Project in CSE 250B
Assignment 2: Logistic Regression with Regularization

Andreas Landstad, Spencer Bliven, Jonas Hoelzler
Computer Science Department
University of California, San Diego
landstad.andreas@gmail.com, sbliven@ucsd.edu, jonas@hoelzler.de

Abstract

This project evaluates an email analytics dataset describing two email campaigns and a control group using machine learning methods. The datasets is analyzed using logistic regression with stochastic gradient descent as optimization method and a real-valued nonlinear regression training method. The results show that the outcome after applying the mens email campaign is 0.77 USD higher than when no email advertisement was applied. For the womens email campaign the outcome was 0.40 USD higher.

1. Description of the Dataset

Kevin Hillstrom, a well-known data mining consultant has published an email analytics dataset with 64,000 records each describing a customer [4, 5]. One third of these customers were randomly chosen to receive an email referred to as the Men's email, a second random third received a different email (the Women's email) and the remaining customers served as a control, receiving neither email. The dataset is almost perfectly equally divided between the two mailings (Men's and Women's) and an untreated control group (see Table 1).

Mailing	Men	Women	None	Total
Count	21,307	21,387	21,306	64,000
%	33.29%	33.42%	33.29%	100.00%

Figure 1. Partitioning of the email analytics dataset

2. Preprocessing: Feature Selection for Regression

The data is given as a comma separated file. To apply the logistic and linear regressions, the data has to be preprocessed. For the learning algorithm, customer information is encoded as $x \in \mathbb{R}^{25}$. The encoding 2 is described at the following.

First, the recency entry shows the number of months since the last purchase from one to twelve months. In this project it is encoded as a bit-vector of eleven bits (bit 1 to 11), where the ith bit is $'1'$ for a recency of i months and $'0'$ otherwise. The 12th bit is left out, since it only represents redundant information. The history segmentation is a categorization of the customers into seven categories, using how many dollars were spent in the last year. It is encoded again as bit-vector 12 to 17, while one bit is again left out due to redundancy. Entry 18 shows the real value of actual dollars spent in the past year. Entry 19 and 20 are true, if the customer purchased men's respectively women's merchandise in the past year. Bit 21 and 22 categorize the customer in its zip code area. Bit 21 is 1, if he is from a rural area, bit 22 is 1, if he is from a urban area. The bit for the suburban area is again left out. Bit 23 is the indicator, if the customer is a new customer mentioned the last twelve months. The channel describes the channel the customer purchased from in the past year. Bit 24 is $'1'$ for the phone, bit 25 is $'1'$ for the web. The bits are not mutually exclusive.

Because the learning rate is the same for every parameter, it is useful to scale the features x_j so that their magnitudes are similar for all j. The features are normalized having mean zero and variance 1, according to [3].

Hillstrom provides three outcome (dependent) variables indicating whether people visited the site during a two-week outcome period, whether they purchased at the site (conversion) during that period, and how much customers spent during the outcome period (zero, for those who didn't). These outcomes serves a labels in this model, where $visit,\ purchase \in \{0, 1\}$ and $spend \in \mathbb{R}$.

2.1. Model of Computation

As one can see in Table 1, three different $treatments$ are applied to the customers: $women's\ clothing$, $men's\ clothing$ and $no\ email$. For each of these $treatments$ the same computational model can be applied. These three sub-models are part of the overall model

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Recency											History segment					
1	2	3	4	5	6	7	8	9	10	11	\$0-\$100	\$100-\$200	\$200-\$350	\$350-\$500	\$500-\$750	\$750-\$1000

18	19	20	21	22	23	24	25
History	Mens	Women	Zip-Code		Newbie	Channel	
H	M	W	U	R	N	P	W

Figure 2. The Encoding of the as $x \in \mathbb{R}^{25}$

$$\begin{aligned} E[spend|x, treatment] = \\ E[spend|purchase, x, treatment] \cdot \\ \cdot p(purchase|visit, x, treatment) \cdot \\ \cdot p(visit|x, treatment) \end{aligned} \tag{1}$$

x is the vector of attribute values described in the section above. $visit$ and $purchase$ are binary random variables that have a certain probability of being true of false for each customer. $spend$ is a real-value random variable for each customer.

Goal of this equation is to find out the expected value of money spent, given a customers with a particular treatment.

The probability $p(visit|x,\ treatment)$ is the probability, that a customer with a particular treatment visits the online store.

