Pattern classification using polynomial and linear regression

Ciuiu, Daniel

Technical University of Civil Engineering, Bucharest, Romania,
Romanian Institute for Economic Forecasting

January 2008

PATTERN CLASSIFICATION USING POLYNOMIAL AND LINEAR REGRESSION

Daniel CIUIU
Technical University of Civil Engineering, Bd. Lacul Tei, no. 124, Bucharest, ROMANIA
E-mail: dciuiu@yahoo.com

Abstract: In this paper we will classify patterns using an algorithm analogous to the *k*-means algorithm and the regression polynomial of the degree *k* (for instance, if *k*=1 we obtain the regression line, and if *k*=2 we obtain the regression parable), and the regression hyper-plane. We will also present a financial application in which we apply these regressions if the points represent the interests for accounts with different terms.
Mathematics Subject Classification (2000): 62J05, 62J02, 68T10
Keywords: Regression, pattern classification, k-means

1. Introduction

First we have two samples $X_1,\ldots,X_n$ and $Y_1,\ldots,Y_n$. The regression polynomial of the degree *k* is (see [4,1,3])

$$Y = \sum_{j=0}^{k} a_j \cdot X^j. \tag{1}$$

$$\sum_{j=1}^{n}\left(Y_j - \sum_{l=1}^{k} a_l \cdot X_j^l \right)^2 \text{ is minimum.} \tag{2}$$

For computing $a_j,\ j=\overline{0,k}$ we will solve the system

$$\sum_{j=0}^{k} a_j \cdot \overline{X^{p+j}} = \overline{X^p \cdot Y},\ p=\overline{0,k}\ , \text{ where } \overline{X^0}=1. \tag{3}$$

Consider now *n* points in $\mathbf{R}^{k+1}$ $X^{(1)},\ldots,$ $X^{(n)}$, where $X^{(i)} = \left(X_1^{(i)}, X_2^{(i)},\ldots, X_{k+1}^{(i)}\right)$. The regression hyper-plane has the equation

$$H\ :\ X_{k+1} = \sum_{j=1}^{k} A_j \cdot X_j + A_0 \text{ such that} \tag{4}$$

$$\sum_{i=1}^{n}\left(X_{k+1}^{(i)} - \sum_{j=1}^{k} A_j \cdot X_j^{(i)} - A_0 \right)^2 \text{ is minimum.} \tag{5}$$

We have to solve the linear system

$$\sum_{j=0}^{k} \overline{X_i \cdot X_j} \cdot A_j = \overline{X_i \cdot X_{k+1}},\ i=\overline{0,k}\ , \tag{6}$$

where $\overline{X_0 \cdot X_i} = \overline{X_i}$ and $\overline{X_0^2}=1$.

For the *n* points of the plane or from $\mathbb{R}^p$ we can find the regression line, the regression polynomial or the regression hyper-plane. But in this case all the *n* points are considered in the same class. A modality to classify *n* points from $\mathbb{R}^p$ in *k* classes is to use the k-means algorithm (see [2]). First each class has only one point, which represents the class. The other points are introduced next into the class represented by the nearest point (the center of gravity of the points from the given class), and we compute the new center of gravity of this class. The next step is to check for each point if the distance to the center of gravity of its class is

minimum. Otherwise we move the point from the current class such that the distance becomes minimum. We compute the centers of gravity for the source class and destination class, and the algorithm stops when no point is moved from its class.

2. The k-means algorithm and the regression

In the k-means algorithm the classes are given by their gravity centers $Y_i,\ i=\overline{1,k}$. These points minimize the sum

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(X_{ij}-Y_i\right)^2, \tag{7}$$

where $X_{ij},\ j=\overline{1,n_i}$ are the n_i points from $\mathbb{R}^m$ that are classified into the class i by the k-means algorithm.

In the same manner we can classify patterns from $\mathbb{R}^2$ using the regression polynomial of the degree k or patterns from $\mathbb{R}^{k+1}$ using the regression hyper-plane. In these cases each class has at least k+1 points, initially exactly k+1 points.

The other points are classified first in the class with the less distance (for the given point $\left|Y-\sum_{j=0}^{k}A_j\cdot X^j\right|$ is minimum in the first case, and $\left|X_{k+1}-A_0-\sum_{j=1}^{k}A_j\cdot X_j\right|$ is minimum in the second case).

