Розглянуті методи класифікації даних

K-Найближчі сусіди

K-Найближчі сусіди працює, перевіряючи відстань від деяких тестових прикладів до відомих значень деяких тренувальних прикладів. Група точок даних / класу, які давали б найменшу відстань між точками навчання та точкою тестування, є вибраний клас.

Лінійний дискримінантний аналіз

Лінійний дискримінантний аналіз працює, зменшуючи розмірність набору даних, проектуючи всі точки даних на лінію. Потім він поєднує ці точки в класи залежно від їх віддаленості від обраної точки або центроїда. Як ви вже можете здогадатися, лінійний дискримінантний аналіз - це лінійний алгоритм класифікації, який найкраще застосовувати, коли дані мають лінійний зв’язок.

**Support Vector Machines**

Support Vector Machines працює, проводячи лінію між різними кластерами точок даних, щоб згрупувати їх у класи. Окуляри на одній стороні лінії будуть одним класом, а точки на іншій стороні належать до іншого класу. Класифікатор намагатиметься максимізувати відстань між проведеною ним лінією та точками по обидва боки від неї, щоб збільшити свою впевненість у тому, які точки належать до якого класу. Коли нанесені точки тестування, сторона лінії, на яку вони потрапляють, є класом, в який вони поставлені.

Моделі CART

Моделі CART популярні з кількох причин: їх легко інтерпретувати, вони легко працюють з змішаними дискретними та безперервними входами, вони нечутливі до монотонних перетворень входів (оскільки точки розділення базуються на ранжуванні точок даних), вони виконують автоматично змінний вибір, вони добре масштабуються до великих наборів даних і їх можна модифікувати для обробки відсутніх вхідних даних.

Однак у моделей CART є і деякі недоліки. Первинний - це те, що вони не дуже точно передбачають в порівнянні з іншими типами моделей. Це частково пов'язано з жадібний характер алгоритму побудови дерева. Пов'язана проблема полягає в тому, що дерева нестабільні: Невеликі зміни вхідних даних можуть мати великий вплив на структуру дерева через ієрархічний характер процесу вирощування дерев, що призводить до помилок у верхній частині, що зачіпає решту дерево. У частотній термінології ми говоримо, що дерева є оцінювачами великих дисперсій. Як рішення цих проблем можемо розглядати:

Випадкові ліси

Один із способів зменшити дисперсію оцінки - усереднити разом багато оцінок. Наприклад, ми можемо навчити M різних дерев на різних підмножинах даних, вибраних випадковим чином із заміною, а потім обчислити:

C:\Users\08041\Desktop\схема\QIP Shot - Screen 2110.png

Список джерел:

1. Kevin P. Murphy. *Machine Learning. A Probabilistic Perspective*. 1st edition. The MIT Press, 2012
2. https://stackabuse.com/overview-of-classification-methods-in-python-with-scikit-learn/