1 Задачи снижения размерности

0.1 Постановка задачи

Как неоднократно отмечалось ранее мы работаем с набором измерений характеристик (признаков) некоторых отобранных объектов из генеральной совокупности. Часто в реальных задачах для описания объекта используется достаточно много таких характеристик и работать с таким набором данных очень непросто. В этом случае пытаются представить первоначальный набор измерений в виде некоторых функций, чаще всего линейных, от других вспомогательных величин гораздо меньшей размерности.

Следуя монографии [**?**], перечислим несколько причин зачем это нужно и как это можно сделать.

1) Во первых хотелось бы **наглядно представить** имеющийся набор данных, чтобы на интуитивном уровне понять специфику имеющейся задачи. Для этого размерность наблюдений должна быть не выше трех.

2) Во-вторых, мы стремимся построить **максимально простую модель**, в рамках которой упрощаются расчеты и интерпретация полученных результатом.

3) Наконец важную роль играет возможность **сжатия объема данных** без существенной потери важной информации. Это становится особенно существенным при записи и хранении больших баз данных.

При формировании новых признаков для описания объектов стараются отобрать наиболее информативные переменные. При этом принимают во внимание следующие обстоятельства:

1) некоторые характеристики являются **сильно коррелированными**, что приводит к дублированию информации;

2) часть характеристик может быть **мало информативной**, т.к. они мало меняются при переходе от одного объекта к другому;

3) некоторые характеристики можно **агрегировать** и переходить к некоторым новым взвешенным характеристикам.

Формально учет этих обстоятельств производится следующим об-

разом. Пусть мы имеем некоторый исходный набор измерений $X = (X_1, \ldots, X_p)$. По определенным правилам строится новый набор признаков $Z(X) = (Z_1(X), \ldots, Z_q(X))$, где $q \ll p$. Задана мера информативности $I_q(Z(X))$ новой системы признаков с учетом старой. Требуется так подобрать новую систему признаков, чтобы максимизировать этот показатель.

Проиллюстрируем эту идею на нескольких типичных примерах.

Метод главных компонент

Образуем несколько новых переменных:

$$Z_k(X) = c_{k1}(X_1 - E(X_1)) + \ldots + c_{kp}(X_p - E(X_p)),$$

$k = \overline{1, q}$, причем

$$\sum_{j=1}^{p} c_{kj}^2 = 1,\ k = \overline{1, q},$$

$$\sum_{j=1}^{p} c_{kj} \cdot c_{lj} = 0,\ k, l = \overline{1, q}, k \neq l.$$

В качестве меры информативности берут величину

$$I_k(Z(X)) = \frac{D(Z_1) + \ldots + D(Z_q)}{D(X_1) + \ldots + D(X_p)}.$$

Факторный анализ

В модели факторного анализа набора исходных наблюдений представляется в виде

$$X_j - E(X_j) = \sum_{k=1}^{q} {}_{jk}Y_k + u_j = \hat{X}_j + u_j,\ j = \overline{1, p},$$

где c_{jk} есть нагрузка общего фактора Y_k на исходный показатель X_j, u_j – остаточная (специфическая) компонента, причем выполнены следующие условия (ограничения): $E(Y_k) = 0$, $E(u_j) = 0$, $D(Y_k) = 1$ и с.в. $Y_1, \ldots, Y_q$, $u_1, \ldots, u_p$ – попарно некоррелированы.

В качестве меры информативности берут величину $I_q(Z(X)) = 1 - ||R_X - R_{\hat{X}}||$, где R_X и $R_{\hat{X}}$ есть корреляционные матрицы случайных векторов X и $\hat{X}$.

Метод экстремальной группировки признаков

Разобъем набор исходных признаков $X_1, \ldots, X_p$ на q непересекающихся групп $S_1, \ldots, S_q$. Признаки внутри группы должны быть сильно коррелированы, а признаки из разных групп должны быть слабо коррелированы. Признаки из одной группы S_k заменяются на некоторый единый общий признак Z_k. В качестве допустимых рассматривают линейные преобразования исходных признаков, причем $D(Z_k) = 1$. Мера информативности задается по правилу:

$$I_q(Z(X); S) = \sum_{X_j \in S_1} r^2(X_j, Z_1) + \ldots + \sum_{X_j \in S_q} r^2(X_j, Z_q).$$

0.2 Метод главных компонент

Одним из наиболее популярных на практике является **метод главных компонент**. Поэтому мы рассмотрим его более подробно.

Пусть мы имеем случайный вектор $X = (X_1, \ldots, X_p)^T$. Предположим, что он является центрированным, т.е. его математические ожидания равны нулю.

Определение 1 *Первой главной компонентой для вектора X называется случайная величина Z вида*

$$Z = c_{11}X_1 + \ldots + c_{1p}X_p,$$

где $c_{11}^2 + \ldots + c_{1p}^2 = 1$, для которой дисперсия будет максимальной.

Пусть $\Sigma = (\sigma_{ij})$ есть матрица ковариаций случайного вектора X. Нам нужно решить некоторую задачу на условный экстремум. Записывая необходимое условие экстремума, после несложных преобразований получаем

$$(\Sigma - \lambda I)c_1 = 0,$$

где $c_1 = (c_{11}, \ldots, c_{1p})$. Последнее уравнение имеет нетривиальное решение, если число λ удовлетворяет уравнению

$$|\Sigma - \lambda I| = 0,$$

т.е. является собственным значением матрицы Σ. Т.к. матрица Σ является положительно определенной, то все ее собственные числа будут вещественными и положительными.

Далее нетрудно показать, что $D(Z) = \lambda$. Поэтому для поиска первой главной компоненты найти нужно наибольшее собственное значение и соответствующий ему собственный вектор. Вектор c_1 есть нормированный собственный вектор, соответствующий найденному собственному значению.

Пусть мы уже нашли $q-1$ главных компонент, $q \geq 2$.

Определение 2 *q-ой главной компонентой для вектора X называется случайная величина Z вида*

$$Z = c_{q1}X_1 + \ldots + c_{qp}X_p,$$

где $c_{q1}^2 + \ldots + c_{qp}^2 = 1$, $c_{k1}c_{q1} + \ldots + c_{kp}c_{qp} = 0$, $1 \leq k \leq q-1$, для которой дисперсия будет максимальной.

Аналогичный анализ показывает, что нужно найти q-ое сверху по величине собственное значение и соответствующий ему нормированный собственный вектор. Это даст нам нужный набор коэффициентов $c_q = (c_{q1}, \ldots, c_{qp})$.