

Active Learning via Sequential Design with Applications to Detection of Money Laundering

Xinwei Deng¹, V. Roshan Joseph¹, Agus Sudjianto², C. F. Jeff Wu^{1,3}

¹ H. Milton Stewart School of Industrial and Systems Engineering,

Georgia Institute of Technology, Atlanta, GA 30332

²Bank of America, Charlotte, NC 28255

Abstract

Money laundering is an act to hide the true origin of funds by sending them through a series of seemingly legitimate transactions. Because it often involves criminal activities, financial institutions have the responsibility to detect and inform about them to the appropriate government agencies in a timely manner. However, detecting money laundering is not an easy job because of the huge number of transactions that take place each day. The usual approach adopted by financial institutions is to extract some summary statistics from the transaction history and do a thorough and time-consuming investigation on those accounts that appear to be suspicious. In this article, we propose an active learning method using Bayesian sequential designs to identify the suspicious accounts. The method uses a combination of stochastic approximation and D -optimal designs to judiciously select the accounts for investigation. The sequential nature of the method helps to identify the suspicious accounts with minimal time and effort. A case study with real banking data is used to demonstrate the performance of the proposed method. A simulation study shows the efficiency and accuracy of the proposed method, as well as its robustness to model assumptions.

Keywords: Pool-based Learning, Stochastic Approximation, Optimal Design, Bayesian Estimation, Threshold Hyperplane.

1 Introduction

Money laundering is an act to hide the true origin of funds by sending them through a series of seemingly legitimate transactions. Its main purpose is to conceal the fact that funds were

³Address for correspondence: C. F. Jeff Wu. Email: jeffwu@isye.gatech.edu

acquired as a result of some form of criminal activity. These laundered funds can in turn be used to foster further illegal activities such as the financing of terrorist activity or trafficking of illegal drugs. Even legitimate funds that are laundered to avoid reporting them to the government, as is the case with tax evasion, lead to substantial costs for society. Financial institutions which have the responsibility to detect and prevent money laundering are facing a challenge to sort through potential suspicious activities among millions of legitimate transactions every day. Once suspicious activities have been detected, an investigation effort can easily take 10 hours to classify a case as suspicious or non-suspicious. Figure 1.1 shows a sample of transaction data. There are all kinds of information in the transaction history. Therefore, investigating every account to detect money laundering is extremely time-consuming and cost prohibitive.

Acct#	D/C	PostDate	TransAmt	TranCode	Description
999999999	D	1/24/2006	\$1,295.00	9059	09059 CHECK CHECK
999999999	D	5/19/2005	\$1,020.00	9059	09059 CHECK CHECK
999999999	D	1/24/2006	\$10,000.00	9059	09059 CHECK CHECK
999999999	D	3/2/2005	\$5.00	9593	09593 RETURNED ITEM CHARGE RETURNED ITEM CHARGE
999999999	D	2/24/2005	\$5.00	9593	09593 RETURNED ITEM CHARGE RETURNED ITEM CHARGE
999999999	D	10/12/2005	\$34.00	9203	09203 OVERDRAFT CHARGE OVERDRAFT CHARGE
999999999	D	7/13/2005	\$60.00	9659	09659 CHECK CARD PURCHASE DR JM LAYTON AND EP LAY5194121949512823
999999999	D	6/10/2005	\$129.36	9905	09905 POS WITHDRAWAL COSTCO WHSE #0001 84426275161089999910830
999999999	D	6/14/2005	\$51.49	9905	09905 POS WITHDRAWAL BED, BATH & BEYO 84426275165089999914310
999999999	D	6/10/2005	\$168.44	9905	09905 POS WITHDRAWAL COSTCO WHSE #0001 84426275161089999910370
999999999	D	7/18/2005	\$34.84	9905	09905 POS WITHDRAWAL COSTCO WHSE #0001 84426275197089999916890
999999999	D	5/24/2005	\$33.20	9905	09905 POS WITHDRAWAL COSTCO GAS #00662 84426275144089999924800
999999999	D	6/22/2005	\$158.65	9905	09905 POS WITHDRAWAL BED, BATH & BEYO 84426275173089999922610
999999999	D	6/10/2005	\$190.64	9905	09905 POS WITHDRAWAL COSTCO WHSE #0001 84426275161089999910750
999999999	C	1/14/2005	\$100.00	9003	09003 DEPOSIT DEPOSIT
999999999	C	8/9/2005	\$20.00	9003	09003 DEPOSIT DEPOSIT
999999999	C	5/11/2005	\$10,000.00	9003	09003 DEPOSIT DEPOSIT
999999999	C	8/31/2005	\$3,300.00	9003	09003 DEPOSIT DEPOSIT 0831CA319P007160134679
999999999	C	6/29/2005	\$2,079.95	9003	09003 DEPOSIT DEPOSIT
999999999	C	10/6/2005	\$2,500.00	9003	09003 DEPOSIT DEPOSIT
999999999	C	1/30/2006	\$22.43	9699	09699 AUTOMATIC DEPOSIT DEPOSIT MERCHANT BANKCD 267917678885
999999999	C	1/30/2006	\$22.43	9699	09699 AUTOMATIC DEPOSIT DEPOSIT MERCHANT BANKCD 267917678885
999999999	C	6/16/2005	\$64.97	9660	09660 REVERSE CHECK CARD PURCHASE THE HOME DEPOT 4715 5166010183470016
999999999	C	7/21/2005	\$151.61	9660	09660 REVERSE CHECK CARD PURCHASE HARDWARE SALES 5202207788501885
999999999	C	9/19/2005	\$24.95	9660	09660 REVERSE CHECK CARD PURCHASE TWX**SPORTS ILLUSTRATED 5259000879500624
999999999	C	4/27/2005	\$14,032.37	9039	09039 DEPOSIT TO CLOSE ACCOUNT DEPOSIT TO CLOSE ACCOUNT
999999999	C	11/30/2005	\$3,243.59	9003	09003 DEPOSIT DEPOSIT
999999999	C	7/6/2005	\$400.00	9003	09003 DEPOSIT DEPOSIT
999999999	C	10/6/2005	\$2,981.07	9003	09003 DEPOSIT DEPOSIT
999999999	C	7/21/2005	\$100.00	9007	09007 MISCELLANEOUS DEPOSIT TRANSFER FROM CHECKING 22782403

