**ЛАБОРАТОРНАЯ РАБОТА №2**

**Кластерный анализ при распознавании образов**

**Цель работы:** Ознакомиться с наглядной интерпретацией многомер­ных образов в виде кластеров на плоскости по произвольно выбранным парам признаков. Освоить статистическое описание образов в много­мерном пространстве признаков и процедуры оценки информативности признаков для их минимизации. Исследовать особенности распознава­ния образов по расстоянию Махаланобиса и разделением пространства признаков плоскостями.

***Теоретические пояснения***

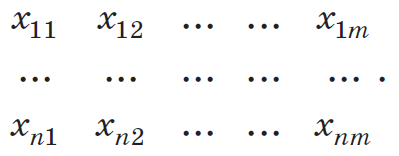
Любой физический объект обладает набором некоторых свойств, которые, собственно, и позволяют отличать один объект от другого. Совокупность свойств, описывающих конкретный объект, называется образом данного объекта. Под классом объектов понимается некоторая совокупность образов, называемых элементами класса, обладающая рядом близких свойств. Измеряемые или вычисляемые свойства объек­тов, позволяющие отличить классы друг от друга, называются призна­ками. В пределе каждый класс может состоять только из одного эле­мента, как, например, при опознавании человека. С другой стороны, вся совокупность образов может быть разделена всего на два класса, например «свой», «чужой».

Пусть число признаков равно *п,* тогда каждый образ можно предста­вить в виде некоторого набора значений признаков или вектора призна­ков видах x= *(х1,*..., *хn),* где хi - значение i-гo признака данного образа. В многомерном пространстве, осями координат которого являются от­дельные признаки, каждый образ определяется точкой, причем рассто­яние от этой точки до начала координат и расстояние между точками определяется выбранной метрикой.

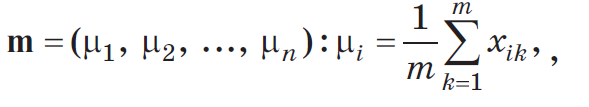
Каждый класс может характеризоваться некоторой областью в *п-*мерном пространстве признаков. Эта область определяется степенью отличия образов, относящихся к данному классу, и может задаваться границами, например плоскостями. Область, в которую с наибольшей вероятностью попадают векторы образов данного класса, обычно назы­вают кластером, а процедуру разбиения пространства признаков на об­ласти, соответствующие разным классам, - кластеризацией совокуп­ности образов.

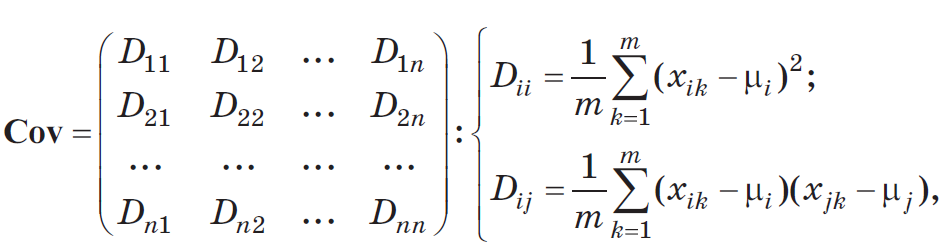
Если образы, относящиеся к одному классу, имеют различные зна­чения признаков, то для описания класса можно использовать статистические характеристики: вектор средних значении или вектор мате­матических ожиданий по всем признакам и вектор среднеквадратич­ных отклонений или дисперсий признаков. Для многомерных векторов принято использовать матрицу ковариации, на диагонали которой на­ходятся дисперсии соответствующих признаков.

Статистические характеристики признаков класса вычисляются следующим образом. Пусть набор n-мерных векторов, полученных в результате предъявления распознающей системе *т* объектов, относя­щихся к одному классу, задан в виде таблицы:



Тогда

(1)

 (2)

где *i, j*=1, ..., *п —* индексы номеров компонент вектора признаков; *т -* число образов, составляющих данный класс; *xik -* значение *i*-гo призна­ка *k*-го образа; - математическое ожидание *i*-й компоненты вектора признаков; *Dii -* дисперсия *i*-го признака; *Dij -* коэффициент ковариации *i*-гo и *j*-го признаков; **m** - вектор математических ожиданий; **Cov** -ковариационная матрица. Из выражения (10) видно, что ковариацион­ная матрица симметрична относительно главной диагонали и, следова­тельно, необходимо вычислять только половину ее элементов. Ковариация характеризует степень линейной зависимости случайных вели­чин. Если ковариация равна нулю, то величины называются некоррелированными.

Таким образом, процесс распознавания включает в себя этап обуче­ния, т. е. определение некоторых характеристик по обучающей репре­зентативной выборке образов, отнесенных к известным классам, и соб­ственно распознавание на основе информации, полученной при обуче­нии.

