**Работа №3:** Классификация объектов на основе дискриминантного анализа

(*Афифи, Эйзен: Статистический анализ. Подход с использованием ЭВМ, гл.5.3 Классификация индивидуума (т.е.объекта) в случае двух популяций*)

а) построение и тестирование классификатора с использованием модельных данных;

 b) построение и тестирование классификатора с использованием данных из репозитория;

**Порядок выполнения работы**

1. ***параметры распределения известны***
2. *Построение дискриминантной функции (ДФ)*

Пусть первая популяция W1 (группа, класс) представлена обучающей выборкой 1 (ОВ1) объема *n*1 из нормального распределения: ;

вторая популяция W2 представлена обучающей выборкой 2 (ОВ2) объема *n*2 из нормального распределения:

∈, ∈ (матрицы ковариаций)**– известны,**  , *m,j*∈ **!!!**

Тогда решающее правило , т.е. дискриминантная функция (ДФ), имеет вид:

**, (\*)**

где вектор коэффициентов, **α**∈, находится по формуле: *.*

Общее правило классификации вектора будет выглядеть следующим образом:

(1)

Формулы для вычисления константы *c* и расстояния Махаланобиса :

*c=*)*/*2,где *,*

, где

(напоминаю: , *m,j*∈

Если известны (или оценены) априорные вероятности:

, ,

то для классификации вектора используют следующее правило:

(2)

**Очевидно, правило (2) совпадает с правилом (1), когда .**

1. ***Численные эксперименты***
   1. Моделируете ОВ1, т.е. выборку заданного объема *n*1 из нормального **трехмерного** распределения, т.е. моделируете последовательность векторов

;

моделируете ОВ2, т.е. выборку заданного объема *n*2 из нормального **трехмерного** распределения, т.е. моделируете последовательность векторов

моделируете тестовую выборку, или последовательность, (ТП) заданного объема *n*, представленную смесью (других) объектов из **двух** известных нормальных **трехмерных** распределений; в этом случаеопределяются как доля объектов

∈, ∈ **– заданные вами значения параметров распределений, *р*=3.**

* 1. Строите дискриминантную функцию ( ДФ, см. (\*)), вычисляете константу *с* и расстояние Махаланобиса (все расчетные формулы – в п.1);
  2. Используя ДФ, построенную по обучающим выборкам ОВ1 и ОВ2, классифицируете векторы тестовой последовательности, ТП. Результаты классификации представляете в виде *Четырехпольной таблицы сопряженности (матрицы соответствий)*

|  |  |  |
| --- | --- | --- |
| Pred→  True ↓ | 1 | 2 |
| 1 |  | 2 |
| 2 |  | 2 |

* 1. Для сравнения с помощью этой же ДФ классифицируете векторы из ОВ1 и ОВ2; результаты представляете в подобной же таблице; проводите сравнительный анализ результатов.
  2. Вычислите вероятность ошибочной классификации по формулам

,

2.6 Анализируете результаты, полученные при выполнении пунктов 2.3, 2.4, 2.5.

**Замечание.** Численные эксперименты проводятся для различных модельных данных –«хорошо» разделенных, «плохо» разделенных и т.д.

Диапазон значений *n*2 , *n*1 , *n*: 100÷1000.

определяются как доля объектов ОВ1 и ОВ2, соответственно.

Любые доп.исследования приветствуются!!!

1. ***параметры распределения неизвестны***

В репозитории находите данные о кредитовании физических лиц

(<https://archive.ics.uci.edu/ml/machine-learning-databases/statlog/german/>);

В файле german.doc найдете описание данных, сами данные – в файле german.data-numeric.

Надеюсь, с извлечением данных проблем не возникнет, отлично разберетесь без меня.

Данные – многомерные, каждый объект характеризуется 24 (23?) признаками и содержит 1000 записей (векторов).

Вы **сами формируете** ОВ1, ОВ2 и ТП.

Затем проводите вычисления по этим данным, используя уже описанную схему.

Отличие заключается в том, что вместо значений параметров распределения придется пользоваться их оценками, исходя из предположения, что обе популяции – нормальные:

Также заменяете вектор параметров ДФ, оценкой ***a*** (это уже **a**, не *alpha*):

Эти оценки (***a***) используюете для получения значений ДФ для каждого вектора из ОВ1 и из ОВ2, после чего строите оценки :

И далее:

Оценки расстояния Махаланобиса (смещенная и несмещенная):

## Оценивание вероятности ошибочной классификации:

= Ф, = Ф

Таблицы сопряженности строятся так же, как и в случае известных параметров.

Далее – выводы по пункту ***b)*** и затем – общие выводы.

Удачи!

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
|  |  |  |  |  |