After the first classification, we take each point and if we have a distance less than those to the current class, we move the point to the new class. The algorithm stops when all the points are not moved.

When we add a new point to a class we compute again the regression polynomial respectively the regression hyper-plane for this class. If we move a point from a class to another one, we have to compute again the regression polynomial respectively the regression hyper-plane for both classes (those from we move and those in which we move the point).

3. A Financial application

For the following application X_1 is the annual interest for an account without term, X_2 is the annual interest for an account with the term one month, X_3 is the annual interest for an account with the term *3* months, X_4 is the annual interest for an account with the term *6* months, X_5 is the annual interest for an account with the term *9* months and X_6 is the annual interest for an account with the term one year. Consider *29* banks as follows.

Bank	X_1	X_2	X_3	X_4	X_5	X_6
ABN-Ambro Romania	*0.25%*	*3.5%*	*3.75%*	*3.75%*	*0*	*3.75%*
AlphaBank	*0.1%*	*6.25%*	*6.5%*	*7%*	*7%*	*7.25%*
Banc Post	*0*	*7.25%*	*7.25%*	*7.15%*	*0*	*7.15%*
Banca Comercială Carpatica	*1%*	*7.5%*	*7.55%*	*7.6%*	*7.75%*	*7.8%*
BCR	*0.25%*	*6%*	*6.25%*	*6.5%*	*6.75%*	*7.5%*
Banca Italo-Romena	*0*	*5.5%*	*5.75%*	*6%*	*6.15%*	*6.25%*
Banca Românească	*0.75%*	*7.3%*	*7.75%*	*8.05%*	*8.1%*	*8.1%*
Banca Transilvania	*0.25%*	*7.5%*	*7.5%*	*7.5%*	*7.75%*	*7.75%*
Bank Leumi Romania	*0.25%*	*7.5%*	*7.5%*	*7.75%*	*7.75%*	*8%*
Blom Bank Egypt	*0.1%*	*6%*	*6.5%*	*6.5%*	*6.75%*	*7%*
BRD-Groupe Société Générale	*0.25%*	*5.5%*	*5.6%*	*5.65%*	*5.65%*	*5.75%*
C.R. Firenze Romania	*0.1%*	*6.5%*	*6.75%*	*7%*	*7.25%*	*7.5%*

CEC	*0.25%*	*7%*	*7%*	*7.25%*	*0*	*7.25%*
Citibank Romania	*1%*	*4.28%*	*4.28%*	*4.28%*	*3.87%*	*3.46%*
Emporiki Bank	*0.5%*	*6.75%*	*7%*	*7.25%*	*7%*	*7%*
Finansbank	*0.1%*	*7.5%*	*8%*	*8%*	*8%*	*8.5%*
HVB-Țiriac Bank	*0.1%*	*6.4%*	*6.3%*	*6.2%*	*6.1%*	*6.1%*
ING Bank	*6.85%*	*5.5%*	*5.75%*	*6%*	*6.25%*	*6.5%*
Libra Bank	*0*	*8%*	*8.1%*	*7.6%*	*7.6%*	*8.5%*
Mind Bank	*0.25%*	*7%*	*7%*	*7.25%*	*7.5%*	*7.75%*
OTP Bank	*0.25%*	*6.25%*	*6.5%*	*7%*	*7%*	*7.25%*
Piraeus Bank	*0.5%*	*7%*	*7.1%*	*7.25%*	*7.1%*	*7.35%*
Pro Credit Bank	*7%*	*7.5%*	*7.65%*	*7.7%*	*0*	*7.85%*
Raiffeisen Bank	*0.25%*	*4%*	*4.25%*	*4.5%*	*4.6%*	*4.75%*
Romanian International Bank	*0.25%*	*6.5%*	*6.75%*	*7%*	*7.5%*	*7.75%*
Romexterra	*0.25%*	*7.5%*	*7.75%*	*7.75%*	*8.1%*	*8.1%*
San Paolo IMI Bank	*0.1%*	*6.5%*	*6.7%*	*6.8%*	*7%*	*7.2%*
Uni Credit Romania	*0.1%*	*5%*	*5%*	*5.25%*	*5.5%*	*5.5%*
Wolksbank	*0.1%*	*4.5%*	*4.75%*	*4.5%*	*3.5%*	*3.25%*

For the polynomial regression we consider $X=X_2$ and $Y=X_6$.