Figure 1.1: A sample of transaction data

One way to overcome this problem is to extract certain statistical features based on the transaction history of each account. If these statistical features are highly representative for the suspiciousness for the transaction history, then they can be used to prioritize the accounts for investigation. The accounts with high priority are investigated thoroughly to

find their suspiciousness level.

The problem can be formulated as follows. Let $\mathbf{x} = (x_1, \dots, x_p)'$ be the vector of feature variables extracted from a transaction history. Let $Y = 1$ if the account is classified as suspicious and $Y = 0$ otherwise. Then, $P(Y = 1|\mathbf{x}) = F(\mathbf{x})$ gives the probability of suspiciousness at a given level of \mathbf{x} . When $F(\mathbf{x})$ exceeds a threshold probability α , we can investigate that account in detail. Assume that $F(\mathbf{x})$ is an increasing function in each x_i . Define the threshold hyperplane $l_{\mathbf{x}}$ at level α as

$$l_{\mathbf{x}} = \{\mathbf{x} : F(\mathbf{x}) = \alpha\}. \quad (1.1)$$

Now for a new account, if \mathbf{x} falls below $l_{\mathbf{x}}$, then we need not investigate that account further. But if \mathbf{x} falls above $l_{\mathbf{x}}$, we must investigate the account in detail. An institution may choose a reasonable α so that only a portion of accounts needs to be investigated. This scientific approach can significantly improve productivity by investigating cases that really matter.

The challenge is not only due to the huge amount of transactions each day, but also due to different kinds of business with money laundering activities. The behaviors of various business categories can be quite different. Even the behaviors of the same business category at different time periods appear to be different in money laundering activities. For example, given a specified suspicious level α , the threshold hyperplane for personal accounts can be completely different from that of small business accounts. Even the importance of statistical features can vary dramatically. Thus, when a new set of accounts is introduced, it is not likely to share the same threshold hyperplane from the past investigation.

It is important to develop a procedure for finding the threshold hyperplane efficiently. The problem is that $F(\mathbf{x})$ is unknown and therefore, $l_{\mathbf{x}}$ is also unknown. Data on \mathbf{x} and Y can be used to estimate $l_{\mathbf{x}}$. For this purpose, a training set of the investigated accounts is needed. However, labelling the suspiciousness (1 or 0) for a large number of accounts is time consuming and extremely expensive. It will be beneficial to find a way to minimize the number of investigated accounts and use them to construct effective threshold hyperplane. This calls for active learning methods (Mackay, 1992; Cohn et al., 1996; Fukumizu, 2000). Here, the learner actively selects data points to be added into the training set. *In this*

article, an active learning method that improves the process of money laundering detection is proposed.

