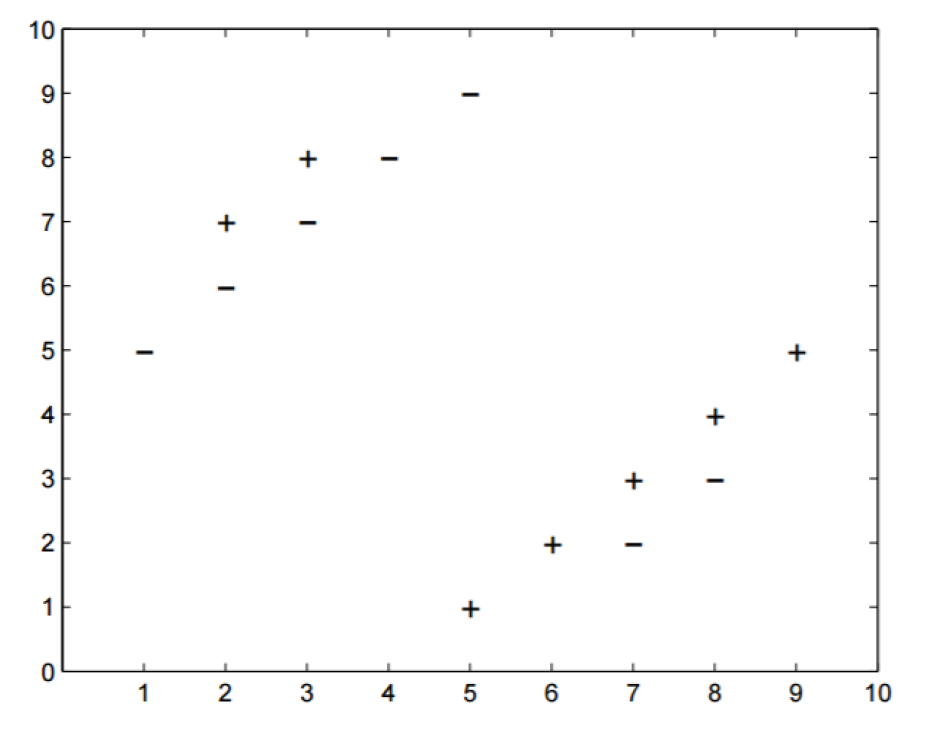
**BUSINESS DATA MINING**

**(IDS 572)**

**Solutions to Homework 8**

**Group Members**

* Amey Pophali (apopha2@uic.edu)
* Karthik Varanasi (vvaran3@uic.edu)
* Mrinal Dhawan ([mdhawa3@uic.edu](mailto:mdhawa3@uic.edu))
* **Solution 1**
* **Part 1**

As per the information provided in the question, a point can be its own neighbor. Therefore, the value of ‘k’ that minimizes the error on the training set is zero. If we take k=0, the error on the training set come out to be zero.

* **Part 2**

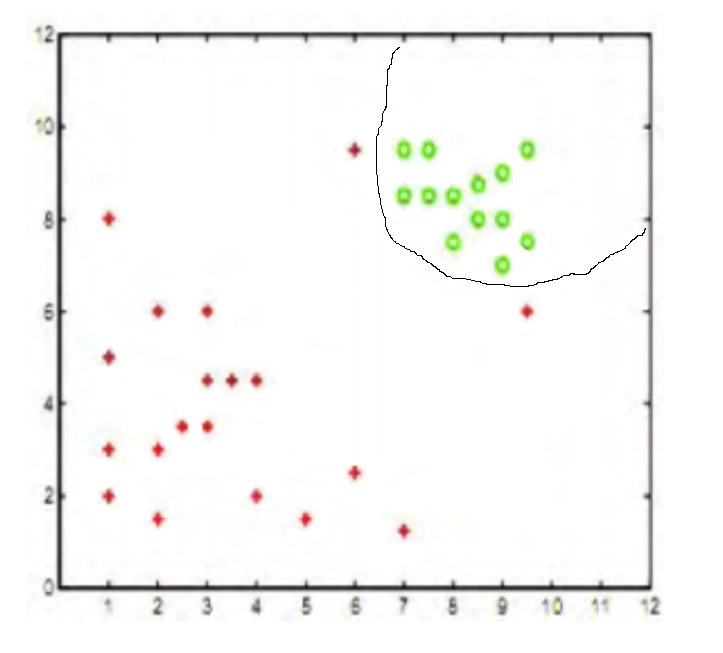
The highest value of k could be 13. If we go with k as 13, our model will wrongly predict all the data points and accuracy of the model would be zero. Hence a large value of k is not recommended. Too small k (k = 3, for example) will lead to overfitting. Therefore, we find an optimal value of k to avoid both of these cases.

