Методи машинного навчання для обробки медичних даних

Студентка групи ДА-61
Материнська Софія

Науковий керівник
ас. Яременко Вадим Сергійович

Об’єкт, предмет

- Методи машинного навчання для обробки медичних даних
- Використання методів машинного навчання(метод наївного Баєса, метод к-найближчих сусідів, дерево рішень, логістична регресія, метод опорних векторів, “random forest”) зокрема штучних нейронних мереж для вирішення задачі класифікації даних

Мета, завдання

- Мета – дослідження методів машинного навчання які дають ефективні результати для обробки медичних даних, а саме тих, що використовуються для задач класифікації
- Завдання – на основі обраних даних побудувати моделі, які будуть давати високі показники ефективності, порівняти їх та зробити висновки

Задача, актуальність

- Задача класифікації
- Причини актуальності використання машинного навчання в сфері медицини:
  1. Доступність цифрових медичних даних.
  2. Перехід медичних закладів на ведення електронного документообігу.
  3. Популярність та широке застосування методів машинного навчання.

Етапи виконання

1. Дослідження сфери
2. Аналіз наявних робіт і методів
3. Вибір даних для побудови моделі
4. Аналіз досліджень на основі цих даних
5. Попередня обробка
6. Побудова моделей обраними методами
7. Висновки з отриманих результатів

1. Дослідження сфери

2. Аналіз наявних робіт і методів
3. Вибір даних
4. Аналіз досліджень на основі цих даних
5. Попередня обробка
6. Побудова моделей
7. Висновки

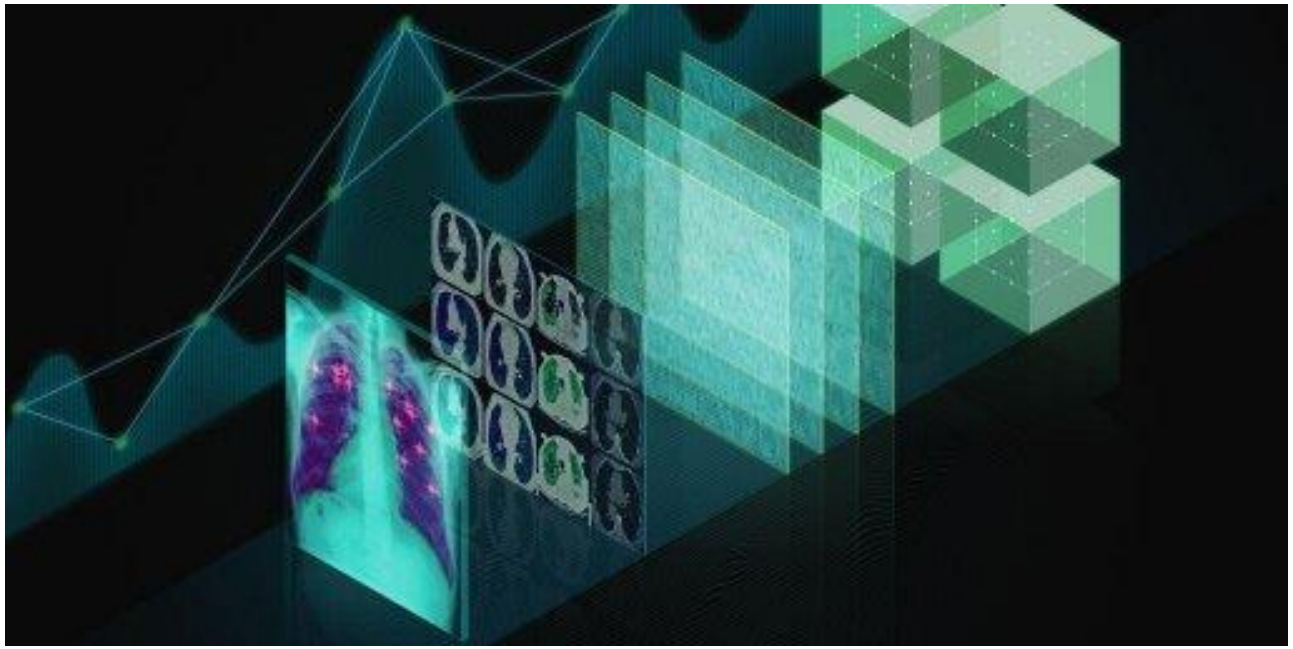
Сфери застосування машинного навчання в медицині:

