

"Програмування систем штучного інтелекту"

Метрики оцінки якості моделей та контроль їх навчання



Вступ до оцінки якості моделей машинного навчання

Критична важливість

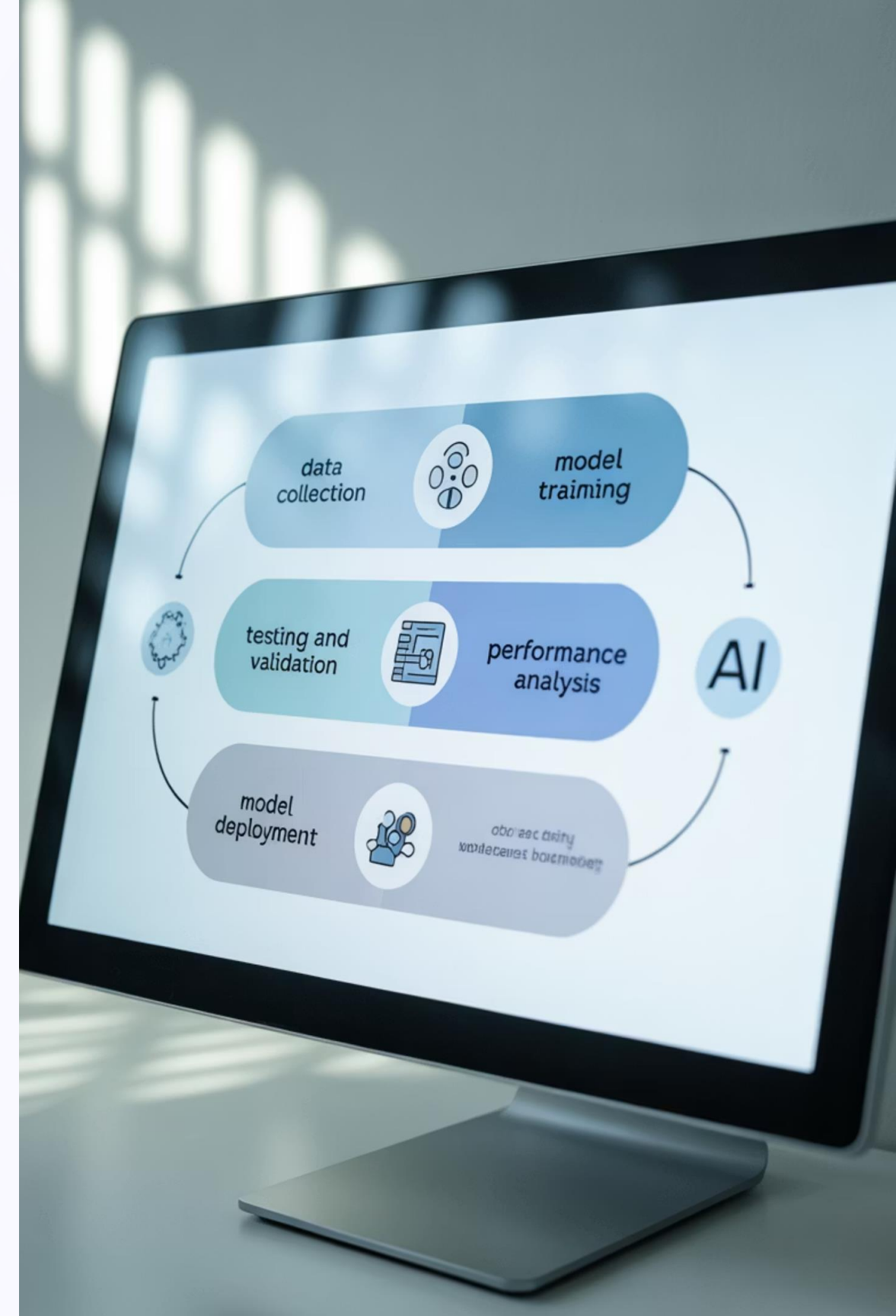
Оцінка якості визначає успіх AI-систем у реальних умовах та їх надійність

Основні цілі

- Вибір найкращої моделі
- Контроль процесу навчання
- Запобігання пере- та недонавчанню

Ключові питання

Наскільки модель точна? Чи узагальнює вона добре? Чи стабільна в роботі?



Основні поняття: Що таке метрика?