* **Part 3**

Taking k equal to 5 or 7 minimizes the error. This model misclassifies four of the points and correctly predicts the rest of them. Hence the error comes out to be 4/14.

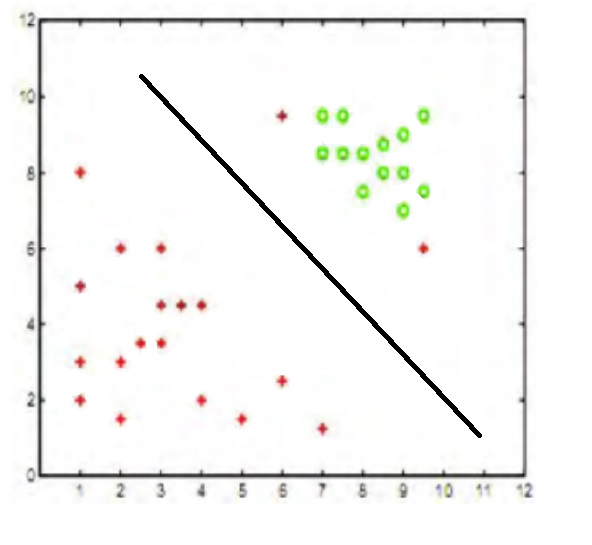
* **Solution 2**
* **Part 1**

If the value of C is very large, all the points would be classified correctly. Therefore, we can draw a parabola with minimum curvature as the classifier or decision boundary. The classifier is shown in the figure below with perfect classification -



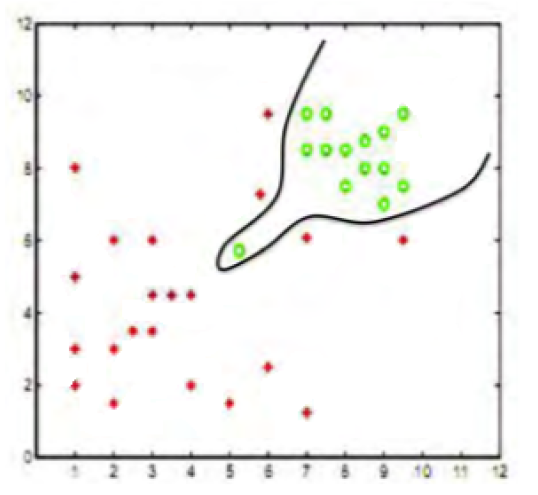
* **Part 2**

For the C value to be nearly zero, we can have a few points misclassified. Hence we can have the classifier as linear as shown below -

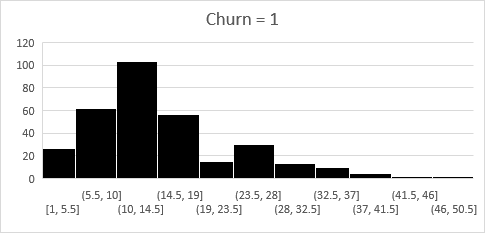


* **Part 3**

The phenomenon that will cause misclassification is called Overfitting. It happens when a very complex model is used to fit the training data. The overly complex model will work accurately on the training data but is deemed to fail on testing data.



* **Solution 3**
* **Part 1**



Wall’s assumption about the dependence of churn rate on the age of the customer is correct. As per Wall, if the customer is with the company for more than 14 months, the customer is less likely to leave the company. If the customer’s age is between 6 to 14 months, the risk of leaving is high.