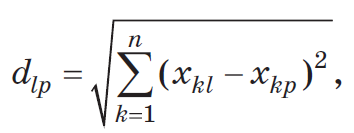
Под распознаванием образа понимается отнесение его к тому классу, расстояние до которого от данного образа в пространстве признаков минимально. Для определения расстояния между точками в метричес­ком пространстве необходимо ввести понятие метрики, т. е. определить процедуру измерения расстояния *dlp,* между точками *l* и *р* в этом про­странстве так, чтобы выполнялись следующие аксиомы:

- симметричность расстояния *(d­lp = dpl* );

- правило треугольника *(dlp* + *dhp >dlp* );

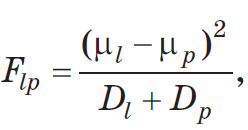
- положительность расстояния (*d*lp >= 0, причем *dlp =* 0, только если *1=р).*

Самой употребительной метрикой является евклидова метрика, в которой расстояние определяется следующим образом:



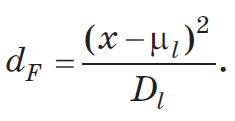
где *хkl, хkp - k-e* координаты точек *l и р* соответственно.

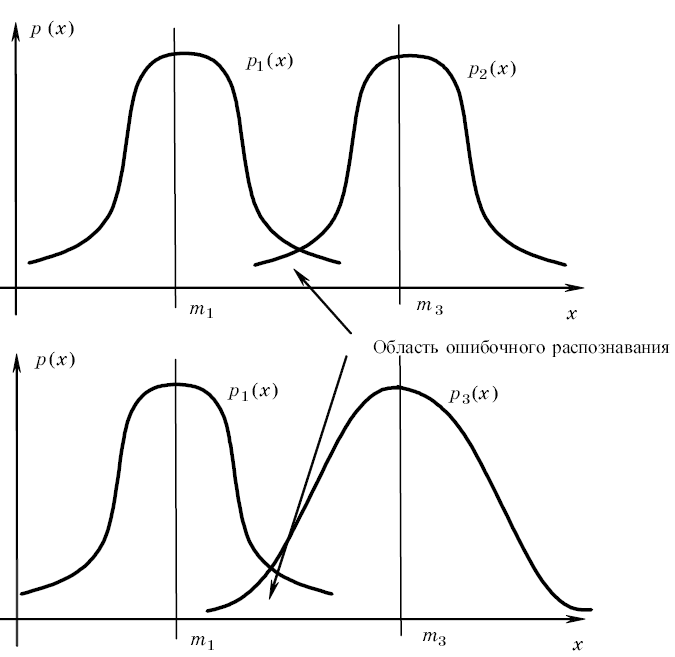
Для одномерных векторов *(k* = 1), т. е. в случае распознавания по одному признаку *dlp = xl – хp* . Расстояние между классами по одному признаку, учитывающее разброс значений признака для образов, отно­сящихся к одному классу, можно вычислить по критерию Фишера



где и D вычисляются по формулам (1), (2). Из приведенного на рис. 1 сравнения плотностей распределения вероятности значений признака для двух классов при одинаковых значениях математического ожида­ния и разных дисперсиях видно, что значение критерия Фишера и веро­ятность правильного распознавания для пары классов 1-2 больше, чем для пары 1-3. Таким образом, с увеличением критерия Фишера растет вероятность правильного распознавания по данному признаку, т. е. критерий Фишера позволяет оценить информативность признака.

Расстояние между распознаваемым образом и классом *l* при одном признаке, учитывающее статистические характеристики признака, можно найти по расстоянию Фишера





*Рис. 1. Плотность распределения вероятности значения признака*

Обобщением расстояния Фишера на многомерный случай является расстояние Махаланобиса



- об­ратная ковариационная матрица для класса *l.*

Методы распознавания можно разделить на две группы. К первой группе относятся методы, основанные на последовательном вычисле­нии расстояний между распознаваемым образом и центрами кластеров в пространстве признаков с той или иной метрикой, например расстоя­ния Махаланобиса. Образ считается принадлежащим классу, расстоя­ние до которого минимально.

Другая группа методов предполагает предварительное разбиение признакового пространства плоскостями или поверхностями более вы­сокого порядка на области, соответствующие известным классам, и определение принадлежности распознаваемого образа одной из облас­тей, например, для двух признаков производится разбиение плоскости набором прямых.

***Порядок выполнения работы***

В файле Train.csv задана обучающая выборка печатных цифр. Необходимо считать данные и на их основе определить статистические характеристики каждой цифры.

Считать тестовые данные из файла test.csv и используя описанный выше подход распознавать символы. Распознавание необходимо провести

1. на основе одного признака (например, площади), используя критерий Фишера;
2. на основе нескольких признаков например количество светлых символов в строках используя расстояние Махалонобиса.