$p(purchase|visit,\ x,\ treatment)$ is the probability that the customer with the particular treatment purchases. $visit$ must be given in this term, since a customer must have visited the store to purchase.

$E[spend|purchase,\ x,\ treatment]$ is the expected real-valued outcome for a customer with a treatment. Trivially, he has to make a purchase to be able to spend money.

These three terms have to be multiplied, since a customer, who spends money has also to visit the store and to purchase something.

3. Logistic and Linear Regression

The three terms of the equation above can be calculated by Logistic Regression and Linear Regression. Logistic Regression can be applied to the second and third term. The first term can be calculated using Linear Regression.

3.1. Logistic Regression

Logistic regression is a common model for binary classification [1]. Given an example $\mathbf{x} \in \mathbb{R}^d$, the probability of a label $y \in \{0, 1\}$ is given by the logistic sigmoid function

$$p(Y = 1 \mid X = \mathbf{x};\ \boldsymbol{\beta}) = \frac{1}{1 + \exp(-\boldsymbol{\beta} \cdot \mathbf{x})}$$

where $\boldsymbol{\beta}$ is a d-dimensional vector. We assume $x_0 = 1$ such that $\boldsymbol{\beta}_0$ acts as an intercept. Given a set of parameters $\boldsymbol{\beta}$, the logistic function can be used for classification:

$$\hat{y} = \arg\max_y\ p(Y = y \mid X = \mathbf{x};\ \boldsymbol{\beta})$$

To learn useful parameters $\boldsymbol{\beta}$ we utilize a training set $(X, Y) = \{(\mathbf{x}_i, y_i) \mid i = 1 \ldots N\}$. The optimal $\boldsymbol{\beta}^*$ is chosen such that it maximizes the regularized log conditional likelihood over the training set

$$\boldsymbol{\beta}^* = \arg\max_{\boldsymbol{\beta}}\ LCL(\boldsymbol{\beta};\ Y|X) - \alpha\ \parallel \boldsymbol{\beta} \parallel^2 \tag{2}$$

$$= \arg\max_{\boldsymbol{\beta}} \sum_{i=1}^{N} \log p(y_i \mid \mathbf{x}_i;\ \boldsymbol{\beta}) - \alpha\ \parallel \boldsymbol{\beta} \parallel^2 . \tag{3}$$

The optimal $\boldsymbol{\beta}^*$ is found numerically using stochastic gradient descent (SGD). An initial $\boldsymbol{\beta}_0 = \mathbf{0}$ is chosen. For each example in the training set, $\boldsymbol{\beta}$ is updated according to the rule

$$\boldsymbol{\beta}_{t+1} = \boldsymbol{\beta}_t + \lambda \nabla \left(LCL_i(\boldsymbol{\beta};\ Y|X) - \alpha\ \parallel \boldsymbol{\beta} \parallel^2 \right) \tag{4}$$

$$= \boldsymbol{\beta}_t + \lambda\ \left((y_i - p_i)\ \mathbf{x} - 2\alpha\boldsymbol{\beta} \right) \tag{5}$$

Choosing learning rate λ

The learning rate is chosen by trial and error. It is gradually decreasing depending on the epochs by

$$\lambda = \frac{1}{t_0 + t} \tag{6}$$

where t is the current epoch. $t_0 = 1/0.1$ seemed appropriate.

Choosing α for Tikhonov Regularization

A common approach to avoid overfitting, is to add the penalty $\alpha\ \parallel \boldsymbol{\beta} \parallel^2$. This idea is called Tikhonov Regularization. αs in 0.1-steps between 0 and 1 were tested using cross-validation. Best results were found for $\alpha = 0$ i.e. no penalty is applied.

3.2. Linear Regression

By using Linear Regression one can predict an real-valued label as a linear combination of features. The linear model is

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_d x_d = \boldsymbol{\beta}\mathbf{x},$$

where $\mathbf{x}$ is an instance in $\mathbb{R}^d$ and y is a scalar. The intercept, β_0, is the predicted value of y when the other coefficients are all 0. Training is done on the coefficients $\boldsymbol{\beta}$ by an algorithm solving the following equation:

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \boldsymbol{\beta}\mathbf{x})^2 .$$

Here pairs $\langle \mathbf{x}_i, y_i \rangle$ are pairs from the training set. After having trained the model, we could predict a value for an x_i by

$$\hat{y}_i = \boldsymbol{\beta}\mathbf{x}$$

We used this to model the real valued expected spending of a customer given one of the three possible treatments (men's email, women's email or no email) and that the customer had done a purchase. ($E(\text{spend} \mid \text{purchase}, x, \text{treatments})$). A possible extension of the training of this model would have been to add a penalty function:

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \alpha \parallel \boldsymbol{\beta} \parallel^2 .$$

This is often helpful for reducing randomness in the training caused by linearly related coefficients. Alpha is the strength of the penalty, or regularization, and we used cross validation(CV) to find out that the alpha giving the lowest SSE on held out training data was $\alpha = 0$ or an α very close to zero. The womens email treatment result table for decimals between 0 and 1 is provided in the next section. The dollar amount is the summed SSE of the different folds in from the CV. The coefficients show clearly that for example recency and amount spent in the past is highly correlated with how much a person is going to spend (Figure 5).

Choosing α for Tikhonov Regularization

The same cross validation for choosing the α was applied to the linear regression (see results in Table 3).

α	Value
0.0	$2.28 \cdot 10^{-6}$
0.1	$3.59 \cdot 10^{-6}$
0.2	$4.19 \cdot 10^{-6}$
0.3	$4.72 \cdot 10^{-6}$
0.4	$4.20 \cdot 10^{-6}$
0.5	$5.32 \cdot 10^{-6}$
0.6	$4.35 \cdot 10^{-6}$
0.7	$3.96 \cdot 10^{-6}$
0.8	$3.93 \cdot 10^{-6}$
0.9	$4.34 \cdot 10^{-6}$
1.0	$4.33 \cdot 10^{-6}$

Figure 3. αs for linear regression with treatment women's email

4. Complexity Analysis

Let N be the number of training examples and d be the dimensionality of the data. Each update step of SGD requires computing the gradient with respect to $\boldsymbol{\beta}$ of regularized LCL for the current example. This takes time $O(d)$. Thus each epoch of SGD takes $O(Nd)$.

Most of the computational complexity comes from the need to repeat SGD multiple times. While this is a constant factor, it is quite significant. To achieve convergence, we observe that 200-250 epochs of SGD with decreasing λ are required. To improve the convergence rate we implemented Newtons Method for gradient descent. While our implementation operated in batch mode rather than stochastically modifying the $\boldsymbol{\beta}$ for each example, convergence was reached in only 10-15 epochs of Newtons Method, resulting in a dramatic overall speedup. While both methods have complexity $O(Nd)$, the decreasing the number of epochs resulted in a speedup from 278 seconds to 1.6 seconds ($N = 4524$, $D = 26$).

5. Evaluation

A trivial prediction of the amount of money a person is likely to spend would be the mean. The performance of such an algorithm is given in Table **??**. While this prediction gives the correct expected spending for each class, the RSSE is quite large. The machine learning method presented here also predicts the expected spending, but has a much lower RSSE on individual examples (Table **??**).

5.1. Evaluation of the weight-vector

In Figure 5 one can see the learned β-coefficients for the three regressions. The recency of the last purchase (coefficient 1 to11) plays a role. If the last visit or the last spending is too far away in into the past, then the coefficients are more negative correlated. It can also be seen, that advertisement is especially useful in the history segments between \$ 100 between \$ 500. If the customer is a newbie (coefficient 23), it is not so likely that hell spend a lot of money. Moreover, people from a rural area seem to be slightly better candidates for efficient advertisement.

Treatment	E[Spend|Trt]		RSSE	
	mean	std	mean	std
Men	1.4632	0.0196	0.2275	0.0115
Women	1.0887	0.0449	0.1991	0.0096
None	0.6905	0.0698	0.1578	0.0275

Figure 6. E[Spend—Trt] means the expected amount that the test examples are predicted to spend per person. This is calculated 5 times, yielding a mean and standard deviation. E[Spend—Trt]= 1/N * sum E[Spend—xi,Trt]. RSSE means the RSSE per example of the predicted-actual spend. $RSSE = 1/N * sqrt(sum(y_i - E[Spend|x_i, Trt])^2$