The regression line is d: $Y=-0.2685+1.12081X$ and the error is 9.01598.

If we consider *2* classes we obtain the regression lines d_1:$Y=-0.04291+1.14965X$ and d_2:$Y=-2.96728+1.44491X$ with the classes C_1={*ABN Ambro Romania, Alpha Bank, BCR, Banca Italo-Romena, Banca Românească, Blom Bank Egypt, BRD-Groupe Société Générale, C.R. Firenze Romania, Finansbank, ING Bank, Mind Bank, OTP Bank, Raiffeisen Bank, Romanian International Bank, San Paolo IMI Bank, Uni Credit Romania*} and C_2={*Banc Post, Banca Comercială Carpatica, Banca Transilvania, Bank Leumi Romania, CEC, Citibank Romania, Emporiki Bank, HVB-Țiriac Bank, Libra Bank, Piraeus Bank, Pro Credit Bank, Romexterra, Volksbank*}, and the error is 1.7142.

If we consider *5* classes we obtain the regression lines d_1:$Y=0.67734+1.04311X$, d_2:$Y=-11.7+2.6X$, d_3:$Y=-7.84474+2.55441X$, d_4:$Y=20.875-1.75X$ and d_5:$Y=0.04704+1.0419X$ with the classes C_1={*Alpha Bank, Blom Bank Egypt, C.R. Firenze Romania, Finansbank, ING Bank, Mind Bank, OTP Bank, Raiffeisen Bank, Romanian International Bank, San Paolo IMI Bank*}, C_2={*Bank Post, Banca Comercială Carpatica*}, C_3={*BCR, Banca Italo-Romena, Citibank Romania, Volksbank*}, C_4={*Banca Românească, Banca Transilvania*} and C_5={*ABN Ambro Romania, Bank Leumi Romania, BRD-Groupe Société Générale, CEC, Emporiki Bank, HVB-Țiriac Bank, Libra Bank, Piraeus Bank, Pro Credit Bank, Romexterra, Uni Credit Romania*}, and the error is 1.08048.

The regression parable is P:$Y=-4.21255+2.5319X-0.12041X^2$ and the error is 8.15035.

If we consider *2* classes we obtain the regression parables P_1:$Y=1.38578+0.6556X+0.02627X^2$ and P_2:$Y=-19.27664+7.34059X-0.4917X^2$ with the classes C_1={*ABN Ambro Romania, Banc Post, Banca Comercială Carpatica, Banca Transilvania, BRD-Groupe Société Générale, CEC, Emporiki Bank, HVB-Țiriac Bank, Libra Bank, Piraeus Bank, Pro Credit Bank, Raiffeisen Bank, San Paolo IMI Bank, Uni Credit Romania*} and C_2={*Alpha Bank, BCR, Banca Italo-Romena, Banca Românească, Bank Leumi Romania, Blom Bank Egypt, C.R. Firenze Romania, Citibank Romania, Finansbank, ING Bank, Mind Bank, OTP Bank, Romanian International Bank, Romexterra, Volksbank*}, and the error is 1.87754.

If we consider *5* classes we obtain the regression parables P_1:$Y=-8.47098+4.73265X-0.35554X^2$, P_2:$Y=-22.46613+8.63262X-0.61202X^2$, P_3:$Y=-33.70345+10.44282X-0.6459X^2$, P_4:$Y=-55.54367+19.06121X-1.43947X^2$ and P_5:$Y=15.60269-5.61478X+0.64371X^2$ with the classes C_1={*ABN Ambro Romania, Alpha Bank, Bank Post, CEC, OTP Bank, Raiffeisen Bank, San Paolo IMI Bank*}, C_2={*Banca Comercială Carpatica, BCR, Banca Italo-Romena, Banca Transilvania, Bank Leumi Romania, ING Bank, Pro Credit Bank,*

Romanian International Bank, Uni Credit Romania}, C_3={*Banca Românească, Finansbank, Libra Bank, Mind Bank, Romexterra*}, C_4={*Blom Bank Egypt, BRD-Groupe Société Générale, C.R. Firenze Romania, Piraeus Bank*} and C_5={*Citibank Romania, Emporiki Bank, HVB-Țiriac Bank, Volksbank*}, and the error is 0.26952.

