Лабораторная работа № 12

Дискриминантный анализ. Метод эталонов. Непараметрический дискриминантный анализ.

Для физико-статистического прогноза метеорологических величин и явлений погоды используются методы, как регрессионного анализа, так и дискриминантного анализа. Остановимся подробнее на последнем методе. При разработке прогноза указанным методом предварительно производится статистическая обработка архивного материала.

В теории статистических прогнозов характеристики состояния атмосферы, $x_1, x_2...$, используемые при составлении прогноза, принято называть прогностическими признаками, или *предикторами*, а прогностические характеристики $y_1, y_2...$- *предиктантами.* И предиктанты и предикторы могут быть как количественными (температура воздуха, влажность, скорость и направление ветра и др.), так и качественными (форма облаков, вид осадков и др.) метеорологическими характеристиками. В перечень предикторов обычно включаются характеристики исходного состояния атмосферы – так называемые первичные прогностические признаки.

Разработка статистических методов прогноза, как правило, начинается с составления предварительного перечня предикторов $x_1, x_2,..., x_n$ и подготовки необходимого архивного материала. Длина архивной выборки N должна быть достаточно большой для обеспеченности оценок характеристик. Для прогноза явления погоды (гроза, шквал, град и т.д.) используются методы дискриминантного анализа. При этом явление погоды – предиктант Y - принимает только два значения – «есть явление» - y_1 и «нет явления» - y_2. Иными словами, предиктант Y представляет собой вектор с двумя координатами y_1, y_2. Предикторы рассматриваются, как многомерные вектора с N координатами, или

$$Y = \begin{Bmatrix} y_1 \\ y_2 \end{Bmatrix},$$

$$x_1 = \begin{Bmatrix} x_{11} \\ x_{12} \\ \\ x_{1N} \end{Bmatrix}, x_2 = \begin{Bmatrix} x_{21} \\ x_{22} \\ \\ x_{2N} \end{Bmatrix} ...., x_n = \begin{Bmatrix} x_{n1} \\ x_{n2} \\ \\ x_{nN} \end{Bmatrix}.$$

Следующим этапом разработки статистических методов прогноза является установление на архивном материале связи между предикторами и предиктантом. В связи с неизбежной ограниченностью объема архивной выборки может оказаться, что из-за низкой статистической значимости выборочных оценок параметров связей и критериев успешности использование всех предикторов является нецелесообразным. Поэтому в процессе разработки статистических методов применяются методы оценки информативности предикторов.

В методе эталонов оценка информативности предикторов осуществляется последовательно для каждого x_i путем построения диаграммы рассеивания. Покажем оценку информативности предикторов на примере.

Пусть дана выборка для трех предикторов x_1, x_2, x_3. Для каждого из предикторов построим диаграмму рассеяния следующим образом. На одной прямой отметим в масштабе все значения i – того предиктора, на двух параллельных – звездочками и ноликами - отсутствие и наличие явления погоды при определенном значении предиктора. При этом вся архивная выборка разделится на два класса A_1 и A_2 . Чем меньше область пересечения классов – тем информативней предиктор. Для двух предикторов наиболее успешных – строится диаграмма рассеяния на плоскости. Эталоном каждого класса считается среднее значение в каждом классе. Решающим правилом для альтернативного прогноза служит следующее условие

если расстояние от точки до эталона первого класса больше, чем расстояние от эталона второго класса, то будет с большей вероятностью прогнозироваться второй класс.

Эталоном первого класса A_1 служит вектор $M_1 = \left\{ \bar{x}_1(A_1), \bar{x}_2(A_1) \right\}$, а эталоном класса A_2 – вектор $M_2 = \left\{ \bar{x}_1(A_2), \bar{x}_2(A_2) \right\}$. В этом случае решающее правило запишется в векторном виде следующим образом

$\tilde{\rho} - M_1 > \tilde{\rho} - M_2$, то прогнозируется A_2. (12.1)

Уравнение дискриминантный прямой в координатах примет вид

$$(\tilde{x}_1 - \bar{x}_1(A_1))^2 + (\tilde{x}_2 - \bar{x}_2(A_1)^2 = (\tilde{x}_1 - \bar{x}_1(A_2))^2 + (\tilde{x}_2 - \bar{x}_2(A_2))^2 \qquad (12.2)$$

Эту прямую необходимо нанести на диаграмму рассеяния. Дискриминантная прямая разделит классы в соответствии с принятым эталоном.

Успешность прогноза можно проверить на независимом материале по параметру Обухова.

$$P\% = \frac{n_{11} + n_{22}}{N} \qquad (12.3)$$, где

n_{11} – количество значений первого класса, попавших в первый класса, n_{22} - количество значений второго класса, попавших во второй класс, а N – длина обучающей выборки. При использовании параметра (критерия) Обухова не учитывались не оправдавшиеся прогнозы. Для прогноза опасных явлений этот критерий не всегда подходит, так как один не оправдавшийся прогноз может причинить большой ущерб потребителю.