ЛЬВІВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ ІВАНА ФРАНКА
Факультет прикладної математики та інформатики
Кафедра обчислювальної математики

Курсова робота

Порівняння лінійних та нелінійних методів статистичного навчання для задачі з Kaggle

Виконав студент III курсу групи
ПМп-31 напрямку підготовки
(спеціальності)
113 – "Прикладна математика"
Середович В.В.

Керівник: Вавричук В.Г.

Львів - 2022

Зміст

Вступ

Сучасний світ демонструє надшвидку динаміку змін, глобалізованість світу та доступність величезної кількості інформаційних ресурсів. Нові можливості підштовхують розвиток оцінки даних якомога глибше. Сфера статистики — це наука про навчання на даних. Статистичні знання допомагають використовувати правильні методи для збору даних, правильного аналізу та ефективного представлення результатів. Будь-якому спеціалісту в ході практичної діяльності доводиться здійснювати операції над кількісними даними, що здійснюються відповідно до математичних законів. Математична теорія змінюється порівняно повільно, проте технологія застосування математичних методів зазнала більш істотних змін. Активне впровадження безперешкодних комунікацій, якісної трансформації у безлічі галузей економіки та цифровізація суспільства у сфери здобутків штучного інтелекту та машинного навчання – лише частина трендів. Прогнозування дозволяє підприємствам ставити розумні та вимірювані цілі на основі поточних та історичних даних. Наявність точних даних і статистичних даних для аналізу допомагає підприємствам вирішувати, які зміни, зростання або покращення будуть визначені як успіх. В межах цієї роботи ми розглянемо деякі існуючі методи на основі алгоритмів, проаналізуємо їх та порівняємо їх між собою.

1 Постановка задачі

Метою даної роботи є дослідження лінійних та нелінійних методів статистичного навчання та їх порівняння. В рамках роботи ми візьмемо набір даних про пацієнтів. Якщо ми визначимо, що існує зв'язок між хворобою серця та віком, холестерином, рівнем цукру, пульс та іншими параметрами, тоді ми можемо заздалегідь передбачити загрозу хвороби пацієнтів, тим самим рятувати життя. Іншими словами, наша мета — розробити точну модель, яку можна використовувати щоб спрогнозувати хворобу на основі параметрів людини. Ми будемо перевіряти методи навчання, використовувати їх для передбачення серцевої недостатності, оцінювати та порівнювати їх.

2 Порівняння машинного та статистичного навчання

Існують відмінності через історичні та соціологічні причини. Статистика є старшою галуззю ніж машинне навчання. Таким чином, ідеї щодо збору та аналізу даних у статистиці корінням сягають ще до того, як існували комп'ютери. Звичайно, ця галузь адаптувалася і з часом підходи до проблем і вибір теми дослідження в машинному та статистичному навчанні часто відрізняються між собою.

Машинне навчання — це область досліджень, присвячена розумінню та створенню методів, які «навчаються», тобто методів, які використовують дані для підвищення продуктивності певного набору завдань.

Модель машинного навчання називається алгоритм, який навчений розпізнавати певні типи закономірностей у даних. Модель навчають на основі набору інформації, надаючи їй алгоритм, який вона може використовувати для аналізу та навчання.

Статистичне навчання є однією з класичних структур для штучного інтелекту та області машинного навчання. Це набір математичних інструментів для вивчення функцій даних. Статистичне навчання базується на статистиці та функціональному аналізі.

Статистична модель — це модель даних, використовується для того, щоб зробити висновок про взаємозв'язки всередині них або щоб створити модель здатну передбачати майбутні значення.

Обидві галузі об'єднує одне питання: **як модель вчиться на даних?** З роботи [3] можемо схарактеризувати машинне та статистичне навчання:

Отже, машинне навчання:

- навчається на даних без явно запрограмованих інструкцій
- може базуватись на основі мільярдів спостережень та атрибутів
- не залежить від припущень і в більшості випадків їх ігнорує
- визначає шаблони з набору даних за допомогою ітерацій, які вимагають набагато менше людських зусиль

Натомість, статистичне навчання:

- базується на меншому наборі даних з кількома атрибутами
- наголошує на передбаченнях, навчанні з наглядом, навчанні без нагляду та навчанні з частковим наглядом
- спирається на інтенсивну математику, яка базується на оцінювачі коефіцієнтів і вимагає хорошого розуміння даних

Загалом, ці дві сфери все більше змішуються між собою та обидві використовуються для розпізнавання образів, виявлення знань і аналізу даних.

3 Оцінка функції f

Для оцінки даних ми будемо мати вхідні та вихідні параметри. Вхідні змінні зазвичай позначаються за допомогою символу X. Вхідні параметри мають такі назви як предиктори, незалежні змінні, ознаки або просто змінні. Вихідна змінна часто називається залежною змінною і зазвичай позначається символом Y.

Загалом, припустимо, що ми спостерігаємо вихідний параметр Y та p різних предикторів, $X_1, X_2, \ldots, X_p$. Ми припускаємо, що між Y та $X = (X_1, X_2, \ldots, X_p)$ є певний зв'язок, який можна записати в загальному вигляді

$$Y = f(X) + \epsilon$$

Тут f є деякою фіксованою, але невідомою функцією від $X_1, \ldots, X_p$, та ϵ випадкова похибка, незалежна від X середнє значення дорівнює нулю. f представляє систематичну інформацію, як X описує Y.