- Розробка ліків
- Персоналізація лікування
- Діагностика

1. Дослідження сфери

2. Аналіз наявних робіт і методів
3. Вибір даних
4. Аналіз досліджень на основі цих даних
5. Попередня обробка
6. Побудова моделей
7. Висновки

- Виявлення раку легень або інсультів на КТ
- Оцінка ризику раптової серцевої смерті на основі електрокардіограм та МРТ серця
- Класифікація уражень шкіри на зображеннях шкіри
- Знаходження показників діабетичної ретинопатії на зображеннях очних яблук



1. Дослідження сфери

2. Аналіз наявних робіт і методів
3. Вибір даних
4. Аналіз досліджень на основі цих даних
5. Попередня обробка
6. Побудова моделей
7. Висновки

Вхідні дані

- Медичні дані, що знаходяться у вільному доступі
- **18** наукових статей, що стосуються застосування методів машинного навчання для діагностування
- Програмні засоби: Python та допоміжні бібліотеки
- Інформація про процес моделювання та можливі проблеми

1. Дослідження сфери
2. Аналіз наявних робіт і методів
3. Вибір даних
4. Аналіз досліджень на основі цих даних
5. Попередня обробка
6. Побудова моделей
7. Висновки

Аналіз 9 робіт (32 методи) показав, що найпопулярніші методи застосовувались з такою частотою:

Скорочення	Повна назва методу	Кількість застосувань в проаналізованих публікаціях
SVM	Метод опорних векторів	5
NB	Метод наївного Баєса	5
DT	Дерево рішень	5
LR	Логістична регресія	4
RF	“Random forest”	4
ANN	Штучна нейронна мережа	4
KNN	Метод к-найближчих сусідів	3

1. Дослідження сфери
2. Аналіз наявних робіт і методів
3. Вибір даних
4. Аналіз досліджень на основі цих даних
5. Попередня обробка
6. Побудова моделей
7. Висновки

kaggle

data.world

Набір даних	Кількість елементів	Кількість атрибутів
Серцево-судинні захворювання	303	13
Хвороби печінки	583	10

UCI Machine Learning Repository
Center for Machine Learning and Intelligent Systems

1. Дослідження сфери
2. Аналіз наявних робіт і методів
3. Вибір даних
4. Аналіз досліджень на основі цих даних
5. Попередня обробка
6. Побудова моделей
7. Висновки

Методи застосовані в дослідженнях та їх точність

Серцево-судинні захворювання

1. LR - 76%, RF – 80%, DT – 83%
2. ANN – 80,17%

Хвороби печінки

1. KNN, DT, RF – 74,2% , NB
2. LR – 68%
3. SVM - 71%, NB – 56%, DT – 66%

1. Дослідження сфери
2. Аналіз наявних робіт і методів
3. Вибір даних
4. Аналіз досліджень на основі цих даних
5. Попередня обробка
6. Побудова моделей
7. Висновки

- Виділення інформативних рис
  - Статистичний аналіз
- Перевірка відсутності значень
  - Заповнення значенням
    - За замовчуванням
    - ***Середнім***
  - Видалення запису
- Нормалізація
  - ***Встановлення значень в діапазоні 0-1***
- Розподіл на набори:
  - ***Тренувальний***
  - Валідаційний
  - ***Тестовий***