Визначення метрики

Метрика – це числова міра якості моделі або ефективності процесу навчання, що дозволяє об'єктивно оцінити продуктивність алгоритму

Метрики без чіткої цілі та контексту застосування є марними

Приклади метрик

- Точність (Accuracy)
- Відтворення (Recall)
- F1-міра
- Логарифмічна втрата

Важливість: правильний вибір метрики під конкретну задачу критично впливає на результат



Метрики для задач класифікації: базові визначення

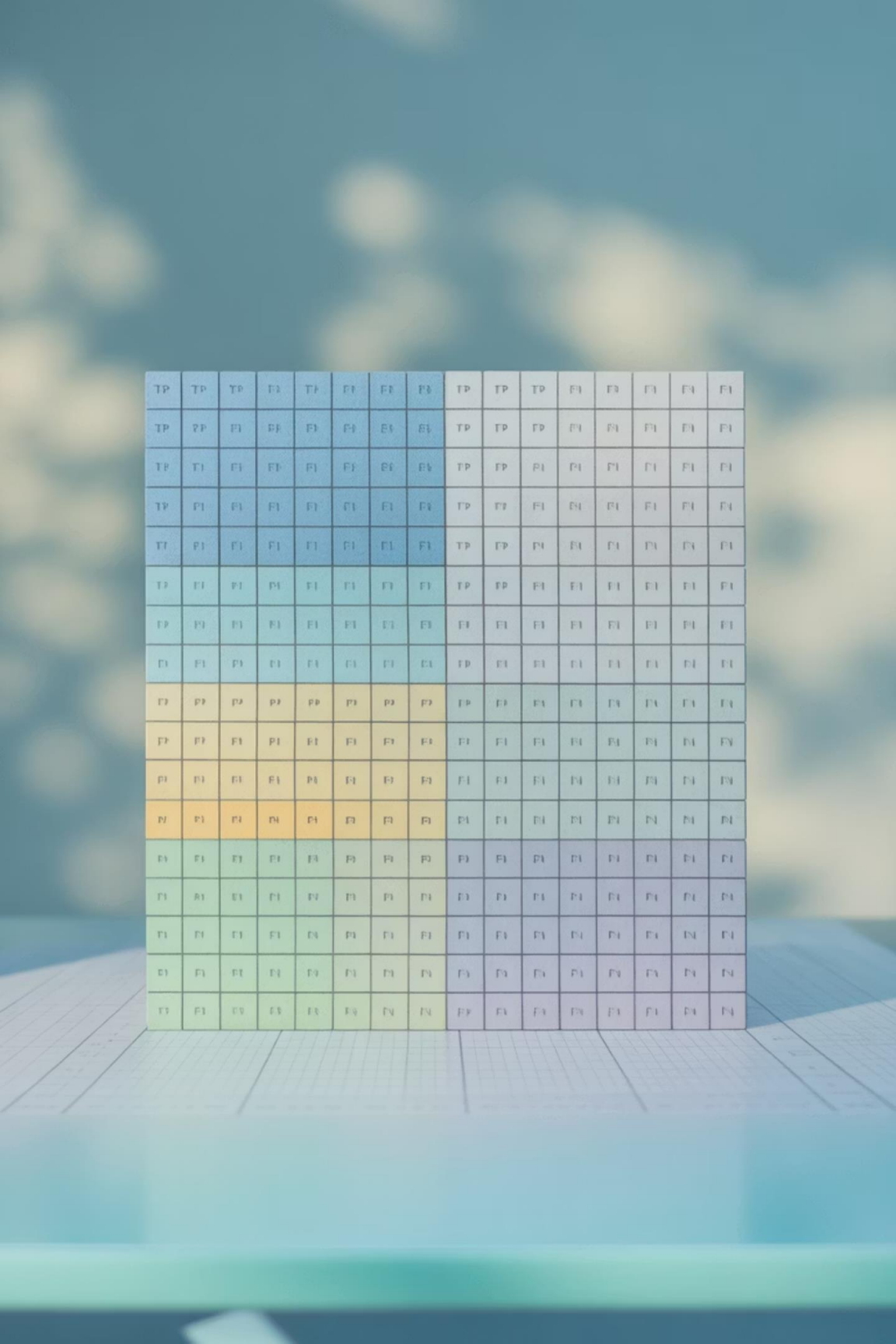
Істинно позитивні (TP)
Правильно класифіковані позитивні випадки

Істинно негативні (TN)
Правильно класифіковані негативні випадки

Хибно позитивні (FP)
Неправильно класифіковані як позитивні (помилка I роду)

Хибно негативні (FN)
Неправильно класифіковані як негативні (помилка II роду)