По суті, **статистичне навчання** належить до набору підходів для оцінки f.

Є два основні критерії, за якими ми оцінюємо, точність f: **передбачення(prediction)** та **завдання висновку(Inference)**.

Передбачення:

У багатьох ситуаціях набір параметрів X є доступний, але Y непросто отримати. В цьому випадку, якщо випадкова похибка в середньому наближена до нуля, можемо прогнозувати Y використовуючи

$$\hat{Y} = \hat{f}(X)$$

де $\hat{f}$ представляє нашу оцінку для f, і $\hat{Y}$ представляє результуючу оцінку Y. В цьому випадку, $\hat{f}$ трактують як чорна скринька (black box), маючи на увазі, що зазвичай вона не зацікавлена в точній формі $\hat{f}$, за умови, що вона дає точні прогнози для Y.

Припустимо, що $X_1, \ldots, X_p$ – це характеристики зразка крові пацієнта, які можна легко виміряти в лабораторії, а Y – це змінна, що кодує ризик тяжкого захворювання пацієнта, побічна реакція на певний препарат. Природно намагатись спрогнозувати Y за допомогою X, оскільки це може допомогти уникнути надання відповідного препарату пацієнтам, які мають високий ризик побічної реакції, тобто тим, для яких оцінка Y є висока.

Точність $\hat{Y}$ як передбачення для Y залежить від двох величин, які ми будемо називати зменшуваною (reducible) похибкою та не зменшуваною (irreducible) похибкою. Загалом, $\hat{f}$ не буде ідеальною оцінкою для f, і ця неточність призведе до певної похибки. Цю похибку можна зменшити, оскільки ми потенційно можемо підвищити точність $\hat{f}$, використовуючи для оцінки f найбільш вдалу методику статистичного навчання. Однак, навіть якби можна було сформувати ідеальну оцінку для f, щоб наша оцінена відповідь мала вигляд $\hat{Y} = f(X)$, отримане передбачення все одно мало б певну похибку. Це відбувається тому, що Y також є функцією ϵ, яку, за визначенням, не можна передбачити за допомогою X. Тому мінливість, пов'язана з ϵ, також впливає на точність наших прогнозів. Це поняття відоме як не зменшувана похибка, тому що, незалежно від того, наскільки добре ми оцінюємо f, ми не можемо зменшити похибку, яку вносить ϵ.

Не зменшувана похибка є завжди більшою за нуль, тому що величина ϵ може містити невиміряні змінні, корисні для передбачення Y, а отже, f не може використовувати їх для свого передбачення. Величина ϵ також може містити невиміряні варіації. Наприклад, ризик побічної реакції може змінюватися для деякого пацієнта в певний день, залежно від зміни самого препарату або загального самопочуття пацієнта в цей день.

Розглянемо задану оцінку $\hat{f}$ і набір предикторів X, що дає прогноз $\hat{Y} = \hat{f}(X)$. Припустимо, що і $\hat{f}$, і X фіксовані. Тоді легко показати, що

$$\begin{aligned} E(Y - \hat{Y})^2 &= E[f(X) + \epsilon - \hat{f}(X)]^2 \\ &= \underbrace{[f(X) - \hat{f}(X)]^2}_{\text{Reducible}} + \underbrace{\text{Var}(\epsilon)}_{\text{Irreducible}} \end{aligned}$$

де $E(Y - \hat{Y})^2$ являє собою середнє квадратичну різницю між прогнозованим і фактичним значенням Y і $\text{Var}(\epsilon)$ та представляє дисперсію, пов'язану з випадкової похибки ϵ.

Завдання висновку.

Зазвичай, нас цікавить розуміння того як зміна Y впливає на $X_1 \ldots, X_p$. У цій ситуації ми хочемо оцінити f, але наша мета не обов'язково полягає в тому, щоб робити прогнози для Y. Натомість ми хочемо зрозуміти зв'язок між X та Y або як змінюється Y як функція від $X_1, \ldots, X_p$. Тепер $\hat{f}$ не можна розглядати як чорну скриньку, тому що нам потрібно знати його точну форму. У цьому контексті нас може цікавити відповідь на такі запитання:

- Які предиктори пов'язані з відповіддю? Часто буває, що лише невелика частина доступних предикторів істотно пов'язана з Y. Визначення

кількох важливих предикторів серед великого набору можливих змінних може бути надзвичайно корисним, залежно від застосування.

- Який зв'язок між відповіддю та кожним окремим предиктором? Деякі предиктори можуть мати позитивний зв'язок з Y, у сенсі що збільшення предиктора пов'язане зі збільшенням значень Y. Інші предиктори можуть мати протилежний вплив. Залежно від складності f, зв'язок між відповіддю та певним предиктором може також залежати від значень інших предикторів.

- Чи можна адекватно підсумувати зв'язок між Y і кожним предиктором за допомогою лінійного рівняння, чи зв'язок є нелінійним? Історично, більшість методів оцінки f мали лінійну форму. У деяких ситуаціях таке припущення є розумним або навіть бажаним. Але часто справжнє відношення є складнішим, і в цьому випадку лінійна модель не може забезпечити точне представлення зв'язку між вхідними та вихідними змінними.

