Министерство образования и науки Российской Федерации

Новосибирский государственный технический университет

Кафедра Программных систем и баз данных

Лабораторная работа №5

по курсу «Статистические методы анализа данных»

Факультет: ПМИ

Группа: ПМ -21, ПМ-23

Студенты: Заборцев В.А, Яковлева М.А.

Преподаватели: Попов А.А., Гультяева Т.А.

Вариант:3

Новосибирск

2015

**Задание:**

1. Сгенерировать экспериментальные данные, в которых в явном виде присутствует эффект мультиколлинеарности. Регрессия на 8 факторах. Эффект мультиколлинеарности создают 4 фактора. Имеется разброс в масштабах факторов.
2. Рассчитать ряд показателей, характеризующих эффект мультиколлинеарности, определить факторы, ответственные за неё.
3. Построить ридж-оценки параметров при различном значении параметра регуляризации, выбрать оптимальное значение. Построить графики изменения квадрата эвклидовой норы оценок параметров и остаточной суммы квадратов от параметра регуляризации.
4. Провести оценивание модели регрессии по методу главных компонентов. Перейти к описанию в исходном пространстве факторов. Сравнить решение с ридж-оцениванием по смешению оценок и точности предсказания отклика.

**Ход работы:**

1. **Генерация экспериментальных данных, с эффектом мультиколлинеарности.**

θ = (1, 1, 1, 1, 1, 1, 1)

f(x)=(1, x1, x2, x3, x4, x5, x6, x7, x5 +10x6 -4x7 +)

Число экспериментов n=50

Matrix X = new Matrix(n, m);

double x1 = 1.0, x2 = 0.5, x3 = 0.7, x4 = 0.4, x5 = 0.3, x6 = 1.0, x7 = 2.0;

double avgU = 0.0;

for (int i = 0; i < n; i++) {

double[] point = new double[]{

math.Distribution.Flat(-x1,x1\*2.0),math.Distribution.Flat(-x2,x2\*2.0),

math.Distribution.Flat(-x3,x3\*2.0),math.Distribution.Flat(-x4,x4\*2.0),

math.Distribution.Flat(-x5,x5\*2.0),math.Distribution.Flat(-x6,x6\*2.0),

math.Distribution.Flat(-x7,x7\*2.0),

};

double[] f = calcF(point);

X[i] = new Matrix(m, f);

u.Add(calcU(f, Tetta));

avgU += u.Last();

}

avgU /= countExperiments;

double sigma2 = avgU \* noize;

for (int i = 0; i < countExperiments; i++) {

e.Add(math.Distribution.Normal(0.0, sigma2));

y.Add(u[i] + e.Last());

}

1. **Расчет показателей, характеризующих эффект мультиколлинеарности**

*а) Определитель информационной матрицы* :

208111,657406803

*Определитель информационной матрицы, нормированной на след:*

2,10594788500701E-25

Matrix XT = X.Transpose();

Matrix XTX = XT \* X;

double detXTX = XTX.Determinant();

Matrix XTX\_Trace = XTX / XTX.Trace();

double detXTX\_Trace = XTX\_Trace.Determinant();

*б) Минимальное собственное число матрицы*:

3,6371927789086E-05

double minL = XTX.MinEiganValue();

Минимальное собственное значение близко к нулю, что говорит о плохой обусловленности матрицы. Чем меньше минимальное собственное значение , тем сильнее мультиколлинеарность.

*в) Мера обусловленности матрицы по Нейману-Голдстейну.*

55695589,8443597

double maxL = XTX.MaxEiganValue();

double minL = XTX.MinEiganValue();

double lmax\_lmin = maxL / minL;

*г) Максимальная парная сопряженность:*

Матрица сопряженности имеет вид:

, где 

Показатель мультиколлинеарности: 0,753977805095502

Matrix R = new Matrix(Tetta.Count(), Tetta.Count());

double r\_max = 0.0;

for (int j = 0; j < Tetta.Count(); j++) {

for (int i = 0; i < Tetta.Count(); i++) {

double sum\_up = 0.0, sum\_d1 = 0.0, sum\_d2 = 0.0;

for (int k = 0; k < countExperiments; k++) {

sum\_up += X[k, i] \* X[k, j];

sum\_d1 += X[k, i] \* X[k, i];

sum\_d2 += X[k, j] \* X[k, j];

}

double rij = sum\_up / (Math.Sqrt(sum\_d1) \* Math.Sqrt(sum\_d2));

if (i != j && Math.Abs(rij) > r\_max)

r\_max = Math.Abs(rij);

R[i, j] = rij;

}

R[j, j] = 1.0;

}

*д) В качестве меры мультиколлинеарности* можно взять , где  вычисляется по формуле:

, элемент матрицы, обратной к сопряженной R.

0,999998926110047

Matrix RInverse = R.Inverse();

r\_max = 0.0;

for (int i = 0; i < Tetta.Count(); i++) {

double ri = Math.Sqrt(1.0 - 1.0 / RInverse[i,i]);

if (ri > r\_max)

r\_max = ri;

}.

Вывод: Эффект мультиколлинеарности присутствует.

1. **Получим ридж-оценки неизвестных параметров модели при различных значениях параметров регуляризации:**

Часто матрицу Λ задают диагональной в виде: .

Будем рассматривать . Параметр выбирался исходя из стабилизации .







При = 4.2105263e-3 ****

1. **Метод главных компонент.**

Перейдем к центрированным переменным:  Выразим главные компоненты в матричном виде через -матрица из всех собственных векторов матрицы . Исключим из матриц  и столбцы, соответствующие собственным значениям с незначительными вкладами . Обозначим полученные матрицы как и соответственно. И посчитаем вектор , с помощью которого найдем .



Вектор собственных значений:

Незначительный вклад даёт , эту строчку и будем исключать.



Вектор оценок:





**Код программы:**