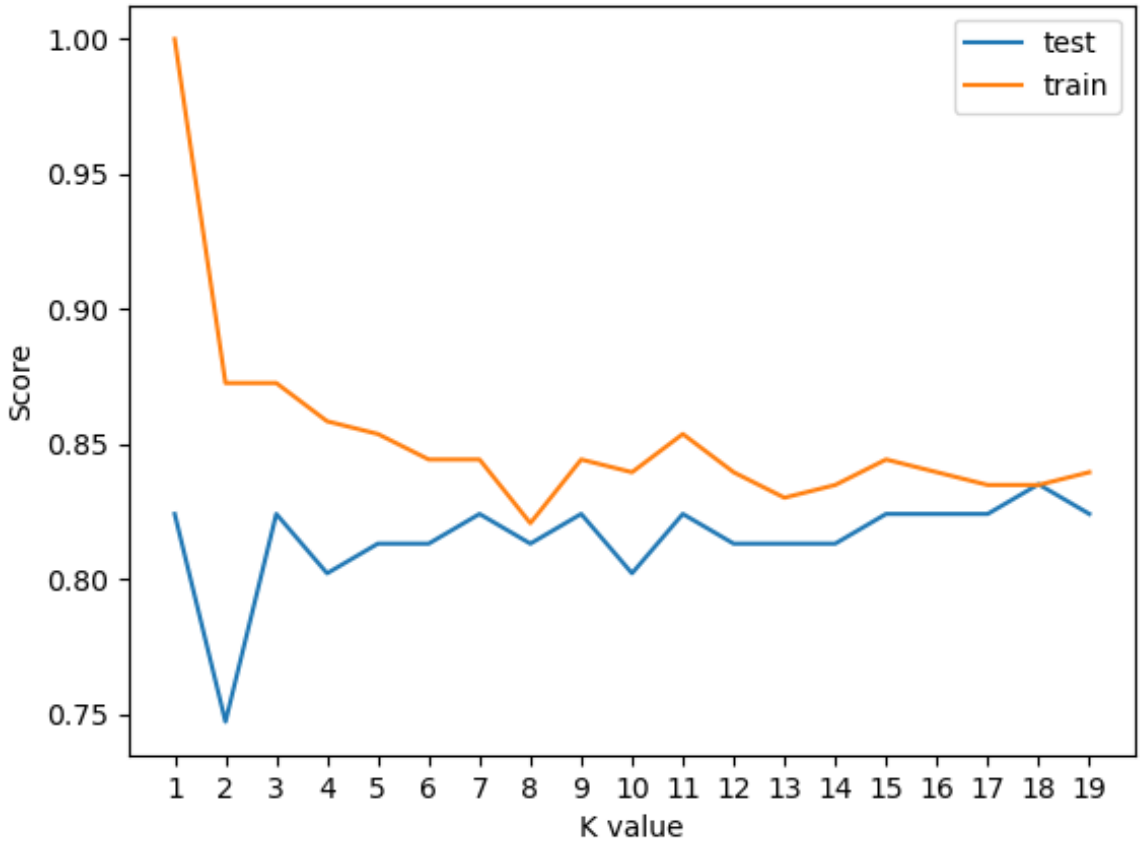
Використані в роботі методи

Повна назва методу	Отримана точність, %	
	Серцево-судинні захворювання	Хвороби печінки
Метод опорних векторів	86,83	72
Метод наївного Баєса	85,71	50,29
Дерево рішень	70,33	62,29
Логістична регресія	85,71	73,14
“Random forest”	82,42	68
Штучна нейронна мережа	91,21	75,43
Метод к-найближчих сусідів	83,52	71,43

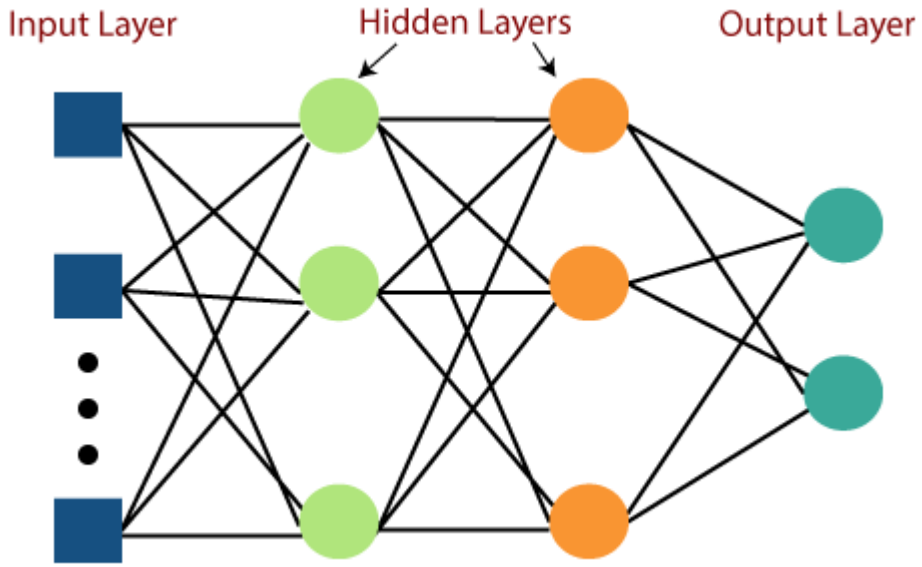
1. Дослідження сфери
2. Аналіз наявних робіт і вибір методів
3. Вибір даних
4. Аналіз досліджень на основі цих даних
5. Попередня обробка
6. Побудова моделей
6.1 Методи
6.2 Гіперпараметри
6.3 Візуалізація
7. Висновки