4 Категоризація методів навчання

4.1 Задачі регресії та задачі класифікації

З підручника з основам статичтичного аналізу [4] можемо схарактеризувати завдання апроксимації як завдання згладжування експериментальних даних. **Апроксимацією** називається процес підбору емпіричної формули $f()$ для встановленої з досвіду функціональної залежності $Y = f(x)$. Формула використовується для аналітичного представлення досліджуваних даних. Тобто завдання алгоритму моделювання полягає в тому, щоб знайти найкращу можливу функцію відображення з огляду на наявний час і ресурси.

Змінні можна поділити як **кількісні** або **якісні** (категоричні). Кількісні змінні набувають числових значень (вік людини, зріст чи дохід, вартість будинку). Якісні змінні набувають значення в одному з K різних класів або категорій (чоловічий чи жіночий, марка A, B або C, так чи ні). Зазвичай, до проблем з кількісною відповіддю, звертаються як до проблем **регресії**. Натомість задачі з якісною відповіддю часто називають проблемами **класифікації**.

4.2 Лінійні та нелінійні методи

Лінійні та нелінійні підходи до оцінки f мають певні характеристики. Припустимо, що існує набір з n різних даних. Дані, які ми будемо використовувати для спостереження та тренування методів навчання, будемо називати навчальними даними.

Модель регресії є лінійною, коли доданками рівняння є константа та один або декілька параметрів помножених на незалежну зміну. Отже, тип рівняння регресії є лінійним за параметрами.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

Для **нелінійної регресії** дані спостереження моделюються функцією, яка є нелінійною комбінацією параметрів моделі та залежить від однієї або кількох незалежних змінних. Дані підбираються методом послідовних наближень. Тобто такий тип регресії використовується для моделювання зв'язку між незалежними змінними та залежними змінними.

Лінійна класифікація належить до класифікації набору точок даних до дискретного класу на основі лінійної комбінації його змінних. Прикладом є логістична регресія. Лінійні методи класифікації використовуються коли навчальні дані не мають складних зв'язків та можуть бути лінійно розділені.

Рис. 4.1: Ліворуч проілюстрована лінійна двокласна класифікація. Тут розділювальна межа визначається як $x^T w = 0$. Праворуч нелінійна двокласова класифікація досягається шляхом введення нелінійних перетворень ознак у нашу модель.

Нелінійну класифікацію використовують для розпізнавання екземплярів, коли навчальні дані мають складніші, нелінійні зв'язки та не можуть бути точно передбачені лінійними методами.

$$model(x, \theta) = w_0 + f_1(x)w_1 + f_2(x)w_2 + \cdots + f_k(x)w_k$$

$f_1, f_2, \ldots, f_k$ є нелінійними параметризованими або не параметризованими функціями та $w_0 - w_k$ представлені в наборі ваг θ.

4.3 Параметричні та непараметричні методи

Загалом, більшість статистичних методів навчання для цього завдання можна схарактеризувати як **параметричні** або **непараметричні**.

Основна ідея **параметричного методу** полягає в тому, що існує набір фіксованих параметрів, які використовуються для визначення ймовірнісної моделі. Для параметричних методів ми попередньо знаємо, що сукупність є нормальною, або якщо ні, то ми можемо легко наблизити її за допомогою нормального розподілу.

Спочатку ми робимо припущення щодо функціональної форми або форми f. Наприклад, одне дуже просте припущення полягає в тому, що f є лінійним у X:

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p. \tag{4.1}$$

Після припущення нам потрібно лише оцінити $p+1$ коефіцієнти $\beta_0, \beta_1, \ldots, \beta_p$. Обравши модель, далі нам потрібна процедура яка використовує навчальні дані для f і t. У випадку лінійної моделі 4.1 нам необхідно оцінити параметри $\beta_0, \beta_1, \ldots, \beta_p$. Тобто ми хочемо знайти значення цих параметрів такі,

що

$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p.$$

Описаний підхід зводить проблему оцінки f до проблеми оцінки набору параметрів. Припущення параметричної форми для f спрощує проблему оцінки f, оскільки, як правило, набагато легше оцінити набір параметрів, таких як $\beta_0, \beta_1, \ldots, \beta_p$ у лінійній моделі 4.1, ніж f.

Непараметричні методи не роблять явних припущень про функціональну форму f. Замість цього вони шукають оцінку f, яка наближається до значення точки даних, наскільки це можливо. Такі підходи можуть мати велику перевагу над параметричними: уникаючи припущення певної функціональної форми для f, вони мають потенціал, щоб точно підігнати ширший діапазон можливих форм для f.

Будь-який параметричний підхід несе за собою ризик, що функціональна форма використана для оцінки f, буде дуже відмінною від істинної f, і в цьому випадку отримана модель буде погано наближати дані. Натомість непараметричні підходи повністю уникають цієї небезпеки, оскільки, не роблять припущень про форму f. Але непараметричні підходи мають недолік: оскільки вони не зводять задачу оцінки f до малої кількості параметрів, для того, щоб отримати точну оцінити f, необхідна велика кількість спостережень.

