3.4.Дискриминантный анализ

Дискриминантный анализ является мощным статистическим средством анализа разделения (дискриминации) многомерных совокупностей наблюдений на группы. Данный анализ был разработан независимо друг от друга известными статистиками Р.Фишером, Г.Хоттелингом и Махаланобисом.

Основной задачей дискриминантного анализа является анализ имеющегося разделения, дискриминации исследуемого множества объектов, характеризующихся набором закоррелированных выходных характеристик, и разделенных на две или большее число групп. В результате такого анализа оцениваются связи между имеющимся делением на группы и вкладами отдельных количественных признаков в это деление.

Линейная дискриминантная функция преобразует множество измерений, входящих в выборку в значение дискриминантной функции (дискриминантную метку). Последним определяется граница разделения совокупностей.

Для нахождения линейной дискриминантной функции строится уравнение регрессии, в котором в качестве зависимых переменных выступают разности между многомерными средними двух анализируемых групп Hj = Uj - Vj. С этой целью решается матричное уравнение вида [C] [a] = [H] (где [C] – ковариационная матрица объединенной выборки размерности m x m; [a] – вектор – столбец коэффициентов дискриминантной функции; [H] – вектор – столбец разностей между средними значениями признаков групп). Уравнение решается с помощью операции обращения и умножения матриц $[a]=[C]^{-1}$ [H], после чего по вычисленным коэффициентам a_1, a_2, …, a_p, …, a_k строится дискриминантная функция

$$DX_1, X_2, \ldots, X_p, \ldots, X_k = \sum_{p=1}^{k} a_p X_p.$$

Дискриминантный индекс D_0, определяющий принадлежность испытуемого объекта к той или иной группе, рассчитывается по формуле

$$D_0 = 1/2 \sum_{p=1}^{k} a_p \, (\bar{u}_p + \bar{v}_p).$$

Он соответствует середине «обобщенного расстояния» между центрами обеих групп в многомерном пространстве изучаемых признаков.

После расчета порогового значения (дискриминантного индекса) D_0 оцениваются значения центров обеих исходных групп D_u и D_v, путем подстановки начала значений u_j, а затем v_j в уравнение дискриминантной функции. «Обобщенное расстояние» между двумя многомерными средними D^2 называют расстоянием (или критерием) Махаланобиса. Оно вычисляется путем вычитания D_u из D_v, что эквивалентно подстановке вектора разности между средними двух групп u_j и v_j в дискриминантную функцию

$$D^2 = \sum_{i=1}^{p} a_j \,(u_j - v_j).$$

Чем больше $D^{2,}$ тем более уверенное разделение можно провести между объектами (А.Б.Кождан, О.И.Гуськов, 1990). Критерий Хотеллинга T^2 для проверки значимости дискриминантной функции выражается через это расстояние. Если $T^2_{расч} > T^2_{табл.}$ и p< a, то нулевая гипотеза о равенстве двух многомерных средних или о равенстве расстояния между ними нулю отвергается. В дискриминантном анализе, получив оценку дискриминантной функции и подставляя в нее значения $X_{1,}$ X_2, … X_p, можем заново, более точно произвести переклассификацию наблюдений, относя их в зависимости от значения функции к той или другой группе. В этом и заключается основное отличие от дисперсионного анализа – в возможности апостериорной переклассификации наблюдений.