The remaining part of the article is organized as follows. In section 2, we give the motivation for the proposed active learning method using sequential designs. Section 3 reviews some existing methods in sequential designs and the concept of optimal designs. The active learning via sequential design is proposed in Section 4. In Section 5, we implemented the proposed method into a real case study for detecting money laundering. Section 6 presents some simulation results to demonstrate the performance of the proposed active learning approach. Some discussions and conclusions are given in Section 7.

2 Motivation

To minimize the number of investigated accounts and use them to construct effective threshold hyperplane, we need to judiciously select the accounts for investigation. This call for the use of active learning in machine learning. Recently, active learning methods using support vector machines (SVM) were developed by several researchers (Tong and Koller, 2001; Schohn and Cohn, 2000; Campbell et al., 2000), which can be applied to the present problem.

For binary response, active learning with SVM is mainly for two-class classification. The decision boundary in SVM implements the Bayes rule $P(Y|\mathbf{x}) = 0.5$, which is a special case of (1.1). In money laundering detection, sometimes the interest lies in values other than $\alpha = 0.5$. It is important to find the threshold hyperplane at a higher value of α such as $\alpha = 0.75$. To address this point, we propose a new active learning method using sequential designs. The sequential nature of the method helps to identify the suspicious accounts with reasonable time and effort. In statistical design of experiments, the locations of training data points are chosen by the users so as to maximize the information in the experiment (Kiefer, 1959; Fedorov, 1972; Pukelsheim, 1993). In sequential designs, the training data points are selected sequentially, i.e., the next point to be selected for training is based on information gathered for previously trained data points. However, in money laundering detection, the problem is different from classic sequential design. Because the accounts

are already available, we cannot select arbitrary setting of accounts to get investigation response. Noting that the motivation of active learning in machine learning is closely related to sequential designs in statistics, in this article, we will exploit the synergies between these two approaches to develop a new active learning procedure based on sequential designs and optimal designs. It provides a more flexible way to get threshold hyperplane for different values of α .

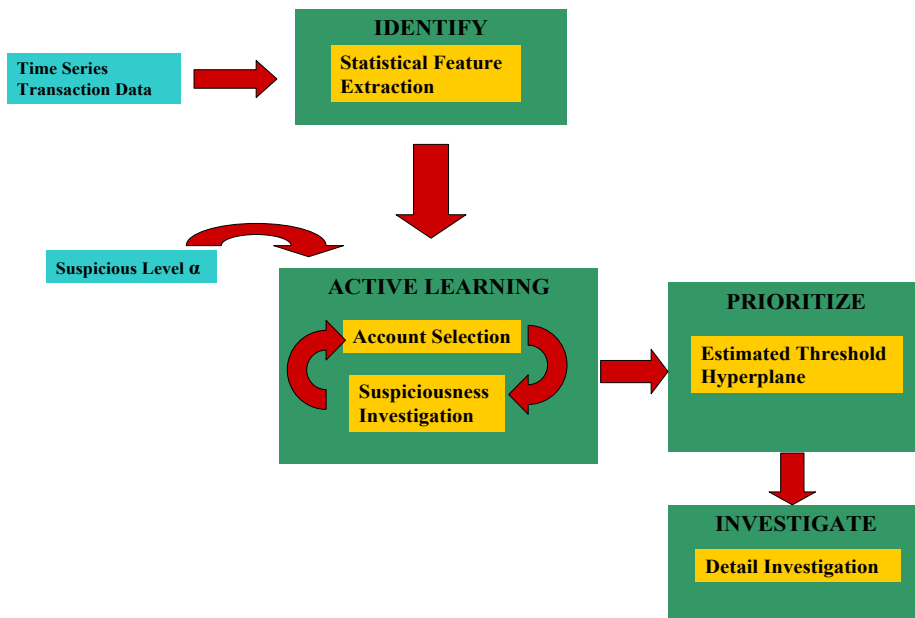


Figure 2.1: Active learning flowchart

Figure 2.1 shows the flowchart of active learning via sequential design applied to a typical money laundering investigation. Each account has its transaction history. Statistical features are extracted in the “IDENTIFY” step. Given a suspicious level α , active learning can be employed to judiciously select accounts for investigation iteratively in the “ACTIVE LEARNING” step. When the training set is ready, the estimated threshold hyperplane is obtained in the “PRIORITIZE” step. Then the investigators can focus on the accounts *above* the threshold hyperplane in “INVESTIGATE” step for detecting money laundering.