4.4 Навчання з наглядом та без нагляду

Для навчання під наглядом для кожного спостереження предиктора $x_i, i = 1...n$ існує відповідна відповідь y_i. Ми хочемо підібрати модель, яка пов'язує відповідь до предиктора, з метою точного прогнозування відповіді (prediction) або розуміння зв'язку між відповіддю та предиктором (inference). Багато класичних статистичних методів навчання, таких як лінійна регресія та логістична регресія працюють у контрольованій області навчання. Для класифікації, контрольована модель навчання передбачає наявність вчителя або керівника, який класифікує навчальні приклади за класами та використовує інформацію про членство в кожному навчальному екземплярі.

Навпаки, навчання без нагляду описує ситуацію, в якій для кожного спостереження $i = 1...n$ існує вектор вимірювань x_i, але без відповідного y_i. В такому випадку, неможливо підібрати модель лінійної або логістичної регресії, оскільки немає змінної відповіді для вимірювань, яку можна було б передбачити. У цьому контексті ми в певному сенсі працюємо наосліп. Ситуація називається неконтрольованою, тому що дозволяє моделі працювати

самостійно, щоб виявляти закономірності та інформацію, які не були виявлені раніше. До неконтрольованих методів належать такі методи як k-найближчих сусідів та деякі нейроні мережі.

5 Оцінка якості методів

Жоден метод не домінує над усіма іншими на будь-якому наборі даних. На певному наборі даних може добре працювати один конкретний метод, але якийсь інший метод може краще працювати на схожому, але іншому наборі даних. Тому, завжди важливо приймати рішення про те, який метод дає найкращі результати на основі кожного окремого датасету.

5.1 Середня квадратична похибка (MSE)

У випадку **регресії**, найбільш часто використовуваною оцінкою алгоритму є середня квадратична похибка (MSE), яка визначається у вигляді 5.1.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}\left(x_i\right)\right)^2 \tag{5.1}$$

де $\hat{f}(x_i)$ – прогноз, який $\hat{f}$ дає для і-го спостереження. MSE буде невеликим, якщо передбачені відповіді дуже близькі до істинних відповідей і, навпаки, великим, якщо для деяких спостережень передбачені та істині відповіді суттєво відрізняються.

MSE в 5.1 обчислюється з використанням тренувальних даних. Загалом, для нас важливіше не те як модель працює на тренувальних даних, а те яка точність дає модель під час застосування методу до тестових даних.

Припустимо, що ми підігнали наш статистичний метод навчання до тренувальних спостережень $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, і отримуємо оцінку $\hat{f}$. Потім, ми можемо обчислити $\hat{f}(x_1), \hat{f}(x_2), \ldots, \hat{f}(x_n)$. Якщо вони близькі до значень $y_1, y_2, \ldots, y_n$, то оцінка MSE, задана 5.1, буде мала. Проте нас насправді не цікавить, чи $\hat{f}(x_i) \approx y_i$. Натомість ми хочемо знати, чи $\hat{f}(x_0)$ приблизно дорівнює y_0, де (x_0, y_0) — це раніше невідоме тестове спостереження, яке не використовувалося для навчання методу статистичного навчання. Таким чином, ми хочемо вибрати метод, який дає найнижчий тестовий MSE, а не найнижчий навчальний MSE.
Іншими словами, якби у нас була велика кількість тестових спостережень, ми могли б обчислити

$$\text{Ave}\left(\hat{f}\left(x_0\right) - y_0\right)^2, \tag{5.2}$$

Середньоквадратичну похибку передбачення для цих тестових спостережень (x_0, y_0).

5.2 Компроміс зсуву та дисперсії

З підручника The Elements of Statistical Learning [1] можемо сказати що MSE, для заданого значення x_0, завжди розкладається на суму трьох основних величин: дисперсії $\hat{f}(x_0)$, квадратичне зміщення $\hat{f}(x_0)$ та дисперсії випадкової похибки ϵ.

$$E\left(y_0 - \hat{f}(x_0)\right)^2 = \mathrm{Var}\left(\hat{f}(x_0)\right) + \left[\mathrm{Bias}\left(\hat{f}(x_0)\right)\right]^2 + \mathrm{Var}(\epsilon) \tag{5.3}$$

Тут позначення $E\left(y_0 - \hat{f}(x_0)\right)^2$ визначає очікуваний тестовий MSE і відноситься до середнього тестового MSE, який ми отримали б, якби ми неодноразово оцінювали f, використовуючи велику кількість навчальних наборів і перевіряли кожен на x_0. Загальний очікуваний тестовий MSE можна обчислити шляхом усереднення $E\left(y_0 - \hat{f}(x_0)\right)^2$ за всіма можливими значеннями x_0 у тестовому наборі.

З рівняння 5.2 маємо, що для того, щоб мінімізувати очікувану похибку тесту, нам потрібно вибрати метод статистичного навчання, який одночасно забезпечує низьку дисперсію та низьке зміщення.

