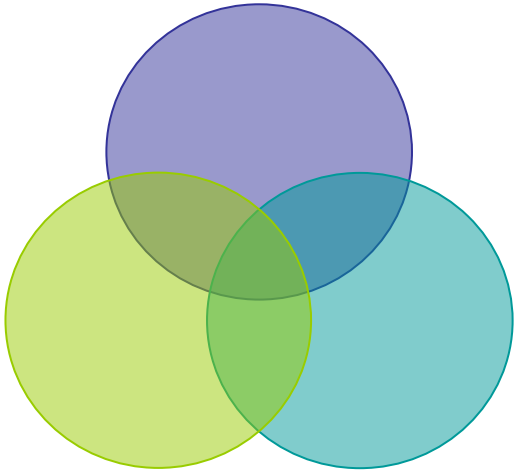
Лекция № 9

ДИСКРИМИНАНТНЫЙ АНАЛИЗ



ВОПРОСЫ

1. Назначение дискриминантного метода анализа данных
2. Математико-статистические идеи метода
3. Исходные данные и основные результаты

1. НАЗНАЧЕНИЕ ДИСКРИМИНАНТНОГО АНАЛИЗА

- **Дискриминантный анализ** это метод многомерной классификации, позволяющий разделять множество испытуемых (объектов) на группы (классы) на основе количественных характеристик объектов.
- **Авторы метода –** П. Махаланобис, Р. Фишер, Г. Хоттелинг

Структура данных для дискриминантного анализа

№	Y	X_1	X_2	$\dots X_i \dots$	X_k
1	y_1	x_{11}	x_{12}	$\dots \dots$	x_{1k}
2	y_2	x_{21}	x_{22}	$\dots \dots$	x_{2k}
3	y_3	x_{31}	x_{32}	$\dots \dots$	x_{3k}
$\dots i \dots$	$\dots y_i \dots$	$\dots x_{i1} \dots$	$\dots x_{i2} \dots$	$\dots \dots$	x_{ik}
N	y_N	x_{N1}	x_{N2}	$\dots \dots$	x_{Nk}

Примечания к данным

Столбцы – независимые переменные Xi,

где $i = 1,2…k;$

– результативный признак (зависимая переменная) Yj, где $j = 1,2…G$.

Строки – показатели N испытуемых:

X_{ik} – количественные;

Yj – номинальные (классифицирующие).

Исходными данными для анализа является группа из N объектов (испытуемых), разделенных на G классов так, что каждый объект отнесен только к одному классу (градации результативного признака – классифицирующей переменной).

Основные задачи метода

1. Интерпретация **различия между классами** (признаки объектов, используемые для этого, называются ***дискриминантными переменными*** и необходимы для получения значения **классифицирующей переменной**).
2. Проведение классификации новых объектов («**распознавание образа**») по измеренным для него дискриминантным переменным.

2.Математико-статистические идеи метода

- Если дискриминатные переменные представить себе как ортогональные оси k-мерного евклидова пространства, то каждый объект будет точкой в этом пространстве, а координатами положения точки будут являться числовые значения его дискриминантных переменных.

- Множество объектов в пространстве признаков можно представить как скопление точек.
- Если несколько классов объектов отличаются по своим дискриминантным признакам, то их можно представить как определенные области пространства признаков, в каждой из которых объекты похожи друг на друга и отличаются от объектов другого класса.
- Для каждого класса определяют положение **центроида** – точки, координаты которой есть средние значения переменных.

- Из геометрической интерпретации задачи дискриминатного анализа следует **правило классификации объектов**: объект приписывается к тому классу, к центроиду которого он ближе всего.
- Таким образом, задача классификации сводится к определению расстояний от каждого объекта (испытуемого) до центроидов каждого класса по известным значениям дискриминантных переменных (признаков).
- **Главный центроид –** точка с координатами N средних значений признаков X_{ik}

- В компьютерных программах задача классификации решается с помощью **дискриминантных функций**. Эти функции представляют ортогональные оси, в максимальной степени различающие центроиды классов.
- **Первая ось** ориентирована в направлении, в котором центроиды классов различаются в максимальной степени.

- Максимальное число дискриминантных функций на 1 меньше числа классов.
- Таким образом, дискриминантные функции позволяют преобразовать *k*-мерное пространство исходных признаков в Q-мерное пространство дискриминантных функций (Q = G - 1).
- Если классов больше двух, то вторая ось ориентирована перпендикулярно первой в направлении максимального разделения классов и т.д.

- **Значения дискриминантных функций** вычисляются для каждого объекта по формуле идентичной линейному уравнению МРА, которая максимизирует различия между классами и минимизирует дисперсию внутри класса.

$$Y_{ij} = b_{jo} + b_{j1}x_{1i} + b_{j2}x_{2i} + \ldots + b_{jk}x_{ki}$$

где Y_{j0} – значение функции ***j*** $(j = 1,2 \ldots G)$ для объекта ***i***, а $b_{jo}, \ldots, b_{jk}$ – канонические коэффициенты для каждой из дискрииминантных переменных.

- Значение дискриминантных функций вычисляются для каждого центроида и для каждого объекта.
- Это позволяет в пространстве дискриминантных функций получить наглядное отображение все объектов вместе с центроидами классов.
- Канонические коэффициенты позволяют оценить относительный вклад переменных в дискриминантную функцию, т.е. оценить различительную способность функции.

3.Исходные данные и основные результаты

Исходными данными для анализа является группа из N объектов (испытуемых), разделенных на G классов так, чтобы в каждом классе содержалось не менее двух объектов. Для каждого из них имеются количественные данные по K переменным. Рекомендуется двукратное превышение числа объектов над числом переменных.

Предполагается нормальное распределение показателей каждого признака.

Между дикриминантными признаками не может быть функциональной зависимости, т.е. значений коэффициентов корреляции равными 1.

Основные результаты дискриминантного анализа

1. Определение статистической значимости различения классов при помощи данного набора дискриминантных переменных.

К основным статистическим показателям относятся: собственные значения дискриминантной функции, **процент дисперсии** дискриминативной возможности, лямбда **λ-Вилкса**, критерий χ^2, статистическая значимость (***p*-уровень**).

Собственное значение деленное на количество классов определяет показатель информативности канонической функции - долю суммарной дисперсии всех объектов по всем переменным.

Лямбда λ-Вилкса определяет долю остаточной дискриминативной способности переменных при учете данного набора канонических функций. Чем меньше λ-Вилкса, тем лучше данная каноническая функция различает объекты.

Критерий χ^2 позволяет определить статистическую достоверность (*p*-уровень) такого различия.

2. **Классификация «известных» и «неизвестных» объектов** при помощи расстояний или значений априорной вероятности.

Качество классификации определяется совпадением действительной классификации и предсказанной для «известных» объектов. Мерой качества может служить вероятность ошибочной классификации как соотношение количества ошибочного отнесения к общему количеству «известных» объектов.

3. Выяснение **вклада каждой переменной** в дискриминантный анализ (по значениям критерия Фишера).

4. Вычисление расстояний между центроидами классов и определения их статистической значимости (по критерию Фишера).

5. Анализ канонических функций, их интерпретация через дискриминантные переменные (по стандартизованным и структурным коэффициентам канонических функций).

Спасибо за внимание!!!

Продолжение смотрите
2 апреля 2015 года