Матриця плутанини є основою для обчислення всіх основних метрик класифікації



Точність (Accuracy)

Математичне визначення

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Переваги

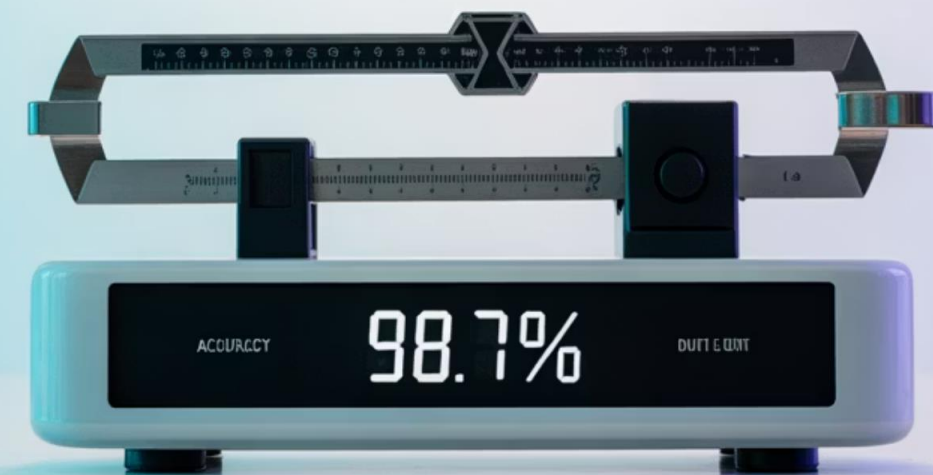
- Інтуїтивно зрозуміла
- Проста у розрахунку
- Широко використовується

Обмеження

- ❏ Не підходить для незбалансованих класів, де один клас значно переважає інші

Практичний приклад

Прогнозування успішності студентів: з 1000 студентів модель правильно передбачила результат для 850 - точність 85%





Precision (Точність) та Recall (Відтворення)

1

Precision

$$Precision = \frac{TP}{TP + FP}$$

Частка правильних позитивних прогнозів серед всіх позитивних прогнозів моделі

2

Recall

$$Recall = \frac{TP}{TP + FN}$$

Частка виявлених позитивних випадків серед всіх дійсно позитивних випадків

Баланс між precision та recall критично важливи для задач з різними ризиками помилок різних типів

F1-міра: гармонійне середнє Precision і Recall

Формула F1-міри

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

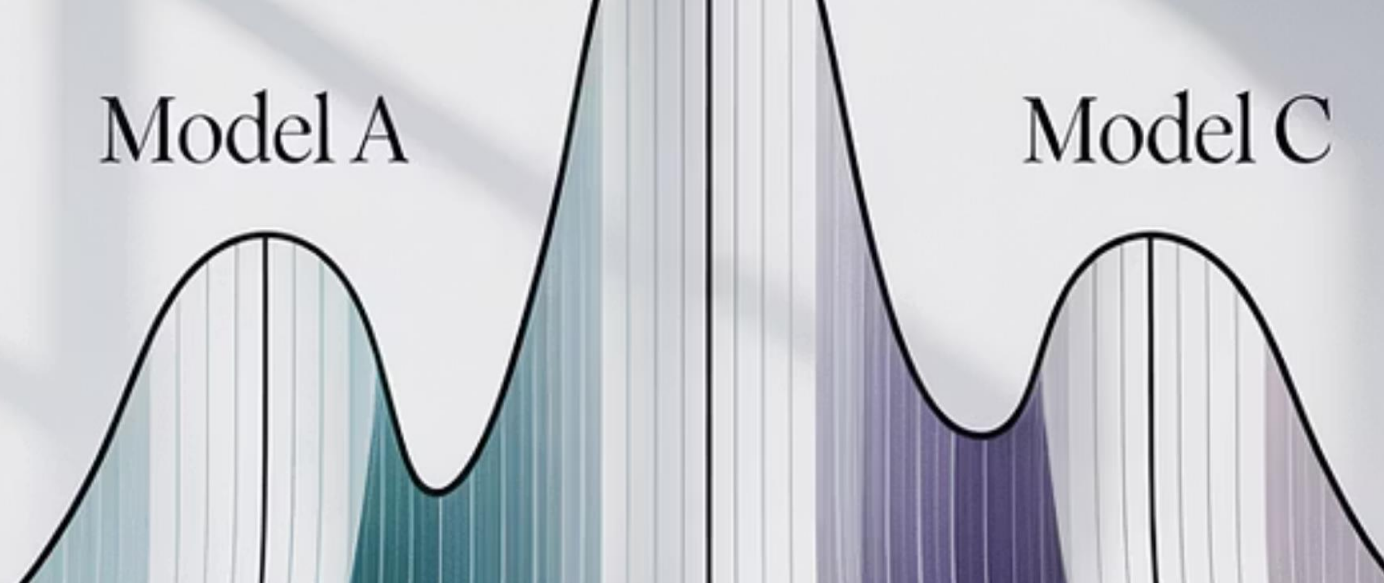
F1-міра забезпечує збалансовану оцінку, враховуючи як точність, так і повноту

