1. Методи агрегації даних (bagging) при налаштуванні класифікаторів.

Агрегація даних - це процес перетворення даних з високим ступенем деталізації до більш узагальненому уявленню. Ув'язнені в обчисленні так звані агрегати - значення, отримані в результаті застосування даних перетворень до певних факторів, пов'язаних з певним виміром. При цьому найчастіше використовується просте підсумовування, обчислення середнього або мінімального значення.

Беггінг (Bagging or Bootstrap Aggregating) – один з основних методів агреграції в машинному навчанні, який використовує паралельне навчання класифікаторів. Загалом, процес беггінгу можна описати наступними кроками:

1. Перш за все, із множини вихідних даних випадковим чином відбирається декілька підмножин, що містять в собі кількість об’єктів, що відповідає кількості об’єктів початкового набору.
2. Варто тримати в голові, що, оскільки відбір реалізується випадковим чином, то набір об’єктів завжди буде різним: деякі приклади потраплять в декілька підмножин, а деякі – в жодну.
3. На основі кожної вибірки будується класифікатор.
4. Алгоритми класифікації текстових даних можна класифікувати за таким чином:
    - класифікатор методом найближчих сусідів;
    - баєсівський класифікатор;
    - метод опорних векторів;
    - класифікація на основі асоціативних правил;
    - класифікатор на основі графових моделей;
    - класифікатор на основі дерева рішень;
    - класифікатор на основіштучних нейронних мереж.
5. Результати роботи класифікаторів агрегуються. Агрегація результатів, зазвичай, відбувається шляхом усереднення або голосування. При чому, перший варіант використовується в задачі регресії, а другий – класифікації. Використання беггінгу дозволяє зменшити відхилення.

Один з класичних алгоритмів класифікації, в основу якого покладено бегінг – це Випадковий Ліс або Random Forest. Випадковий ліс - це алгоритм, що базується на деревах рішень, який включає створення декількох дерев, а потім поєднання їх результатів для покращення можливостей узагальнення моделі. Метод об'єднання дерев відомий як метод ансамблю. Ансамбль - це нещо інше, як поєднання слабких учнів (окремих дерев) для створення сильного учня. Випадкові ліси стійкі до перенавчання. Кількість дерев можна збільшувати як завгодно, швидкість роботи все еще буде високою. Як працює випадковий ліс: навчальна

вибірка для поточного дерева складається шляхом відбору зразків (samples) із заміною, близько однієї третини випадків залишаються поза зразком. Ці дані, що мають назву “out-of-bag”, використовуються для отримання неупередженої оцінки помилок класифікації по мірі того, як нові дерева додаються до лісу. Цей підхід також використовується для отримання оцінок важливості змінних.

Після побудови кожного дерева всі дані пропускаються вниз по дереву, а близькість обчислюється для кожної пари випадків. Якщо два випадки потрапили в одну й ту ж кінцеву вершину (лист), значення їх близькості збільшується на один. В кінці пробігу, показники близькості нормалізуються шляхом ділення на кількість дерев. У випадкових лісах відсутня потреба в крос-валідації або окрема тестова вибірка для отримання об'єктивної оцінки помилки тестового набору. 53 Внутрішньо, під час запуску, вона оцінюється наступним чином: кожне дерево конструюється за допомогою іншого зразка бутстреп з вихідних даних. Близько однієї третини випадків залишаються поза піднабором та не використовуються при побудові k-го дерева. Кожен об’єкт, що не був використаний при побудові дерева k-го дерева, пропускається через алгоритм, щоб отримати класифікацію. Таким чином, класифікація об’єктів тестової вибірки проводиться в середньому третиною дерев. В кінці пробігу для кожного об’єкта обирають j-тий клас, який отримав більшу кількість голосів кожного разу, коли цього об’єкта не було в навчальній вибірці. Частка випадків, коли j не дорівнює дійсному класу n, усереднена по всіх випадках, дає нам оцінку помилки

У кожному дереві, вирощеному в лісі, відкладемо випадки out-of-bag і підрахуємо кількість голосів, «відданих» за правильний клас. Тепер випадково перемішаємо значення змінної m і пропустимо через у дерево. Віднімемо кількість голосів за правильний клас у змінних даних out-of-bag від кількості голосів за правильний клас на недоторканих даних. Середнє значення цього числа у всіх деревах лісу є значенням важливості для змінної m.