By looking at the data, it is evident that churn is 1 for a lot of age values that are between 6 and 14 months. Moreover, churn rate is 0 for a lot of age values that are more than 14. These facts support Wall’s statement.

* **Part 2**

We start with loading the data into R studio. Second step would be to change the relevant attributes to Factors. Going forward, we get rid of the irrelevant columns i.e. the ID column. The R code for all these tasks is provided here -

*>d <- read.csv("D:/UIC Fall/Data Mining/HW/8/Assignments/HW7/Supplement/UV6696-XLS-ENG - Copy.csv")*

*>d$Churn = factor(d$Churn)*

*>d=d[,2:13]*

*>str(d)*

Now we run the logistics regression. The logistics regression summary is shown below

*>Logit = glm(Churn~., data = d, family = "binomial")*

*>summary(Logit)*

Summary -

Call:

glm(formula = Churn ~ ., family = "binomial", data = d)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.0050 -0.3542 -0.2956 -0.2328 3.0650

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -2.763e+00 1.069e-01 -25.841 < 2e-16 \*\*\*

Age 1.271e-02 5.370e-03 2.367 0.017937 \*

CHI.Score.Month.0 -4.661e-03 1.223e-03 -3.811 0.000138 \*\*\*

CHI.Score.0.1 -1.027e-02 2.474e-03 -4.152 3.29e-05 \*\*\*

Support.Cases.Month.0 -1.520e-01 1.049e-01 -1.449 0.147342

Support.Cases.0.1 1.699e-01 9.045e-02 1.879 0.060255 .

SP.Month.0 1.566e-02 1.022e-01 0.153 0.878157

SP.0.1 -5.164e-02 7.850e-02 -0.658 0.510634

Logins.0.1 2.897e-04 2.092e-03 0.138 0.889885

Blog.Articles.0.1 2.745e-04 1.961e-02 0.014 0.988831

Views.0.1 -1.098e-04 4.071e-05 -2.697 0.006989 \*\*

Days.Since.Last.Login.0.1 1.724e-02 4.289e-03 4.020 5.82e-05 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 2553.2 on 6347 degrees of freedom

Residual deviance: 2440.4 on 6336 degrees of freedom

AIC: 2464.4

Number of Fisher Scoring iterations: 7

From the summary, we can conclude that the relevant variables in the regression are Age, CHI Score (Month 0), CHI Score.0.1, Views0.1, and Days Since Last Login. We will consider only these variables while calculating the overall impact on Churn rate.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Intercept** | **Age** | **CHI (Month 0)** | **CHI.Score** | **Views** | **Days.Since.Last.Login** |
| -2.76E+00 | 1.27E-02 | -4.66E-03 | -1.03E-02 | -1.10E-04 | 1.72E-02 |
| **ID** | 672 |  |  |  |  |
|  | p/1-p | 3.94E-02 |  |  |  |
|  | **P (Churn =1)** | 0.04 |  |  |  |

The above table shows the value of probability calculated for the event Churn = 1 for the customer ID 672. The predicted probability that the person will leave between

December 2011 and February 2012 is very low. By looking at the actual data, we can see that the customer has not left. Hence our model is in line with the actual data.

For ID 354 -

|  |  |  |
| --- | --- | --- |
| ID | 354 |  |
|  | p/(1-p) | 5.02E-02 |
|  | P | 0.05 |

For ID 5203 -

|  |  |  |
| --- | --- | --- |
| ID | 5203 |  |
|  | p/(1-p) | 4.38E-02 |
|  | p | 0.04 |

As per the calculations, the probability for both the customer IDs 354 and 5203 to leave between December 2011 and February 2012 is low. Hence as per the predicted probability, they should not leave.

As per the actual data, both the customers have not left.

* **Part 3**

We can predict the probabilities for all the customers by the following command.

>*pred=predict(Logit,d, type="response")*

*>View(pred)*

*>pred = as.data.frame(pred)*

The top 100 customer list is attached below -

**

By looking at the coefficients of the relevant attributes, we can conclude that the top three factors that drive the churn rate are, Day since last login, CHI Score.0.1 and Age.