**Общая информация**

Дискриминантный анализ (discriminant analysis) – это метод многомерного статистического анализа, используемый для принятия решения о том, какие переменные различают (дискриминируют) две или более возникающие совокупности (группы).

Например, некий исследователь в области образования может захотеть исследовать, какие переменные относят выпускника средней школы к одной из трех групп: (1) поступающий в колледж, (2) поступающий в профессиональную школу или (3) отказывающийся от дальнейшего образования или профессиональной подготовки. Для этой цели исследователь может собрать данные о различных переменных, связанных с учащимися школы. Затем можно использовать Дискриминантный анализ для определения того, какие переменные дают наилучшее предсказание.

В этом и заключается основная идея ДА - определить, отличаются ли совокупности по среднему какой-либо переменной (или линейной комбинации переменных), и затем использовать эту переменную, чтобы предсказать для новых членов их принадлежность к той или иной группе.

*При применении дискриминантного анализа обычно имеются несколько переменных, и задача состоит в том, чтобы установить, какие из переменных вносят наибольший вклад в дискриминацию между совокупностями. В этом случае мы имеем матрицу общих и внутригрупповых дисперсий и ковариаций. Мы можем сравнить эти две матрицы с помощью многомерного F-критерия для того, чтобы определить, имеются ли значимые различия между группами.*

**Виды ДА**

Вероятно, наиболее общим применением дискриминантного анализа является включение в исследование многих переменных с целью определения тех из них, которые наилучшим образом разделяют совокупности между собой.

Опишу 2 возиожных анализа

*Пошаговый анализ с включением.* В нем модель дискриминации строится по шагам. На каждом шаге просматриваются все переменные и находится та из них, которая вносит наибольший вклад в различие между совокупностями. Эта переменная должна быть включена в модель на данном шаге, и происходит переход к следующему шагу.

*Пошаговый анализ с исключением*. Двигается в обратном направлении. В этом случае все переменные будут сначала включены в модель, а затем на каждом шаге будут устраняться переменные, вносящие малый вклад в предсказания. Тогда в качестве результата успешного анализа можно сохранить только важные переменные в модели, то есть те переменные, чей вклад в дискриминацию больше остальных.

*F для включения, F для исключения. Эта пошаговая процедура руководствуется соответствующим значением F для включения и соответствующим значением F для исключения. Значение F-статистики для переменной указывает на её статистическую значимость при дискриминации между совокупностями. то есть она является мерой вклада переменной в предсказание членства в совокупности.*

**Дискриминантные функции для двух групп**

*Для двух групп дискриминантный анализ может рассматриваться также как процедура множественной регрессии (и аналогичная ей).*

Ядром дискриминантного анализа является построение так называемой дискриминантной функции

Group = С + b1\*x1 + b2\*x2 + ... + bm\*xm

где

* С - константа
* b1...bm - коэффициенты регрессии
* x1 и хn - значения переменных

Необходимо определить такие коэффициенты, чтобы по значениям дискриминантной функции можно было с максимальной четкостью провести разделение по группам.

Описанное уравнение регрессии составляется на основе тех объектов, о которых известна групповая принадлежность, что позволяет максимально точно подобрать его коэффициенты.

*Интерпретация результатов задачи с двумя совокупностями тесно следует логике применения множественной регрессии: переменные с наибольшими регрессионными коэффициентами вносят наибольший вклад в дискриминацию.*

**Дискриминантные функции для нескольких групп**

**Для нескольких групп можно оценить несколько дискриминантных функци.**

Например, когда имеются три совокупности, можно оценить:

1. Функцию для дискриминации между совокупностью 1 и совокупностями 2 и 3, взятыми вместе
2. Другую функцию для дискриминации между совокупностью 2 и совокупности 3.

Например, мы можем иметь одну функцию, дискриминирующую между теми выпускниками средней школы, которые идут в колледж, против тех, кто этого не делает (но хочет получить работу или пойти в училище), и вторую функцию для дискриминации между теми выпускниками, которые хотят получить работу против тех, кто хочет пойти в училище.

**Процедура дискриминантного анализа**

1. Разделение выборки на две части.

2. Выбор переменных – предикторов.

3. Вычисление параметров дискриминантной функции.

Разделение происходитт в процентном соотношении 70 к 30. Обучающая и тестовые выборки.

Затем происходит выбор переменных

Замечу, что если число переменных достаточно велико (например, несколько сотен), то не представляется возможным применить дискриминантный анализ ко всем переменным одновременно.

Поэтому

• Далее на начальном этапе дискриминантного анализа для предикторов формируется корреляционная матрица. В данном контексте она имеет особый смыс и называется общей внутригрупповой корреляционной матрицей и содержит средние коэффициенты корреляции для двух или более корреляционных матриц (каждая для одной группы).

• помимо общей внутригрупповой корреляционной матрицы можно также вычислить ковариационные матрицы для отдельных групп, либо общую внутригрупповую ковариационную матрицу для всей выборки.

*Вычисление параметров дискриминантной функции производится одним из описанных ранее анализов.*

**Классификация**

Как только модель установлена и получены дискриминирующие функции, возникает вопрос о том, как хорошо они могут предсказывать? В этом помогут **функции классификации**

Они предназначены для определения того, к какой группе наиболее вероятно может быть отнесен каждый объект. Имеется столько же функций классификации, сколько групп. Каждая функция позволяет нам для каждого образца и для каждой совокупности вычислить веса классификации по представленной формуле формуле:

Si = ci + wi1\*x1 + wi2\*x2 + ... + wim\*xm

i обозначает соответствующую совокупность, а индексы

m переменных;

ci являются константами для i-ой совокупности,

wij - веса для j-ой переменной при вычислении показателя классификации для i-ой совокупности;

xj - наблюдаемое значение для соответствующего образца j-ой переменной.

Величина Si является результатом показателя классификации

В общем случае наблюдение считается принадлежащим той совокупности, для которой получен наивысший показатель классификации.

*Если мы хотим определить вероятность принадлежности, можем воспользоваться расстоянием Махаланобиса.*

*Для каждой совокупности в выборке мы можем определить положение точки, представляющей средние для всех переменных рассматриваемой модели в многомерном пространстве. Эти точки называются центроидами группы.*

*Для каждого наблюдения вычисляем его расстояние Махаланобиса от каждого центроида группы.*

*В общем, расстояние Махаланобиса является мерой расстояния между двумя точками в пространстве, определяемым двумя или более коррелированными переменными.*

*Наблюдение принадлежит к той группе, к которой оно ближе, т.е. когда расстояние Махаланобиса до нее минимально.*

Следует заметить, что Классификация действует лучшим образом для выборки, по которой была проведена оценка дискриминантной функции (апостериорная классификация), чем для свежей выборки (априорная классификация). Поэтому оценивание качества процедуры классификации лучше не производить по той же самой выборке, по которой была получена дискриминантной функция.

**По итогу,** дискриминантный анализ – это очень полезный инструмент для поиска переменных, позволяющих относить наблюдаемые объекты в одну или несколько реально наблюдаемых групп, и для классификации наблюдений в различные группы.