Підбір гіперпараметрів

Параметра k для методу k-найближчих сусідів

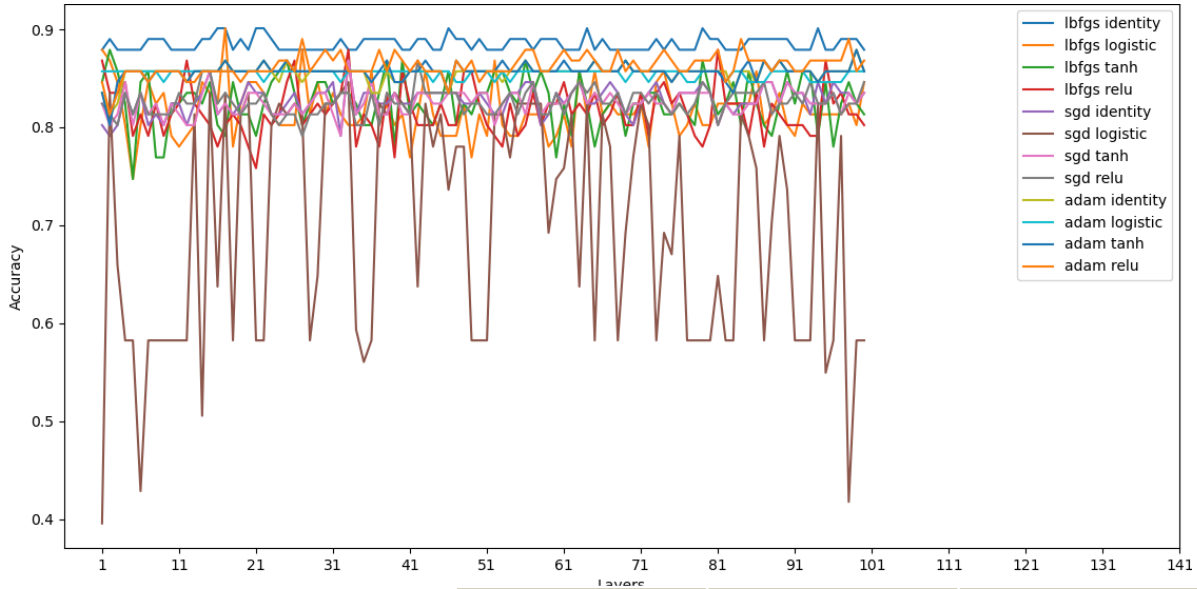


Гіперпараметри багатошарового перцептрона	**Серцево-судинні захворювання**	**Хвороби печінки**
Кількість прихованих шарів і нейронів в них	4 – в 1-му шарі 7 – в 2-му	2 – в 1-му шарі 8 – в 2-му
Функція активації	Relu	Tanh
Функція зворотного поширення	lbfgs	lbfgs



Модель багатошарового перцептрона

1. Дослідження сфери
2. Аналіз наявних робіт і вибір методів
3. Вибір даних
4. Аналіз досліджень на основі цих даних
5. Попередня обробка
6. **Побудова моделей**
   - **6.1 Методи**
   - **6.2 Гіперпараметри**
   - ***6.3 Візуалізація***
7. Висновки



