкретному наборі.
3. **Крос-валідація**. Використання методів крос-валідації, таких як k-fold крос-валідація, допомагає оцінити стабільність і узагальненість моделі. Крос-валідація дає змогу перевірити модель на різних підмножинах даних і зменшити ймовірність перенавчання.
4. **Підбір параметрів моделі**. Для боротьби з перенавчанням важливо оптимізувати параметри моделі, зокрема параметри регуляризації (λ). Занадто велике значення λ може призвести до недонавчання, а занадто мале — до перенавчання.
5. **Зменшення складності моделі**. Якщо модель має занадто багато параметрів, це може призвести до перенавчання. Зменшення кількості ознак, використання технік відбору ознак або застосування методів зменшення розмірності, таких як PCA, можуть допомогти зменшити складність моделі.
6. **Раннє припинення (Early stopping)**. Якщо модель навчається дуже довго і починає показувати погіршення на тестових даних, можна використовувати метод раннього припинення, щоб зупинити навчання, коли ефективність на тестових даних починає знижуватися.

**5.**

Параметр регуляризації λ\lambdaλ має значний вплив на модель логістичної регресії, оскільки він контролює ступінь регуляризації, що додається до функції вартості. Регуляризація допомагає запобігти перенавчанню, обмежуючи величину коефіцієнтів моделі. Ось як зміна значення λ\lambdaλ може вплинути на модель:

1. **Мале значення λ\lambdaλ**:
   * Якщо λ\lambdaλ дуже мале (або рівне нулю), регуляризація майже не впливає на модель, і вона може надмірно підлаштовуватися під тренувальні дані (перенавчання). Це може призвести до того, що модель буде занадто складною і погано узагальнюватиметься на нові дані.
   * В такому випадку модель може дуже точно відтворювати тренувальні дані, але з поганими результатами на тестових даних.
2. **Велике значення λ\lambdaλ**:
   * Якщо λ\lambdaλ дуже велике, регуляризація стане домінуючим фактором і значно зменшить величину коефіцієнтів моделі, що може призвести до **недонавчання (underfitting)**. Модель стає занадто простою і не здатна відображати складність даних.
   * Це може призвести до зниження якості прогнозів як на тренувальних, так і на тестових даних, оскільки модель не здатна точно відображати залежності в даних.
3. **Оптимальне значення λ\lambdaλ**:
   * Ідеальне значення λ\lambdaλ — це компроміс між перенавчанням і недонавчанням. Воно дозволяє регуляризації обмежити модель, не зменшуючи її здатність до узагальнення. Параметр λ\lambdaλ слід налаштовувати за допомогою методів, таких як **крос-валідація**, щоб знайти оптимальне значення для даного набору даних.

**7.** **Матриця помилок (confusion matrix)** — це інструмент для оцінки якості моделі класифікації, який показує, скільки правильних і неправильних передбачень модель зробила для кожного класу. Вона допомагає детально проаналізувати, як модель працює з різними класами і де виникають помилки.

**8.** **Логістична регресія** — це один з основних методів для класифікації в машинному навчанні, і вона має певні переваги та недоліки порівняно з іншими методами класифікації, такими як дерева рішень, SVM (метод опорних векторів), або нейронні мережі.

**9.** Логістична регресія є ефективним методом класифікації, але її доцільно використовувати в певних ситуаціях, а в інших випадках можуть бути більш ефективними інші методи класифікації. Ось коли краще використовувати логістичну регресію та коли варто звернутися до інших методів:

**Коли доцільно використовувати логістичну регресію:**

1. **Лінійно роздільні дані**:  
   Логістична регресія працює добре, коли існує лінійний розподіл між класами, тобто коли дані можна розділити за допомогою прямої або гіперплощини. У таких випадках вона є дуже ефективною і швидкою.
2. **Інтерпретованість моделі важлива**:  
   Логістична регресія є прозорою та легко інтерпретованою моделлю. Якщо потрібно зрозуміти, як кожна ознака впливає на ймовірність належності до певного класу, логістична регресія буде чудовим вибором.
3. **Малі та середні набори даних**:  
   Якщо кількість спостережень не надто велика, логістична регресія є швидким та ефективним методом, оскільки її обчислення не є обтяжливим для ресурсів.
4. **Ймовірнісні прогнози**:  
   Логістична регресія видає ймовірності того, до якого класу належить спостереження. Якщо важливо отримати ймовірнісні оцінки, а не лише категорії, логістична регресія буде дуже корисною.
5. **Проблеми з двома класами (бінарні задачі)**:  
   Логістична регресія найкраще працює для бінарних класифікаційних завдань, де потрібно передбачити два можливі класи (наприклад, позитивний/негативний, 1/0).

**10.**

Коефіцієнти моделі логістичної регресії відображають вплив кожної ознаки на ймовірність належності спостереження до одного з класів. Якщо коефіцієнт додатний, збільшення ознаки підвищує ймовірність належності до класу 1, якщо від'ємний — зменшує цю ймовірність. Інтерпретація коефіцієнтів стає зрозумілішою, якщо використовувати експоненціальний коефіцієнт eβje^{\beta\_j}eβj​, який показує, на скільки змінюються шанси належності до класу 1 при одиничній зміні ознаки. Це дозволяє оцінити, чи збільшує чи зменшує певна ознака ймовірність належності до класу 1. Наприклад, коефіцієнт β1=0.5\beta\_1 = 0.5β1​=0.5 означає, що при збільшенні ознаки на одиницю шанси належності до класу 1 зростають.