3 Review of Sequential Designs

The problem of estimating the threshold hyperplane is closely related to the problem of stochastic root-finding. Suppose we want to find the root of an unknown univariate function $E(Y|x) = F(x)$. The root can be estimated from the data $(x_1, Y_1), \dots, (x_n, Y_n)$. In sequential designs, the data points are chosen sequentially, i.e., x_{n+1} is selected based on x_1, x_2, \dots, x_n and their corresponding response Y_1, Y_2, \dots, Y_n . There are two approaches to generating sequential designs: stochastic approximation and optimal design.

In stochastic approximation methods, the x 's are chosen such that x_n converges to the root as $n \rightarrow \infty$. Robbins and Monro (1951) proposed the stochastic approximation procedure given by

$$x_{n+1} = x_n - a_n(Y_n - \alpha), \quad (3.1)$$

where $\{a_n\}$ is a pre-specified sequence of positive constants. They also established the conditions under which x_n converges to the root. This stochastic approximation method is also one of the classical pattern classification methods (Duda et al., 2001). An interesting modification of the Robbins-Monro procedure for binary data was proposed by Joseph (2004). Wu (1985) proposed another stochastic approximation method known as the “logit-MLE method”, in which $F(x)$ is approximated by a parametric function $H(x|\boldsymbol{\theta})$. Then, determination of x_{n+1} is a two-step procedure. First, a maximum likelihood estimator $\hat{\boldsymbol{\theta}}_n$ of $\boldsymbol{\theta}$ is found from $(x_1, Y_1), (x_2, Y_2), \dots, (x_n, Y_n)$. Then x_{n+1} is chosen as $H(x_{n+1}|\hat{\boldsymbol{\theta}}_n) = \alpha$. Ying and Wu (1997) showed the convergence of x_n almost surely irrespective of the function $F(x)$. Because of the efficient utilization of the complete data, the logit-MLE performs better than the Robbins-Monro procedure. Joseph et al. (2007) proposed a stochastic approximation method that gives more weights to data points closer to the root via a Bayesian scheme.

In the optimal design approach to sequential designs, first a parametric model for the unknown function is postulated. Then, the x points are chosen sequentially based on some optimality criteria (Kiefer, 1959; Fedorov, 1972; Pukelsheim, 1993). For example, Neyer (1994) proposed a sequential D -optimality-based design. Here x_{n+1} is chosen so that the determinant of the estimated Fisher information is maximized. It is well known that a D -

optimal criterion minimizes the volume of the confidence ellipsoid of the parameters (Silvey, 1980). The root is solved from the final estimate of the function $F(x)$.

The performance of the optimal design approach is model dependent. It performs best when the assumed model is the true model, but the performance deteriorates as the model deviates from the true model. One attractive property of the stochastic approximation methods, including the logit-MLE, is the robustness of their performance to model assumptions. This is because as n becomes large, the points get clustered around the root which enable the estimation of root irrespective of the model assumption. Understandably, the performance of the stochastic approximation method is not as good as the optimal design when the assumed model in the latter approach is valid. This point was confirmed by Young and Easterling (1994) through extensive simulations. In this article, we propose a new sequential design approach that combines the advantages of both approaches. Our approach is expected to be robust to model assumptions as in stochastic approximation methods as well as produce comparable performance to optimal design approach when the model assumptions are valid.

One shortcoming of the aforementioned methods is that they can only be applied to univariate problems. But in money laundering detection example and other applications (e.g., junk email classification), more than one statistical feature of the data are of interest. Therefore, it is important to extend the existing methods to multivariate problems. We propose a simple approach to account for the multivariate nature of the data. The methodology is explained in the next section.

4 Methodology

4.1 Active Learning via Sequential Design

In pool-based active learning (Lewis and Gale, 1994), there is a pool of unlabelled data. The learner has access to the pool and can request the true label for a certain number of data in the pool. The main issue is in finding a way to choose the next unlabelled data point to get the response. The proposed active learning via sequential design attempts to get “close in”

on the region of interest efficiently, meanwhile improves the estimation accuracy of $l_{\mathbf{x}}$ for a given α .

For the ease of exposition, we explain the methodology with two variables $\mathbf{x} = (x_1, x_2)^T$. It can be easily extended to more than two variables. We assume that each variable has a positive relationship with the response, i.e., for larger value of x_j , the probability is higher to get the response $Y = 1$. Define a *synthetic variable* z by $z = wx_1 + (1 - w)x_2$, where w is an *unknown* weight factor in $[0, 1]$. By doing this we can convert the multivariate problem into a univariate problem, so that the existing methods for sequential designs can be easily applied.