Практичні приклади: виявлення дефектів у програмному забезпеченні, медична діагностика, системи безпеки

Коли використовувати

- Незбалансовані набори даних
- Потреба збалансувати помилки двох типів
- Порівняння різних моделей





Логарифмічна втрата (Log Loss)

01

Оцінка невизначеності

Логарифмічна втрата оцінює якість імовірнісних прогнозів моделі

02

Інтерпретація

Чим менша втрата, тим вища впевненість моделі у правильних прогнозах

03

Застосування

Критично важлива для моделей, що видають імовірності замість бінарних рішень

Приклад: система прогнозування імовірності захворювання, де важлива не лише правильність, але й впевненість прогнозу

ROC-крива та AUC (Площа під кривою)

ROC-крива

Графік залежності True Positive Rate від False Positive Rate при різних порогах класифікації

- Показує компроміс між чутливістю та специфічністю
- Дозволяє вибрати оптимальний поріг

Важлива для порівняння моделей незалежно від конкретного порогу класифікації

AUC метрика

Інтегральна метрика якості бінарного класифікатора

- $AUC = 0.5$: випадковий класифікатор
- $AUC = 1.0$: ідеальний класифікатор
- $AUC > 0.8$: хороша модель





Контроль навчання: пере- та недонавчання

Переобучення (Overfitting)

Модель "запам'ятовує" шум у тренувальних даних замість вивчення справжніх закономірностей, що призводить до поганого узагальнення

- Висока точність на тренуванні
- Низька точність на тестуванні

Недообучення (Underfitting)

Модель не здатна вивчити основні закономірності в даних через недостатню складність або недостатній час навчання

- Низька точність на тренуванні
- Низька точність на тестуванні

Різниця між метриками на тренувальній та валідаційній вибірках - ключовий індикатор проблем

Візуалізація контролю навчання: графіки навчання

1

Криві навчання

Графіки точності/втрати на тренувальній та валідаційній вибірках протягом епох навчання

