Лабораторная работа №3

РАЗДЕЛЕНИЕ ОБЪЕКТОВ НА ДВА КЛАССА ПРИ ВЕРОЯТНОСТНОМ ПОДХОДЕ

Цель работы: изучить особенности классификации объектов при вероятностном подходе и научиться находить ошибку классификации.

Порядок выполнения работы

1. Изучение теоретической части лабораторной работы.
2. Выполнение классификации случайной величины и определение ошибки классификации.
3. Защита лабораторной работы.

Исходные данные:
1. Две случайные величины, распределенные по закону Гаусса.
2. Априорные вероятности отнесения каждой из случайных величин к первому из двух классов, в зависимости от того, для какого из них определяется ошибка классификации.

Выходные данные: вероятность ложной тревоги, вероятность пропуска обнаружения ошибки, вероятность суммарной ошибки классификации. Результаты работы программы должны представляться в графическом виде.

Примечание. Результат работы представить графически.

На основе апостериорных вероятностей можно разработать метод автоматической классификации. Примером апостериорной плотности вероятности является случай одномерного гауссового распределения, выражаемого формулой (1).

$$p(x/j)=\frac{1}{\sigma_j\sqrt{2\pi}}\exp[-\frac{1}{2}(\frac{x-\mu_j}{\sigma_j})^2]. \qquad (1)$$

Плотность распределения является функцией двух параметров: μ_j – математическое ожидание и σ_j – среднеквадратичное отклонение. Эти параметры могут быть вычислены по N опытам, в каждом из которых измеряется величина x_k $(k=1, 2,...N)$, а затем вычисляются $\hat{\mu}_j=\frac{1}{N}\sum_{k=1}^{N}x_k$; $\hat{\sigma}^2{}_j=\frac{1}{N}\sum_{k=1}^{N}(x_k-\hat{\mu}_j)^2$.

Пусть задано сепарабельное пространство признаков, которое по определению может быть разделено на классы. X – вектор, представляющий k-й класс, сепарабельного пространства. Априорная вероятность того, что X относится к классу с номером k, есть $P(X_k)$. Она считается заданной самой постановкой задачи.

Задача заключается в том, чтобы отнести неизвестный предъявляемый объект X к одному из известных классов C_k с минимальной ошибкой. Для этого выполняют n измерений в соответствии с признаками, выбранными надлежащим образом. В результате получают вектор измерений X_m, для которого можно найти условную вероятность или ее плотность: $p(X_m / C_k)$.

Решение об отнесении неизвестного объекта к классу с номером k можно считать оправданным, если для любого j выполняется условие

$$p(C_k / \vec{X}_m) \geq p(C_j / \vec{X}_m) \ \forall j.$$

Эти вероятности могут быть вычислены согласно теореме Бейеса по тем условным вероятностям $p(\vec{X}_m / C_k)$, которые получаются непосредственно в процессе измерений:

$$P(C_k / \vec{X}_m) = \frac{P(C_k) p(\vec{X}_m / C_k)}{p(X_m)}, \quad P(C_j / \vec{X}_m) = \frac{P(C_j) p(\vec{X}_m / C_j)}{p(X_m)}.$$

Откуда следует решающее правило:

$$P(C_k) p(\vec{X}_m / C_k) \geq P(C_j) p(\vec{X}_m / C_j).$$

Рассмотрим случай, когда весь набор возможных решений сводится к двум, т. е. предъявленный объект может быть отнесен к одному из двух имеющихся классов. На рис. 1 показаны плотности распределения случайной величины X_m в случае ее отнесения к классам C_1 и C_2. $P(C_1)$ – это вероятность отнесения X_m к классу C_1, а $P(C_2)$ – вероятность отнесения случайной величины к классу C_2. Рассмотрим вероятности ошибок, которые могут возникать при такой процедуре. Очевидно, что на прямой AB неравенство Бейеса выполняется, и можно заключить, что X_m принадлежит классу C_1.

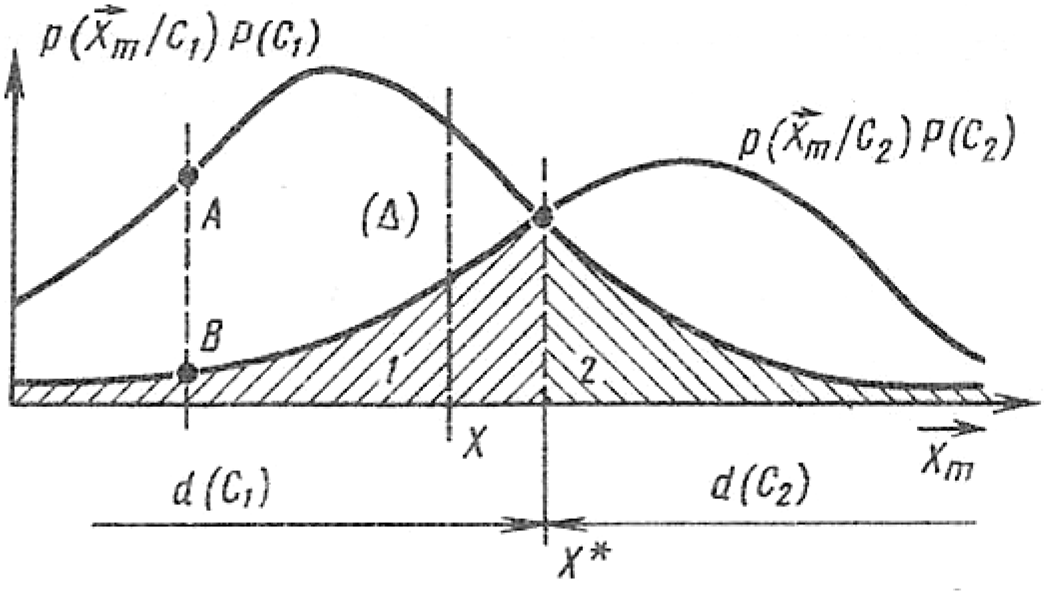

Рис. 1. Плотности распределения случайной величины

Рассмотрим линию раздела, обозначенную Δ. Любая точка, для которой $X_m < X$, считается принадлежащей классу C_1, в то время как все точки, для которых $X_m > X$, относятся к классу C_2. Однако вероятность того, что в первом случае точка может принадлежать классу C_2, отлична от нуля (область 1), так же как и то, что во втором случае точка X принадлежит классу C_1 (область 2). Для класса C_1 зона 1 является зоной ложной тревоги, а зона 2 является зоной пропуска обнаружения. Они определяются соответственно выражениями:

$$P_{л.т} = \int_{-\infty}^{x} P(C_2)\, p(\vec{X}_m / C_2) d\vec{X}_m; \qquad P_{п.о} = \int_{x}^{\infty} P(C_1)\, p(\vec{X}_m / C_1) d\vec{X}_m.$$

Суммарная ошибка классификации представляется суммой этих двух вероятностей. Если перемещать линию Δ, разделяющую два решения, вдоль оси *X*, то она должна достичь точки X^*, в которой имеет место равенство $P(C_1)\, p(\vec{X}_m / C_1) = P(C_2)\, p(\vec{X}_m / C_2)$, показывающее, что при бинарных ценах правило максимума правдоподобия обеспечивает оптимальную классификация по отношению к возможности ошибочного решения.