As in the case of Wu's logit-MLE method, assume the model

$$F(\mathbf{x}|\boldsymbol{\theta}) = \frac{e^{(z-\mu)/\sigma}}{1 + e^{(z-\mu)/\sigma}}, \quad (4.1)$$

which has three parameters $\boldsymbol{\theta} = (\mu, \sigma, w)^T$. As noted before, its convergence is independent of the logit model assumption. By the definition in (1.1), the threshold hyperplane $l_{\mathbf{x}}$ is

$$l_{\mathbf{x}} = \{\mathbf{x} = (x_1, x_2)^T : \frac{z - \mu}{\sigma} = \log\left(\frac{\alpha}{1 - \alpha}\right), \text{ where } z = wx_1 + (1 - w)x_2\}. \quad (4.2)$$

Let \mathcal{X} be the pool of data. Suppose we have $(\mathbf{x}_1, Y_1), (\mathbf{x}_2, Y_2), \dots, (\mathbf{x}_n, Y_n)$ in the training set. Based on this training data, we can estimate the threshold hyperplane $l_n = \{\mathbf{x} : F(\mathbf{x}|\hat{\boldsymbol{\theta}}_n) = \alpha\}$ by

$$l_n : \hat{w}_n x_1 + (1 - \hat{w}_n)x_2 = \hat{\mu}_n + \hat{\sigma}_n \log\left(\frac{\alpha}{1 - \alpha}\right), \quad (4.3)$$

where $\hat{\boldsymbol{\theta}}_n = (\hat{\mu}_n, \hat{\sigma}_n, \hat{w}_n)^T$ is estimated from the labelled data $(\mathbf{x}_1, Y_1), (\mathbf{x}_2, Y_2), \dots, (\mathbf{x}_n, Y_n)$. The details of the estimator $\hat{\boldsymbol{\theta}}_n$ are described in Section 3.2. Now, using the idea in stochastic approximation, we choose the next point from \mathcal{X} as the closest to the estimated hyperplane. Note that we have to choose the closest point because none of the points in \mathcal{X} may fall on the hyperplane. Thus

$$\mathbf{x}_{n+1} = \arg \min_{\mathbf{x} \in \mathcal{X}} |F(\mathbf{x}|\hat{\boldsymbol{\theta}}_n) - \alpha|. \quad (4.4)$$

There can be multiple points satisfying (4.4) because $\mathbf{x} \in \mathbb{R}^2$. Moreover, as pointed out in the previous section, the stochastic approximation method produces points clustered around

the true hyperplane, which leads to poor estimation of some of the parameters in the model. We can overcome these problems by integrating the above approach with the optimal design approach.

First, we choose k_0 points as candidates which are closest to the estimated threshold hyperplane l_n . Denote them as $\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \dots, \tilde{\mathbf{x}}_{k_0}$. Then, we select the next point as the one maximizing the determinant of the Fisher information matrix among the candidates. Thus

$$\mathbf{x}_{n+1} = \arg \max_{\mathbf{x} \in \{\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \dots, \tilde{\mathbf{x}}_{k_0}\}} \det(I(\hat{\boldsymbol{\theta}}_n, \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n, \mathbf{x})). \quad (4.5)$$

The Fisher information matrix for $\boldsymbol{\theta}$ can be calculated as

$$I(\boldsymbol{\theta}, \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n) = \sum_{i=1}^n \frac{e^{g(\mathbf{x}_i)}}{(1 + e^{g(\mathbf{x}_i)})^2} \frac{\partial g(\mathbf{x}_i)}{\partial \boldsymbol{\theta}} \frac{\partial g(\mathbf{x}_i)}{\partial \boldsymbol{\theta}^T}, \quad (4.6)$$

where $g(\mathbf{x}) = (z - \mu)/\sigma$, $z = wx_1 + (1 - w)x_2$ and $\boldsymbol{\theta} = (\mu, \sigma, w)^T$. The foregoing approach inherits the advantages of both stochastic approximation and optimal design. The stochastic approximation method in (4.4) can produce reasonable estimates of μ and σ , but can be very poor in the estimation of w . Because the D -optimality criterion in (4.5) ensures that the chosen points are well-spread, we can get a better estimate of w .