The regression hyper-plane is H:X_6=$-0.96942+0.01774X_1-0.28119X_2+0.08868X_3+1.3183X_4+0.04368X_5$, and the error is 3.88028.

If we consider *2* classes we obtain the regression hyper-planes H_1:X_6=$-4.66687-0.22427X_1-8.95837X_2+9.5893X_3+1.02818X_4-0.09291X_5$ and H_2:X_6=$-0.2763+0.00181X_1-0.34901X_2+0.14269X_3+0.01683X_4+1.2312X_5$ with the classes C_1={*ABN Ambro Romania, Banc Post, BCR, CEC, Libra Bank, Piraeus Bank, Pro Credit Bank*} and C_2={*Alpha Bank, Banca Comercială Carpatica, Banca Italo-Romena, Banca Românească, Banca Transilvania, Bank Leumi Romania, Blom Bank Egypt, BRD-Groupe Société Générale, C.R. Firenze Romania, Citibank Romania, Emporiki Bank, Finansbank, HVB-Țiriac Bank, ING Bank, Mind Bank, OTP Bank, Raiffeisen Bank, Romanian International Bank, Romexterra, San Paolo IMI Bank, Uni Credit Romania, Volksbank*}, and the error is 0.38177.

If we consider *4* classes we obtain the regression hyper-planes H_1:X_6=$-9.5613+0.91968X_1-18.27318X_2+23.07002X_3-2.52669X_4-0.19166X_5$, H_2:X_6=$-0.59823+0.07785X_1-0.03927X_2+0.30234X_3-0.38595X_4+1.23798X_5$, H_3:X_6=$-2.68207+0.19553X_1-0.16804X_2+2.13261X_3-0.53362X_4-0.04692X_5$ and H_4:X_6=$5.31396-5.52515X_1+7.92617X_2-9.62765X_3+7.19706X_4-4.8613X_5$ with the classes C_1={*ABN Ambro Romania, Alpha Bank, Banc Post, Banca Comercială Carpatica, BCR, Banca Italo-Romena*}, C_2={*Bank Leumi Romania, Blom Bank Egypt, BRD-Groupe Société Générale, C.R. Firenze Romania, Romanian International Bank, San Paolo IMI Bank, Uni Credit Romania, Volksbank*}, C_3={*Banca Românească, Banca Transilvania, CEC, Citibank Romania, Emporiki Bank, Finansbank, HVB-Țiriac Bank, ING Bank, Romexterra*} and C_4={*Libra Bank, Mind Bank, OTP Bank, Piraeus Bank, Pro Credit Bank, Raiffeisen Bank*}, and the error is 0.00579.

4. Conclusions

The applied *k*-means algorithm finds the minimum of error because there exists a finite number of classifications, and when we move a point to another class we obtain a smaller error. The error is smaller even if we only move the point and we consider the same regression polynomial.

We can see that if we consider the same degree of the polynomial and we increase the number of classes the error decrease. We can explain this as follows. Suppose that at a given moment we have k optimal classes given by their regression polynomials of the degree d. From some classes with at least $d+2$ points we can move $d+1$ points to a new class given by the interpolation polynomial.

If we take the same number of classes and we increase the degree of the polynomial the error generaly decrease. This can be explain because we find the polynomials with the degree at most d (particularly if the dominant coefficient is zero, we obtain a polynomial of the degree at most $d-1$).

References

[1] Ciucu, G. and Craiu, V.: *Statistical Inference*, Didactic and Pedagogic Publishing House, Bucharest, 1974 (Romanian).

[2] Dumitrache, I., Constantin, N. and Drăgoicea, M.: *Neural Networks*, Matrix Rom, Bucharest, 1999 (Romanian).

[3] Petrehuş, V. and Popescu, A.: *Probabilities and Statistics*, UTCB Publishing House, Bucharest, 1997 (Romanian).

[4] Saporta, G.: *Probabilités, analyse des donées et statistique*, Editions Technip, Paris, 1990.