2

Діагностика переобучення

Зростання тренувальної точності при одночасному падінні валідаційної точності

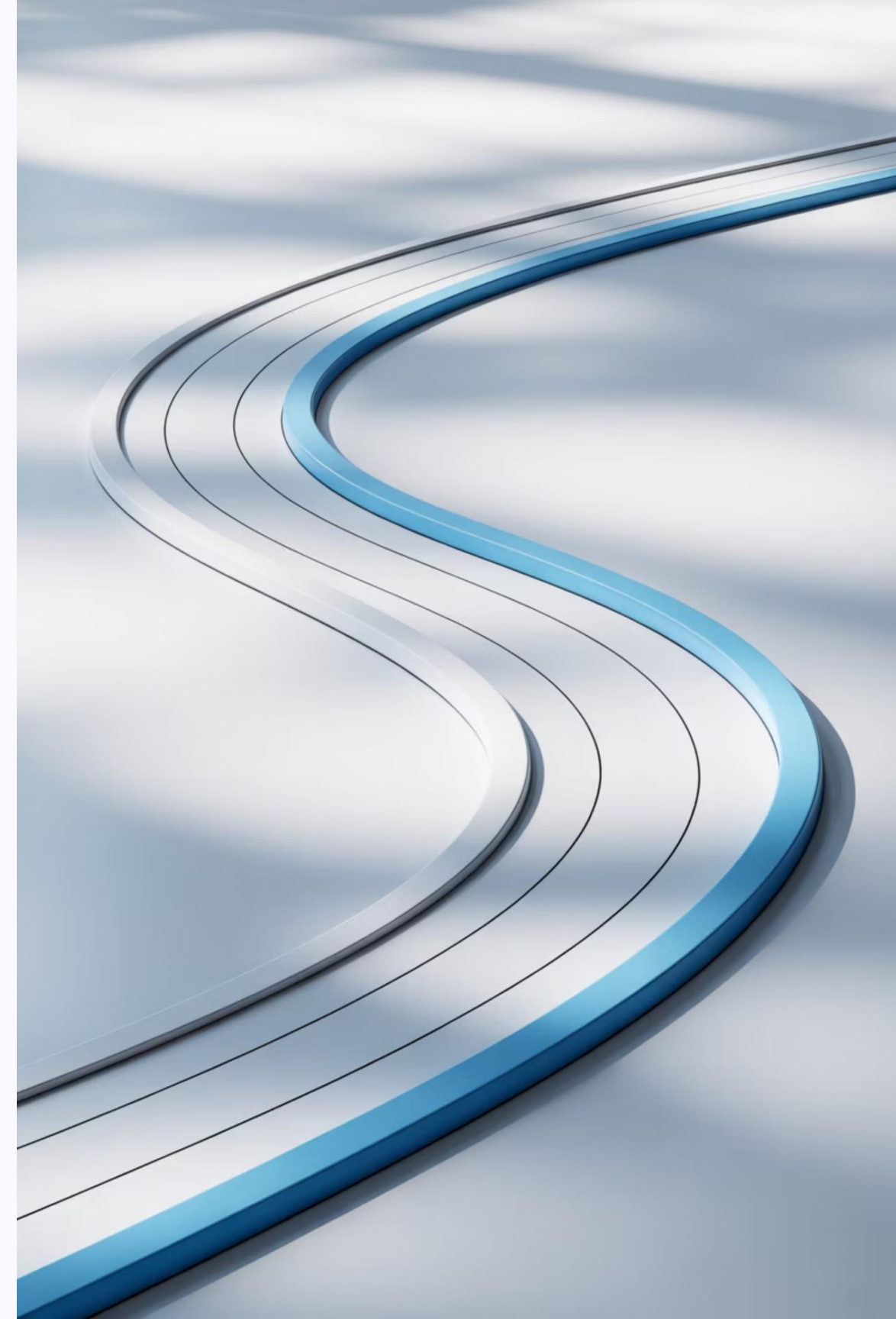
3

Практичні рішення

Регуляризація, рання зупинка, збільшення обсягу даних, dropout



Моніторинг кривих навчання в реальному часі дозволяє вчасно виявити проблеми та скоригувати процес



Метрики для задач регресії

$\frac{f}{dx}$

Середньоквадратична помилка (MSE)

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Чутлива до викидів, штрафує великі помилки сильніше



Середня абсолютна помилка (MAE)

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

Менш чутлива до викидів, легше інтерпретується



Коефіцієнт детермінації (R^2)

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

Показує частку дисперсії, пояснену моделлю

Приклад застосування: прогнозування цін на нерухомість, де кожна метрика надає різну інформацію про якість моделі

Property
Value Trend



Приклади застосування метрик у реальних проектах



Прогнозування успішності студентів

Використання accuracy та F1-міри для збалансованої оцінки ризиків неуспішності та підтримки студентів



Виявлення дефектів у ПЗ

Precision та Recall критичні - важливо знайти всі баги (високий Recall) без хибних спрацювань



Оцінка мовних моделей

Perplexity, BLEU, BERTScore для оцінки якості генерації тексту та перекладу



Контроль якості анотацій

Cohen's Kappa та Krippendorff's alpha для оцінки узгодженості між анотаторами

Сучасні підходи до оцінки моделей AI

Гібридна оцінка

Комбінування автоматичних метрик з експертною людською оцінкою для більш повного розуміння якості моделі

LLM як судді

Використання великих мовних моделей для автоматизованої оцінки якості відповідей інших AI-систем

Етичні аспекти

Виявлення та усунення упереджень, забезпечення справедливості та інклюзивності AI-систем

Масштабовані рішення

Розробка стратегії для надійної оцінки великих моделей з мільярдами параметрів





Висновки та рекомендації

1

Відповідність цілям

Вибір метрик має точно відповідати специфіці задачі та бізнес-цілям проекту

2

Контроль навчання

Постійний моніторинг процесу навчання - ключ до створення стабільної та надійної моделі

3

Комплексний підхід

Використовуйте комбінацію різних метрик для отримання повної картини продуктивності

4

Постійна адаптація

Регулярно аналізуйте результати та адаптуйте підходи до оцінки відповідно до нових вимог

Наступні кроки: Питання? Обговорення практичних кейсів та hands-on практика на наступних заняттях



101100101

110010101101101101

010010010100101

1101101011010

Дякую за увагу!

Запитання