The improved estimation in our approach can be shown by considering the following version of the problem. Assume that there is at least one point in \mathcal{X} that lies in the hyperplane l_n . Then, the selected point \mathbf{x}_{n+1} is the solution of the following optimization problem:

$$\begin{aligned} \max_{\mathbf{x}} \det(I(\hat{\boldsymbol{\theta}}_n, \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n, \mathbf{x})) \\ \text{s.t. } \hat{w}_n x_1 + (1 - \hat{w}_n) x_2 = \hat{\mu}_n + \hat{\sigma}_n \log\left(\frac{\alpha}{1 - \alpha}\right). \end{aligned} \quad (4.7)$$

As shown in the Appendix, it is equivalent to

$$\begin{aligned} \max_{\mathbf{x}} \boldsymbol{\eta}_x^T I^{-1}(\hat{\boldsymbol{\theta}}_n, \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n) \boldsymbol{\eta}_x \\ \text{s.t. } \hat{w}_n x_1 + (1 - \hat{w}_n) x_2 = \hat{\mu}_n + \hat{\sigma}_n \log\left(\frac{\alpha}{1 - \alpha}\right), \end{aligned} \quad (4.8)$$

where $\boldsymbol{\eta}_x = (-1/\sigma, -\log(\alpha/(1 - \alpha)), (x_1 - x_2)/\sigma)^T$. The objective function in (4.8) is precisely the estimated variance of the hyperplane where the data is collected. It gives us

more accurate estimation when the response is acquired at the point with largest uncertainty. Thus, we select the point to be labelled such that the expected information gain is maximized. Note that the objective function in (4.8) is associated with \mathbf{x} only through $\boldsymbol{\eta}_x$. It maximizes a quadratic form in terms of $(x_1 - x_2)$. Therefore, the optimal value is achieved on the boundary of the feasible region of $(x_1 - x_2)$. The point selected by (4.5) is expected to be not close to the previous selected ones when they are projected onto the estimated threshold hyperplane l_n . This is why the proposed approach can provide a more stable estimation of the parameter w . A pseudo code of the proposed approach is shown in Figure 4.1.

Figure 4.1: The proposed active learning algorithm

Input: α value.

Suppose n data points are in the training set.

While Check stopping criterion,

Step 1, Find efficient estimators $\hat{\boldsymbol{\theta}}_n$ of $\boldsymbol{\theta}$ from $(\mathbf{x}_i, Y_i)_1^n$.

Step 2, Choose k_0 candidate points which are closest to the estimated threshold hyperplane $\hat{l}_{\mathbf{x}} = \{\mathbf{x} : F(\mathbf{x}|\hat{\boldsymbol{\theta}}_n) = \alpha\}$.

Step 3, Select the next point \mathbf{x}_{n+1} by (4.5).

Step 4, Get the response Y_{n+1} for \mathbf{x}_{n+1} .

Step 5, Set $n = n+1$.

End

Output: $\hat{l}_{\mathbf{x}} = \{\mathbf{x} : F(\mathbf{x}|\hat{\boldsymbol{\theta}}_n) = \alpha\}$.

4.2 Estimation

Since (4.1) is a probabilistic model, it is tempting to consider maximum likelihood estimation (MLE) for the parameter $\boldsymbol{\theta}$. Suppose the labelled data are $(\mathbf{x}_1, Y_1), (\mathbf{x}_2, Y_2), \dots, (\mathbf{x}_n, Y_n)$. It is known that the existence and uniqueness of MLE can be achieved only when successes

and failures overlap (Silvapulle, 1981; Albert and Anderson, 1984; Santner and Duffy, 1986). However, even when we are able to compute the MLE, they may suffer from low accuracy due to the small sample size, especially for nonlinear models. Use of a Bayesian approach with proper prior distribution for the parameters can overcome these problems.

We use the following priors:

$$\mu \sim \text{N}(\mu_0, \sigma_\mu^2), \quad \sigma \sim \text{Exponential}(\sigma_0), \quad w \sim \text{Beta}(\alpha_0, \beta_0). \quad (4.9)$$

A normal prior is specified for the location parameter μ . The scale parameter σ is nonnegative since each x_i is assumed positively related with the response Y . Therefore, an exponential prior with mean σ_0 is used as the prior for σ . Because w is a weight factor in $[0, 1]$, a beta distribution is a reasonable prior for w .

Assuming μ, σ and w are independent with each other, the overall prior for $\boldsymbol{\theta}$ is the product of the priors for each of its components. Thus, the posterior distribution is

$$f(\boldsymbol{\theta}|\mathbf{Y}) \propto \prod_{i=1}^n \left(\frac{e^{(z_i - \mu)/\sigma}}{1 + e^{(z_i - \mu)/\sigma}} \right)^{Y_i} \left(\frac{1}{1 + e^{(z_i - \mu)/\sigma}} \right)^{1 - Y_i} e^{-\frac{(\mu - \mu_0)^2}{2\sigma_\mu^2}} \lambda_0 e^{-\lambda_0 \sigma} w^{\alpha_0 - 1} (1 - w)^{\beta_0 - 1}, \quad (4.10)$$