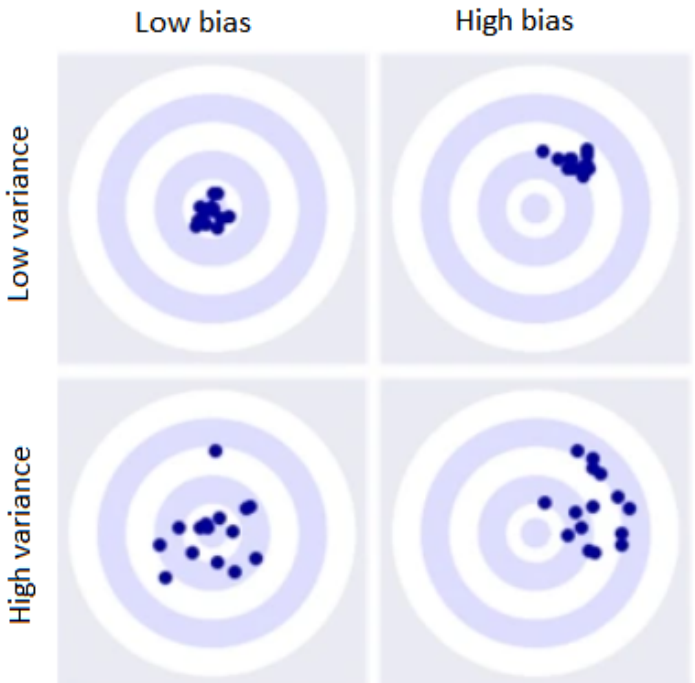

Рис. 5.1: Пояснення дисперсії та зсуву на прикладі гри в дартс.

Дисперсія належить до величини, на яку змінилася б $\hat{f}$, якби ми оцінили її за допомогою іншого набору навчальних даних. Оскільки навчальні

дані використовуються відповідно до статистичного методу навчання, різні набори навчальних даних призведуть до різних $\hat{f}$. Але в ідеалі оцінка f не повинна сильно відрізнятися між навчальними наборами. Однак, якщо метод має високу дисперсію, то невеликі зміни в навчальних даних можуть призвести до великих змін у $\hat{f}$. Загалом, більш гнучкі статистичні методи мають більшу дисперсію та призводять до меншої упередженості.

5.3 Налаштування класифікації

Припустимо, що ми прагнемо оцінити f на основі навчальних спостережень $\{(x_1, y_1), \ldots, (x_n, y_{n)})\}$, де тепер $y_1, \ldots, y_n$ є якісними. Найпоширенішим підходом для кількісного визначення точності нашої оцінки $\hat{f}$ є **коефіцієнт помилок навчання**, які ми отримуємо при застосованні нашої оцінки $\hat{f}$ до навчальних спостережень:

$$\frac{1}{n}\sum_{i=1}^{n} I\left(y_i \neq \hat{y}_i\right) \tag{5.4}$$

Тут $\hat{y}_i$ – це передбачена мітка класу для i-го спостереження за допомогою $\hat{f}$, $I\left(y_i \neq \hat{y}_i\right)$ – це індикаторна змінна, яка дорівнює 1, якщо $y_i \neq \hat{y}_i$ і нуль, якщо $y_i = \hat{y}_i$. Якщо $I\left(y_i \neq \hat{y}_i\right) = 0$, то i-е спостереження було класифіковано правильно за нашим методом класифікації, інакше його було неправильно класифіковано. Отже, рівняння 5.4 обчислює частку неправильних класифікацій.

Рівняння 5.4 називають коефіцієнтом помилок навчання, оскільки воно обчислюється на основі даних, які були використані для навчання нашого класифікатора. Рівень помилок тесту, пов'язаний з набором тестових спостережень у вигляді (x_0, y_0), визначається як

$$\text{Ave}\left(I\left(y_0 \neq \hat{y}_0\right)\right) \tag{5.5}$$

де $\hat{y}_0$ – це передбачена мітка класу, яка є результатом застосування класифікатора до тестового спостереження з предиктором x_0. Хорошим класифікатором є той, для якого тестова помилка 5.4 найменша.

5.4 Перехресна перевірка (Cross validation)

За відсутності дуже великого призначеного тестового набору, який можна використати, щоб оцінити частоту похибок тесту, можна використовувати ряд методик, щоб оцінити цю кількість, використовуючи наявні навчальні дані. Деякі методи роблять математичне корегування коефіцієнта похибок навчання, щоб оцінити частоту похибок тесту.

Підхід набору перевірки

Припустимо, що ми хотіли б оцінити похибку тесту, пов'язану з підгонкою конкретного статистичного методу навчання до набору спостережень.

Рис. 5.2: Схематичне відображення підходу до набору перевірки. Набір з n спостережень випадковим чином розбивається на навчальний набір (синій колір) і набір перевірки (коричневий). Статистичний метод тренується на навчальному наборі, а його ефективність оцінюється на наборі перевірки.

Підхід включає випадковий поділ доступного набору спостереження на дві частини, навчальний набір і набір для перевірки. Модель тренується на навчальному наборі, а підібрана модель використовується для прогнозування відповіді на спостереженнях в наборі перевірки. На основі цієї перевірки, зазвичай рахується середньоквадратична похибка для оцінки точності отриманої моделі.

Перехресна перевірка проблем класифікації

Перехресна перевірка також може бути дуже корисним підходом у випадку класифікації, коли Y є якісним. Тоді перехресна перевірка працює майже так само, за винятком того, що замість використання MSE для кількісної оцінки похибки тесту ми використовуємо кількість помилково класифікованих спостережень. Наприклад, у налаштуваннях класифікації частота помилок LOOCV приймає форму

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{Err}_i$$

де $\mathrm{Err}_i = I\left(y_i \neq \hat{y}_i\right)$. Коефіцієнт похибки у k-кратному CV і частоти похибок набору визначаються аналогічним чином.