Heart1	identity	logistic	tanh	relu
lbfgs	2.6977 s test set - 0.9011 train set - 0.8491 Bias - 0.0520 17 neurons	18.6049 s test set - 0.9011 train set - 1.0000 Bias - -0.0989 18 neurons	17.4155 s test set - 0.8791 train set - 1.0000 Bias - -0.1209 3 neurons	28.5483 s test set - 0.8791 train set - 1.0000 Bias - -0.1209 34 neurons
sgd	30.8591 s test set - 0.8681 train set - 0.8538 Bias - -0.0144 34 neurons	41.4602 s test set - 0.8462 train set - 0.5283 Bias - 0.3179 34 neurons	41.5007 s test set - 0.8681 train set - 0.8491 Bias - 0.0191 34 neurons	62.8697 s test set - 0.8681 train set - 0.8443 Bias - 0.0238 43 neurons
adam	16.7942 s test set - 0.8791 train set - 0.8396 Bias - 0.0395 100 neurons	37.7403 s test set - 0.8571 train set - 0.8538 Bias - 0.0034 2 neurons	22.022 s test set - 0.8791 train set - 0.8396 Bias - 0.0395 100 neurons	51.1394 s test set - 0.8901 train set - 0.8774 Bias - 0.0128 28 neurons

```
Logistic Regression: test - 85.7143, train - 84.4340, time - 0.0140 s.
KNN: test - 83.5165, train - 83.4906, time - 0.6004 s.
SVM: test - 86.8132, train - 88.6792, time - 0.0090 s.
Naive Bayes: test - 85.7143, train - 84.9057, time - 0.0040 s.
Decision tree: test - 70.3297, train - 100.0000, time - 0.0140 s.
Random forest: test - 82.4176, train - 100.0000, time - 2.2958 s.
```

1. Дослідження сфери
2. Аналіз наявних робіт і методів
3. Вибір даних
4. Аналіз досліджень на основі цих даних
5. Попередня обробка
6. Побудова моделей
7. **Висновки**
   - ***7.1 Результати***
   - **7.2 Проблеми**

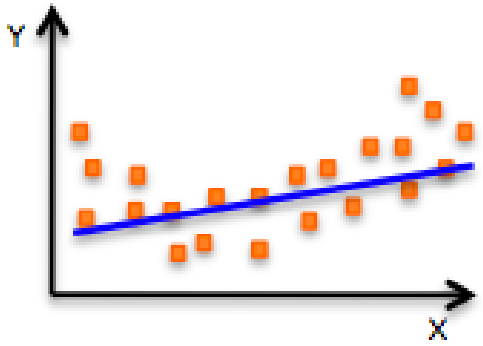
Результати

	Нейронна мережа		Стандартні методи машинного навчання		Результати отримані зі сторонніх досліджень	
	1 прихований шар	2 приховані шари				
Серцево-судинні захворювання	90.11	91.21	86.83%	SVM	83%	DT
Хвороби печінки	74.86	75.43	73.14%	LR	74.2%	RF

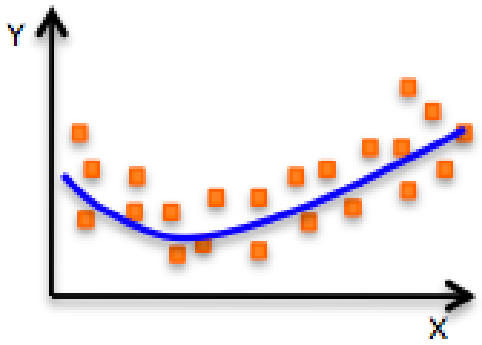
1. Дослідження сфери
2. Аналіз наявних робіт і методів
3. Вибір даних
4. Аналіз досліджень на основі цих даних
5. Попередня обробка
6. Побудова моделей
7. Висновки
7.1 Результати
7.2 Проблеми

Проблеми

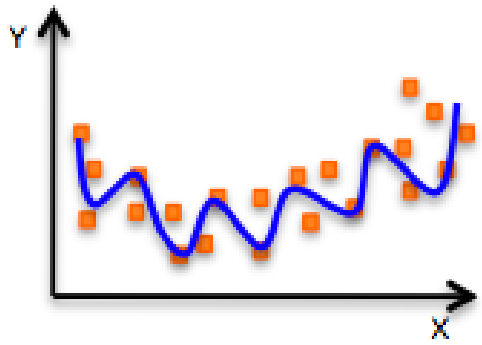
- Пошук наборів даних – обмежена кількість
- ***Достовірність даних***
- Компроміс між узагальненням та підлаштуванням під дані
- Оцінка ефективності



Надмірне узагальнення



Точна модель



Підлаштування під дані

- Дякую за увагу!