where $z_i = wx_{i1} + (1 - w)x_{i2}$ and $\mathbf{x}_i = (x_{i1}, x_{i2})^T$. Finding the posterior mean of the parameters is difficult because it involves a complicated multidimensional integration. The maximum-a-posterior (MAP) estimators are much easier to compute. The MAP estimators of μ, σ , and w are obtained by solving

$$\hat{\boldsymbol{\theta}}_n = (\hat{\mu}_n, \hat{\sigma}_n, \hat{w}_n)^T = \arg \max_{\boldsymbol{\theta}} \log f(\boldsymbol{\theta}|\mathbf{Y}), \quad (4.11)$$

where

$$\begin{aligned} \log f(\boldsymbol{\theta}|\mathbf{Y}) \triangleq & \sum_{i=1}^n \frac{z_i - \mu}{\sigma} Y_i - \sum_{i=1}^n \log(1 + \exp(\frac{z_i - \mu}{\sigma})) \\ & - \frac{(\mu - \mu_0)^2}{2\sigma_\mu^2} - \lambda_0 \sigma + (\alpha_0 - 1) \log(w) + (\beta_0 - 1) \log(1 - w). \end{aligned} \quad (4.12)$$

Because proper prior distributions are employed, the optimization in (4.11) is well defined even when $n = 1$. Thus, this Bayesian approach allows us to implement a *fully* sequential procedure, i.e., the proposed active learning method can begin from $n = 1$. This would not

have been possible with a frequentist approach (Wu, 1985), for which some initial sample is necessary before the active learning method can be called. One advantage of using initial sample is that the approach will be more robust to prior specifications. In Section 6, we report a simulation study to compare the use of initial sample against a fully sequential procedure.

5 Case Study

Financial institutions invest much resources and efforts into detection of money laundering. We applied the proposed method to some real transaction data from a financial institution. The data in this example consists of 92 accounts from personal customers. It keeps the recent two-year transaction history for each customer. By working with expert investigators, we got a large set of summary variables. Then using multi-stage modelling and dimension reduction on these summary variables, we extracted two statistical features $\mathbf{x} = (x_1, x_2)^T \in \mathbb{R}^2$. Based on discussions with expert investigators, these two features can be highly representative of the suspiciousness for the transaction history, where x_1 describes the velocity and amount of money flowing through the account, and x_2 measures the differences of the transaction behaviors among the peer comparisons. For reasons of confidentiality, we do not disclose more details about the data being used here. Variables x_1 and x_2 are standardized to have zero mean and unit variance. The standardized data is shown in Figure 5.1.

We need to specify the prior for μ , σ , and w in (4.9) before active learning can be started. Here we use a heuristic procedure for doing this. First consider the prior for w . Assuming equal importance of x_1 and x_2 on the response, we would like the mean of w to be 0.5. Thus, set $\alpha_0/(\alpha_0 + \beta_0) = w_0 = 0.5$, which implies $\alpha_0 = \beta_0$. To get a flat prior, we take $\alpha_0 = \beta_0 = 3/2$. Thus, $w \propto w^{\frac{1}{2}}(1-w)^{\frac{1}{2}}$. Now consider the priors for μ and σ . Choose two extreme points (i.e., two accounts) \mathbf{x}_l and \mathbf{x}_u based on the lowest and highest values of z (denoted by z_l and z_u) through the mapping $z = w_0x_1 + (1-w_0)x_2$. We assume $\alpha_l = 5\%$ suspicious level for \mathbf{x}_l and $\alpha_u = 95\%$ suspicious level for \mathbf{x}_u . Plugging them into the model

