**"Дейтасет Ирис"**

Наборът от данни "Ирис" съдържа четири характеристики (дължина и ширина на чашките и листата) на 50 броя от три вида Ирис (Iris Setosa, Iris virginica и Iris Versicolor). Тези измервания са използвани за създаване на линеен дискриминантен модел за класификация на видовете. Наборът от данни се използва често в примери за добив на данни, класификация и кластеризация, както и за тестване на алгоритми.

Информация за оригиналната статия и приложения на набора от данни може да бъде намерена в UCI Machine Learning Repository -- Iris Data Set. Наборът от данни съдържа множество от 150 записа под пет атрибута: дължина на чашката, ширина на чашката, дължина на листата, ширина на листата и вид. Колоните в този набор от данни са:

Id (Идентификационен номер)

SepalLengthCm (Дължина на чашката в см)

SepalWidthCm (Ширина на чашката в см)

PetalLengthCm (Дължина на листата в см)

PetalWidthCm (Ширина на листата в см)

Species (Вид)

Първоначално използван като примерен набор от данни, върху който е приложен линеен дискриминантен анализ на Фишър, този набор става типичен тестов случай за множество статистически методи за класификация в машинното обучение, като например метода на опорните вектори.

Въпреки това, използването на този набор от данни за анализ на клъстъри не е обичайно, тъй като той съдържа само два клъстера с доста ясно разграничие. Единият клъстер съдържа Ирис сетоза, докато другият клъстер съдържа както Ирис вирджиника, така и Ирис версиколор, и не може да бъде разделен без информацията за видовете, използвана от Фишър. Това прави набора от данни добър пример за обяснение на разликата между контролирани и неконтролирани методи в добива на данни: линейният дискриминантен модел на Фишър може да бъде получен само когато видовете на обектите са известни; класовете и клъстерите не са задължително едно и също нещо.

Въпреки това, всички три вида Ирис са различими в проекцията върху нелинейния и разклоняващ се основен компонент. Наборът от данни се апроксимира с най-близкото дърво с някакво отклонение за излишния брой възли. След това се строи така наречената "метрокарта". Точките се проектират в най-близкия възел. За всеки възел се изготвя кръгова диаграма на проектираните точки. Площта на кръговата диаграма е пропорционална на броя на проектираните точки. От диаграмата става ясно, че абсолютното мнозинство на образците от различните видове Ирис принадлежат на различни възли. Само малка част от Ирис вирджиника се смесва с Ирис версиколор (смесените синьо-зелени възли в диаграмата). Следователно три вида Ирис (Ирис сетоза, Ирис вирджиника и Ирис версиколор) могат да бъдат разграничени чрез неконтролирани методи на нелинеен анализ на основни компоненти. За да ги различите, е достатъчно просто да изберете съответните възли на основното дърво.