Дискриминантный линейный анализ (Linear discriminant analysis)

Разделы: Алгоритмы

Линейный дискриминантный анализ (ЛДА), а также связанный с ним **линейный дискриминант Фишера** — методы статистики и машинного обучения для нахождения линейных комбинаций признаков, наилучшим образом разделяющих два или более класса объектов или событий. Полученная комбинация может быть использована в качестве линейного классификатора или для сокращения размерности пространства признаков перед последующей классификацией.

ЛДА представляет собой раздел многомерного статистического анализа, содержанием которого является разработка методов решения задач различения (дискриминации) объектов наблюдения по набору признаков. Иными словами, он позволяет изучать различия между двумя и более группами объектов по нескольким признакам одновременно.

ЛДА тесно связан с дисперсионным анализом и регрессионным анализом, также пытающимися выразить какую-либо зависимую переменную через линейную комбинацию других признаков или измерений. В этих двух методах зависимая переменная — численная величина, а в ЛДА она является величиной номинальной (меткой класса). Помимо того, ЛДА имеет схожие черты с методом главных компонент и факторным анализом, которые ищут линейные комбинации величин, наилучшим образом описывающие данные.

Можно выделить три вида задач дискриминантного анализа:

- определение дискриминирующих признаков (т.е. признаков, которые позволяют отнести наблюдение к той или иной группе);
- построение дискриминирующей функции;
- прогнозирование будущих событий, связанных с попаданием объекта в ту или иную группу на основе значений его признака (например, предсказание выживаемости пациента после операции).

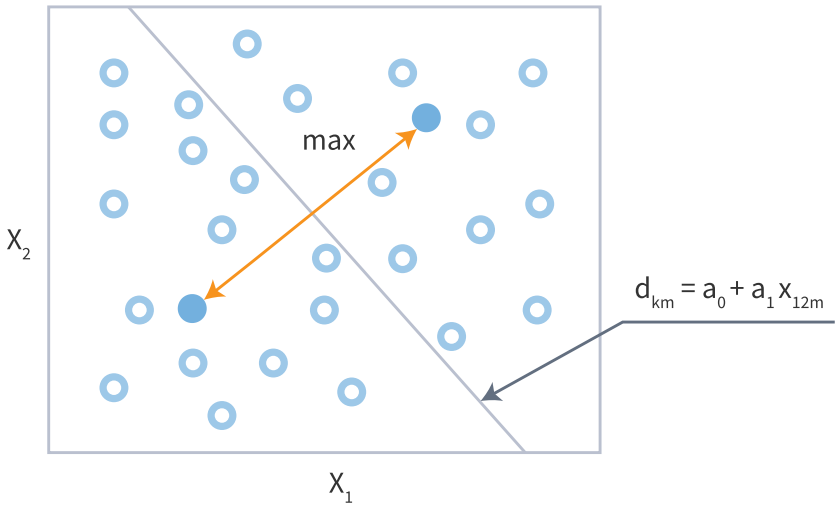
Основной целью дискриминации является поиск линейной комбинации признаков (называемых дискриминантными признаками), которые позволили бы наилучшим образом разделить рассматриваемые группы.

Рассмотрим линейную функцию, называемую канонической дискриминантной функцией (КДФ):

$$d_{km} = a_0 + a_1 x_{1km} + a_2 x_{2km} + \ldots + a_p x_{pkm},$$

где d_{km} — значение дискриминирующей функции для m-го наблюдения k-й группы, $m = 1..n$, $k = 1..g$, x_{2ikm} — значение дискриминантного признака для m-го наблюдения k-й группы, p — число дискриминантных признаков (размерность многомерного пространства).

С геометрической точки зрения КДФ определяет гиперповерхности в p-мерном пространстве. При $p = 2$ она будет прямой, а при $p = 3$ — плоскостью. Коэффициенты a_i первой КДФ выбираются так, чтобы центроиды различных групп как можно больше отличались друг от друга.



Для случая, представленного на рисунке, прямая должна разбить пространство признаков (x_1, x_2) таким образом, чтобы расстояние между центроидами результирующих подмножеств было максимально возможным.

Коэффициенты второй КДФ выбираются так же, но при этом налагается дополнительное ограничение, чтобы значения второй функции не коррелировали со значениями первой. Аналогично определяются и другие функции. Отсюда следует, что любая КДФ d имеет нулевую внутригрупповую корреляцию с $d_1, \ldots d_{g-1}$.

Если число групп равно g, то число КДФ будет на единицу меньше числа групп. Однако удобно использовать одну, две или три КДФ, поскольку графическое изображение объектов в этом случае будет представлено в одно-, двух- и трехмерных пространствах. Такое представление особенно полезно в случае, когда число дискриминантных признаков велико по сравнению с числом групп.

ЛДА широко используется для решения задач классификации и распознавания образов, понижения размерности входных данных. Хотя он и работает с информацией, которая определяет принадлежность объекта к одному из классов, но сам по себе классификатором не является, а используется как часть линейной классификационной модели.

Преимущество метода — сравнительная простота реализации и интерпретации результатов. Недостаток — чувствительность к распределению исходных данных, когда даже небольшое их изменение приводит к значительным изменениям результатов классификации.

Основные идеи дискриминантного анализа были сформулированы Роналдом Фишером в 1936 г.