5.5 Перехресна перевірка Leave One Out

В підручнику An Introduction to Statistical Learning [2] представлено підхід до набору валідації LOOCV. Він передбачає поділ набору спостережень

на дві частини. Однак замість створення двох підмножин порівнянного розміру, для набору перевірки використовується одне спостереження (x_1, y_1), а решта спостережень $\{(x_2, y_2), \ldots, (x_n, y_n)\}$ складають навчальний набір. Метод статистичного навчання тренують на $n-1$ навчальних спостережень, а для виключеного спостереження робиться передбачення $\hat{y}_1$, використовуючи його значення x_1. Оскільки (x_1, y_1) не використовувався в процесі підбору, $\text{MSE}_1 = (y_1 - \hat{y}_1)^2$ надає неупереджену оцінку помилки тесту. Але навіть попри те, що MSE_1 є неупередженим для тестової помилки, це погана оцінка, оскільки вона дуже змінна, оскільки базується на одному спостереженні $(x_1, y_1).1)$.

Ми можемо повторити процедуру, вибравши (x_2, y_2) для даних перевірки, навчаючи статистичну модель на основі $n-1$ спостереженнях $(x_1, y_1), (x_3, y_3), \ldots, (x_n,$ і обчислення $\text{MSE}_2 = (y_2 - \hat{y}_2)^2$. Повторення цього підходу n разів створює n квадратів похибок, $\text{MSE}_1, \ldots, \text{MSE}_n$. Оцінка LOOCV для тестового MSE є середнім з цих n оцінок похибок тесту:

$$\text{CV}_{(n)} = \frac{1}{n} \sum_{i=1}^{n} \text{MSE}_i$$

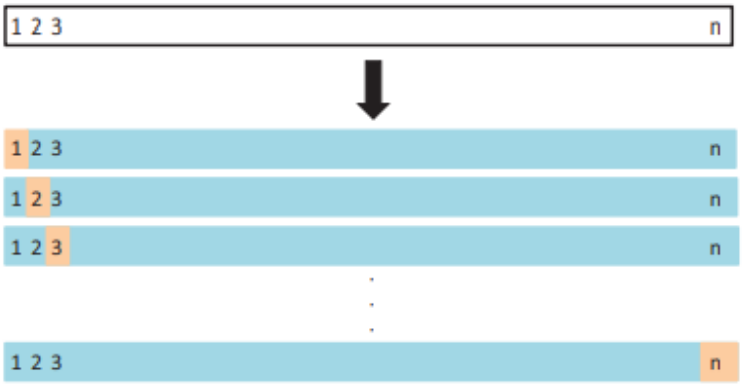

Рис. 5.3: Набір з n точок даних багаторазово розбивається на навчальний набір (синій), що містить усі спостереження, крім одного, і набір перевірки, який містить лише це спостереження (коричневий). Тестова похибка оцінюється шляхом усереднення n отриманих середньоквадратичних похибок. Перший тренувальний набір містить всі, крім спостереження 1, другий навчальний набір містить все, крім спостереження 2 і так далі...

5.6 Перехресна перевірка K Fold

Перехресна перевірка K-Fold передбачає випадковий поділ множини спостережень на k груп, або згини, приблизно рівного розміру. Перший згин розглядається як набір перевірки, а метод тренують на решті $k-1$ згинів. Далі рахують середньоквадратичну похибку MSE_1 на основі спостережень

у розгорнутому згині. Ця процедура повторюється k разів, кожного разу розглядається інша група спостережень як набір для перевірки. Цей процес призводить до k оцінок похибки тесту, $MSE_1, MSE_2, \ldots, MSE_k$. $K-$кратна оцінка CV обчислюється шляхом усереднення значень,

$$CV(k) = \frac{1}{k} \sum_{i=1}^{k} MSE_i.$$

6 Тренування і оцінка методів

6.1 Оцінка датасету

Ми будемо використовувати датасет з ресурсу Kaggle, а саме дані про 918 пацієнтів з параметрами: (вік, стать, рівень холестерину, стенокардія, рівень цукру та пульс) вихідним параметром є хворе серце в людини чи ні. Датасет поділений на тренувальні та тестові в пропорції 80/20.

На гістограмі 6.1 зображено розподіл здорових та хворих пацієнтів по відношенню до віку. Як можна побачити, люди похилого віку частіше мають серцеві захворювання.