Якщо значення цього показника від дерева до дерева незалежні, то стандартна помилка може бути обчислена за допомогою стандартного обчислення. Кореляції цих оцінок між деревами були обчислені для ряду наборів даних і виявилися досить низькими, тому ми класично вираховуємо стандартні похибки, ділимо оцінку рядку за стандартною помилкою, щоб отримати z-бал, і призначати рівень значущості до z-оцінки припускаючи нормальність. Якщо кількість змінних є дуже великою, ліси можуть бути запущені один раз з усіма змінами, а потім запустити знову, використовуючи лише найважливіші змінні з першого запуску. У деяких наборах даних помилка прогнозування між класами вельми незбалансована. Деякі класи мають низьку помилку прогнозування, інші - високу.

Це відбувається зазвичай, коли один клас набагато більшим, ніж інший. Тоді випадкові ліси, намагаючись мінімізувати загальну частоту помилок, будуть підтримувати низький рівень помилок у великому класі, дозволяючи меншим класам мати більший рівень помилок. Наприклад, при відкритті ліків, 54 коли дана молекула класифікується як активна чи ні, загальним є те, що кількість активів перевищує 10 на 1, до 100 до 1. У таких випадках частота помилок на цікавому класі (активах) буде дуже високим. Зазвичай, користувач може виявляти дисбаланс за виводами коефіцієнтів помилок для окремих класів. Щоб штучно збалансувати помилку в таких випадках, окремим класам назначають вагові коефіцієнти. Чим більша вага класу, тим менша помилка на ньому. Окрім цього, використовується метод усічення (pruning), коли видаляються вузли нижчих рівнів.

Окремим розширенням ідеї методу випадкових лісів є ротаційний ліс або Rotation Forest, основна особливість якого полягає в підготовці даних для навчання простих класифікаторів. До піднаборів, отриманих випадковим розбиттям початкової вибірки, застосовується метод головних компонент. Основною метою методу, за словами розробників, є заохочення одночасно індивідуальної точності та різноманітності в ансамблі.

2. Перевірити, чи належить ядерна функція k(x, y) = exp(x + y^2) - ln(tg(x/y) + 2) до класу ядерних функцій Мерсена.

Перевірку виконати на множині X={2,4,8,6,4,1}, Y = {4,6,8,5,3,1}. Аргументи тригонометричних функцій задано в радіанах.

$x := (2 \quad 4 \quad 8 \quad 6 \quad 4 \quad 1)$

$y := (4 \quad 6 \quad 8 \quad 5 \quad 3 \quad 1)$

$k(x,y) := e^{x+y^2} - \ln\left[\left[\tan\left(\left(\frac{x}{y}\right)\right) + 2\right]\right]$

$i := 0..5$

$j := 0..5$

$M_{i,j} := k(x_{0,i}, y_{0,j})$

$$M = \begin{pmatrix} 6.566 \times 10^{7} & 3.186 \times 10^{16} & 4.607 \times 10^{28} & 5.32 \times 10^{11} & 5.987 \times 10^{4} & 21.773 - 3.142i \\ 4.852 \times 10^{8} & 2.354 \times 10^{17} & 3.404 \times 10^{29} & 3.931 \times 10^{12} & 4.424 \times 10^{5} & 147.263 \\ 2.649 \times 10^{10} - 3.142i & 1.285 \times 10^{19} & 1.859 \times 10^{31} & 2.146 \times 10^{14} & 2.415 \times 10^{7} & 8.102 \times 10^{3} - 3.142i \\ 3.585 \times 10^{9} & 1.739 \times 10^{18} & 2.515 \times 10^{30} & 2.905 \times 10^{13} & 3.269 \times 10^{6} - 3.142i & 1.096 \times 10^{3} \\ 4.852 \times 10^{8} & 2.354 \times 10^{17} & 3.404 \times 10^{29} & 3.931 \times 10^{12} & 4.424 \times 10^{5} & 147.263 \\ 2.415 \times 10^{7} & 1.172 \times 10^{16} & 1.695 \times 10^{28} & 1.957 \times 10^{11} & 2.203 \times 10^{4} & 6.12 \end{pmatrix}$$

$M1 := \text{eigenvals}(M)$

$$M1 = \begin{pmatrix} -0.939 + 7.787i \times 10^{-3} \\ 1.859 \times 10^{31} \\ 3.529 \times 10^{7} + 0.339i \\ 4.656 \times 10^{3} + 0.305i \\ -2.438 - 0.161i \\ -0.483 - 0.109i \end{pmatrix}$$

Відповідь: Ні