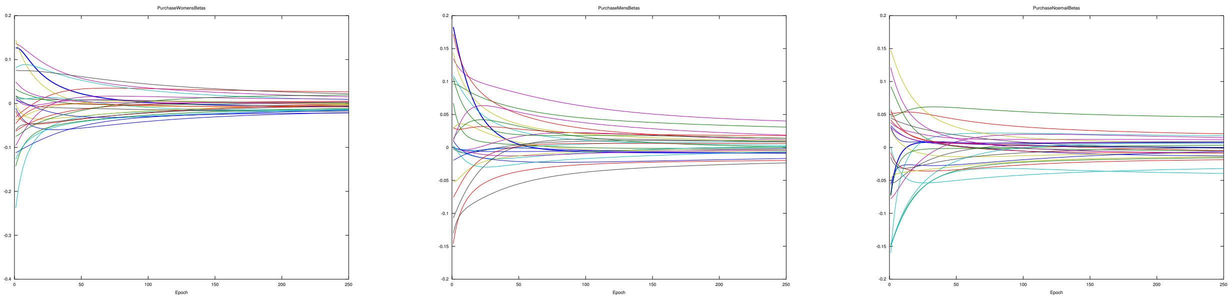

Figure 4. Purchase βs Conversion

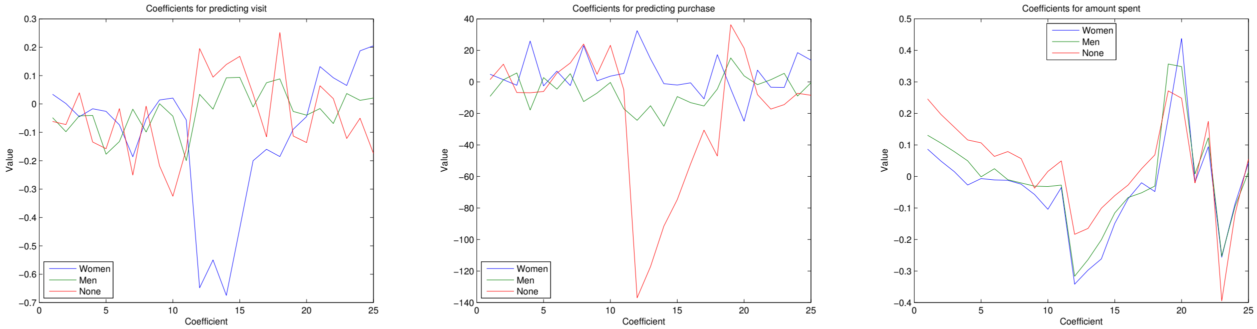

Figure 5. The β-Coefficients for the three treatments for predicting visit, purchase and amount spend (from left to right)

Treatment	E[Spend|Trt]		RSSE	
	mean	std	mean	std
Men	1.4226	0.0638	0.2686	0.0479
Women	1.0772	0.0794 7 0.2279	0.0431	
None	0.6528	0.0306	0.1735	0.0418

Figure 7. The trivial results for the same testing as in Figure **??**

6. Conclusion

The results show that e-mail advertisement makes a difference. The men's email leads to a higher outcome of 0.77 USD, the women's mail to a higher outcome of 0.40 USD. That means, that no more than this amount should be invested to increase sales. Table 8 shows the maximum amount one should spend for the Men's and Women's email campaigns to increase the sales for different numbers of customers.

It is recommended to send advertisement emails to customers who have previously purchased merchandise, spent money recently and are in a history segment between $ 100 and $ 500

Unfortunately it is very hard to predict the habitual buying behavior of single customers, since this behavior is very randomly distributed and depends on too many other influences. Using our simple models however, we can sometimes see which features that are important in regards to for example how much a person will spend given a treatment. The question of which campaign that has the greatest influence could have been answered with simple statistics. Machine learning only strengthens the confidence in the validity of this result and gives us an opportunity to see what factors were influential.

References

[1] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[2] L. Bottou. Stochastic gradient descent examples on toy problems, March 2011.

[3] C. Elkan. Nearest neighbor classification, 2011.

[4] K. Hillstrom. MineThatData E-Mail Analytics And Data Mining Challenge, March 20 2008.

[5] N. Radcliffe. Hillstrom's MineThatData Email Analytics Challenge: An Approach Using Uplift Modelling. *Stochastic Solutions Limited*, 1:1–19, 2008.

Number of Customers	500	1,000	2,500	5,000	10,000	20,000	50,000	100,000
USD for Men’s mail	386,-	773,-	1,932,-	3,864,-	7,727,-	15,454,-	38,635,-	77,270
USD for Women’s mail	199,-	398,-	996,-	1,991,-	3,982,-	7,964,-	19,991,-	39,820,-

Figure 8. Maximum amount one should spend for the Men’s and Women’s email campaigns to increase sales depending on the number of customers