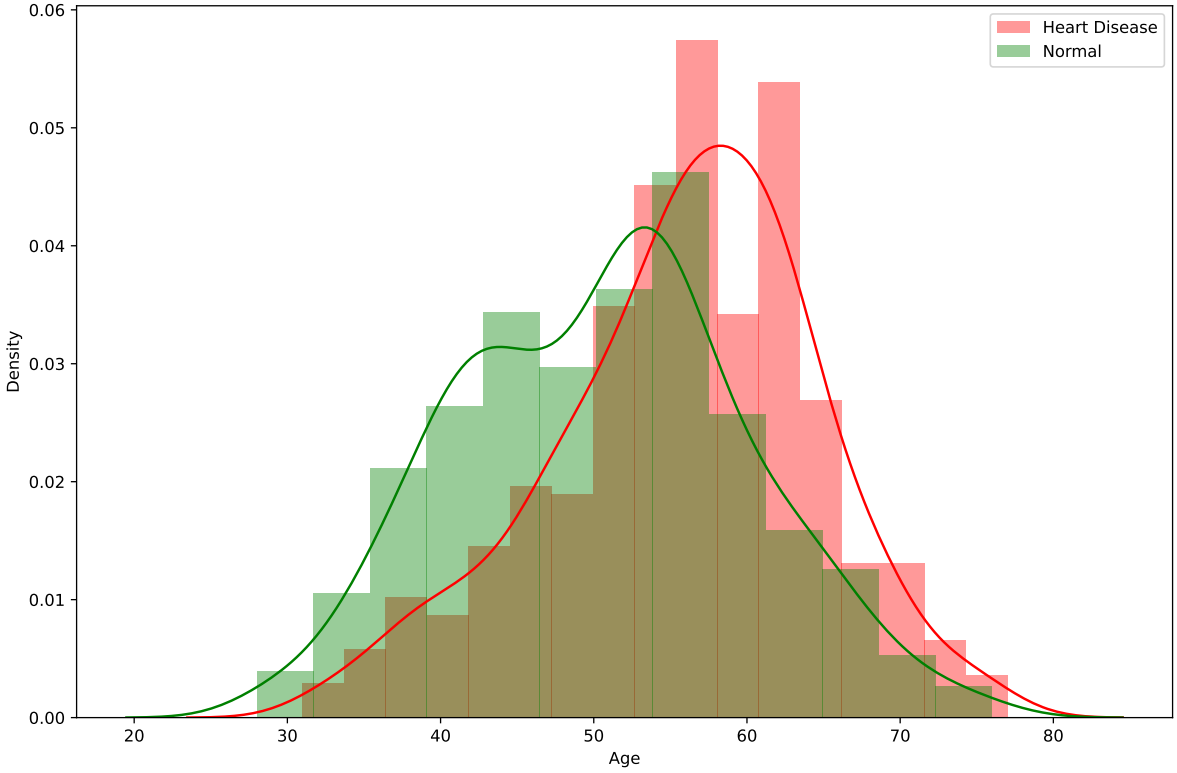

Рис. 6.1: Гістограма віку хворих та здорових пацієнтів.

Відповідно до графіка кореляції між параметрами датасету 6.2 можна сказати, що існує позитивна кореляція серцевих захворювань людей з низькою максимальною частотою серцебиття та людей похилого віку.

Таким чином, на основі попередніх спостережень можна зробити висновок про те які параметри є більш важливі для побудови ефективної моделі виявлення серцевих захворювань.

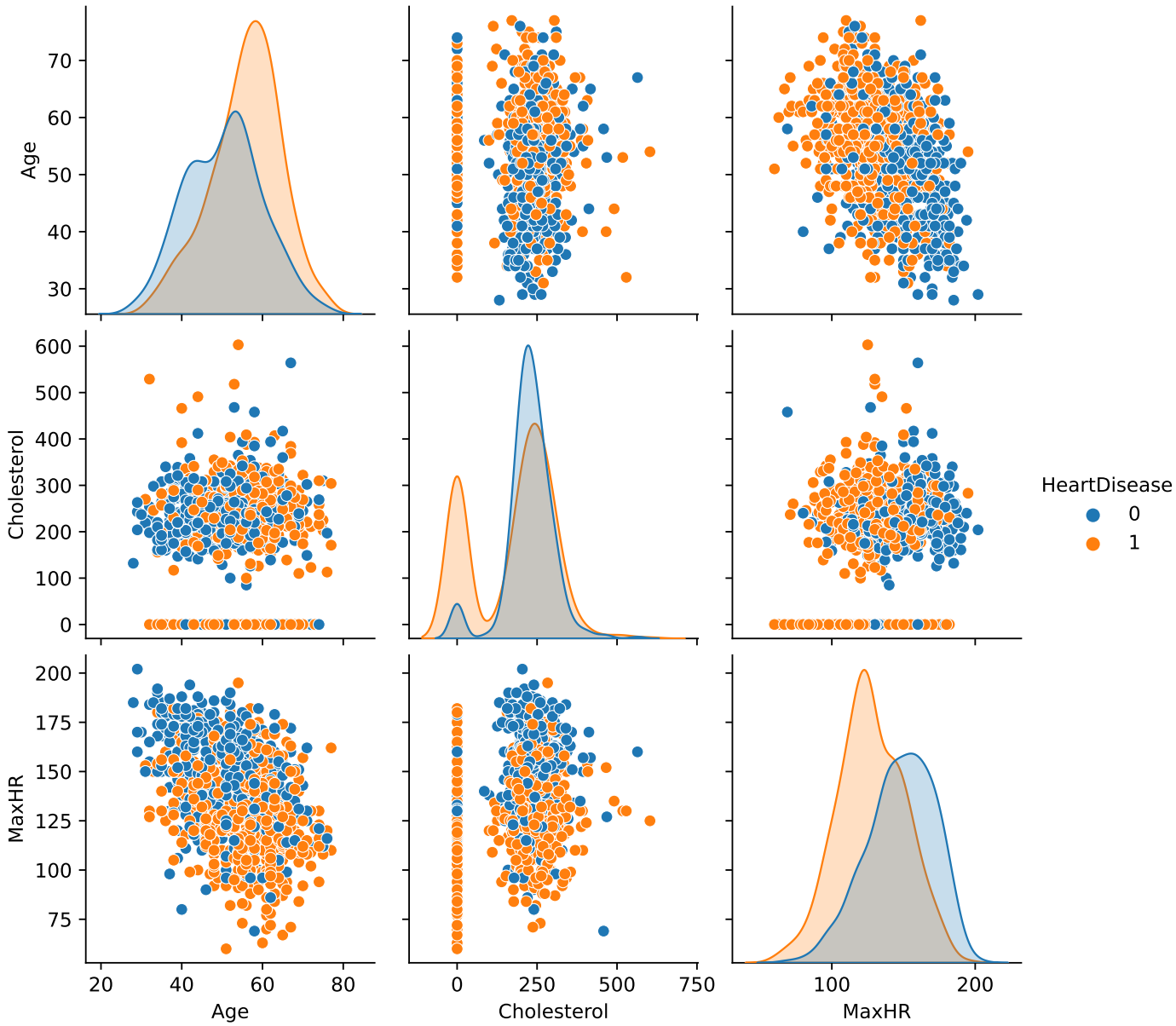

Рис. 6.2: Графік кореляції між параметрами датасету.

6.2 Оцінка методів

Дана задача є задачею класифікації. Для порівняння методів були розглянуті лінійні методи класифікації, а саме: логістична регресія та лінійний дискримінантний аналіз. Серед нелінійних методів були вибрані: метод k найближчих сусідів, квадратичний дискримінантний аналіз та нейронні мережі. Алгоритм логістичної регресії був реалізований самостійно коли інші алгоритми взяті з бібліотеки sklearn.

Модель	Точність	MSE	RMSE	Time (seconds)
К найближчих сусідів	72.28%	0.2771	0.5264	6.9e-7
Лінійний дискримінантний аналіз (LDA)	84.78%	0.1521	0.39	2.92e-3
Квадратичний дискримінантний аналіз (QDA)	74.45%	0.2554	0.5054	1.92e-3
Нейронні мережі	83.69%	0.163	0.4037	1.88
Логістична регресія	82.60%	0.1739	0.417	6,3e-2

Табл. 1: Оцінка передбачень моделі на тестовому наборі даних.

В таблиці 6.2 ми спостерігаємо ефективність лінійних та нелінійних моделей. Найкращий результат за точністю показує лінійний дискримінантний аналіз. Інший лінійний метод, а саме логістична регресія також показує добрий результат. Гіршим себе проявляє К наближених сусідів. Також можемо помітити що лінійні алгоритми загалом працюють швидше за нелінійні.

Модель	Leave One Out	K-Fold
К найближчих сусідів	76.6%	76.4%
Лінійний дискримінантний аналіз (LDA)	86.2%	85.8%
Квадратичний дискримінантний аналіз (QDA)	58.3%	68.8%
Нейронні мережі	84.6%	83.8%
Логістична регресія	85.7%	85.5%

Табл. 2: Таблиця оцінки моделей перехресної перевірки.

З таблиці перехресної перевірки 6.2 ми підтверджуємо наші спостереження щодо якості метода для нашої задачі. Лінійні методи проявляють себе краще.

7 Висновок

В ході цієї роботи ми розглянули зв'язок між статистичним та машинним навчанням. Розглянули ідею оцінки функції, різні категорії на які розподіляються статистичні методи та способи оцінювати та порівняння їх між собою.

Для поставленої задачі необхідно було обрати датасет разом з ним лінійні та нелінійні методи статистичного навчання. Також ми використали методи оцінки для наших статистичних навчань такі як перехресні перевірки, коефіцієнт помилок навчання та середня квадратична похибка.

За результатами експериментів можна сказати що для даного конкретного дадасету лінійні методи проявили себе краще як за часом виконання, так і за точністю результатів. Також для простих датасетів з переважно лінійними зв'язками в даних доцільніше використовувати лінійні методи навчання, адже вони менш схильні до перетренування.

Підсумовуючи наведене можна стверджувати що і лінійні та нелінійні методи мають свої переваги та недоліки. Чим більш обмежувальний та негнучкий є метод, тим простіше зрозуміти зв'язок між предикторами та відповіддю та він буде краще інтерпретуватись. А більш гнучкі методи краще апроксимує, але є більш вибагливими за кількістю тренувальних даних та потребують більше часу на тренування. Вибір найкращого підходу може бути однією з найскладніших частин виконання статистичного навчання. Для того, щоб оцінити ефективність статистичного методу навчання нам потрібен спосіб виміряти, наскільки добре його передбачення фактично відповідають спостережуваним даним.

Література

[1] Trevor Hastie, Robert Tibshirani та Jerome Friedman. *The Elements of Statistical Learning Data Mining, Inference, and Prediction*. https://link.springer.com/book/10.1007/978-0-387-84858-7. Springer, 2009.

[2] Gareth James та ін. *An Introduction to Statistical Learning with Applications in R*. https://www.statlearning.com/. Springer, 2013.

[3] Matthew Stewar. "The Actual Difference Between Statistics and Machine Learning". В: (2019). https://towardsdatascience.com/the-actual-difference-between-statistics-and-machine-learning-64b49f07ea3.

[4] Т. В. Борздова. "Основы статистического анализа и обработка данных". В: (2011). https://elib.bsu.by/bitstream, с. 4—12.