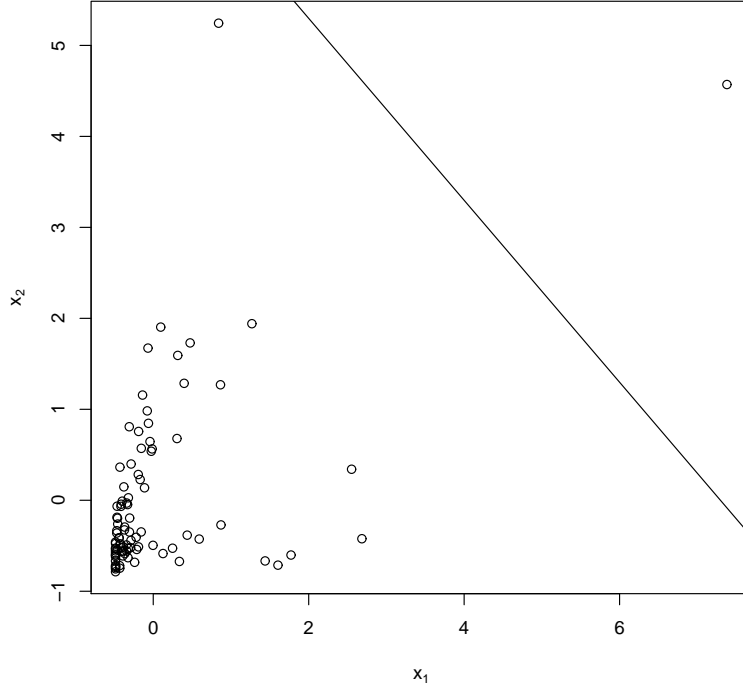


Figure 5.1: The standardized data. (Black line: the initial estimated threshold hyperplane by w_0, μ_0 and σ_0 .)

(4.1), we obtain

$$z_l = \mu + \sigma \log \frac{\alpha_l}{1 - \alpha_l},$$

$$z_u = \mu + \sigma \log \frac{\alpha_u}{1 - \alpha_u}.$$

Now, μ_0 and σ_0 are obtained by solving the above equations as

$$\mu_0 = \frac{z_l \log \frac{\alpha_u}{1 - \alpha_u} - z_u \log \frac{\alpha_l}{1 - \alpha_l}}{\log \frac{\alpha_u}{1 - \alpha_u} - \log \frac{\alpha_l}{1 - \alpha_l}} = \frac{z_l + z_u}{2}, \quad (5.1)$$

$$\sigma_0 = \frac{z_u - z_l}{\log \frac{\alpha_u}{1 - \alpha_u} - \log \frac{\alpha_l}{1 - \alpha_l}}. \quad (5.2)$$

We take σ_μ^2 as the sample variance of z_i , $i = 1, \dots, n$, where $z_i = w_0 x_{i1} + (1 - w_0) x_{i2}$. This completes the prior specification for the three parameters.

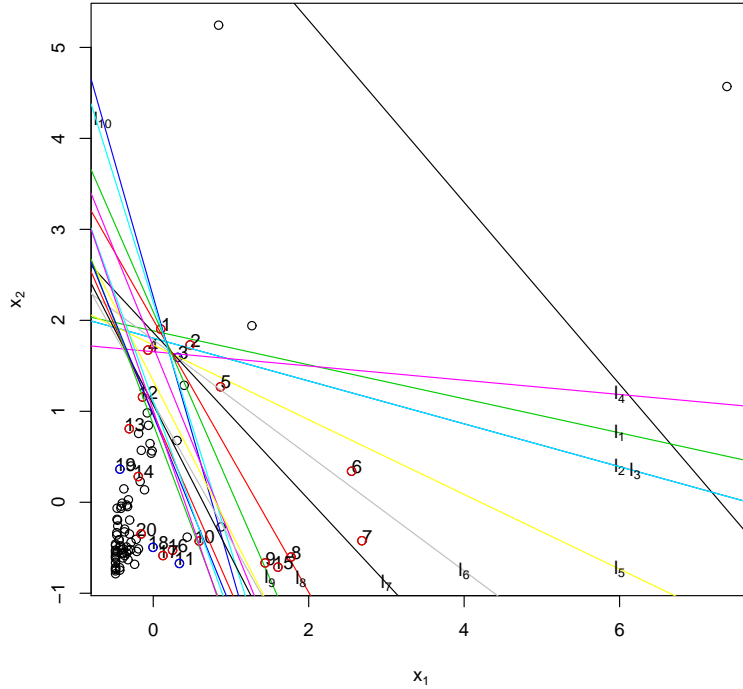


Figure 5.2: Active Learning via Sequential Design. (For example, yellow line l_5 stands for the estimated threshold hyperplane at iteration 5.)

Now the active learning method can be started. Suppose our objective is to find the threshold hyperplane with $\alpha = 0.75$. The initial estimated hyperplane based on only the prior is shown in Figure 5.1. The points are then selected one at a time using the procedure described in the previous section. In this example, we took $k_0 = 15$ in (4.5). The performance of the proposed method for the first 20 points is shown in Figures 5.2 and 5.3.

Figure 5.2 shows a series of estimated threshold hyperplanes using the proposed approach. The red data point in the figure means it is selected and the response is 1. The blue one means it is selected and the response is 0. At the beginning, there were large changes in the threshold hyperplane. In about 10-15 points it started to converge. The final estimated threshold hyperplane (i.e., after 20 points) is shown in Figure 5.3. The points above this hyperplane should be given higher priority and be investigated thoroughly for their suspiciousness. There

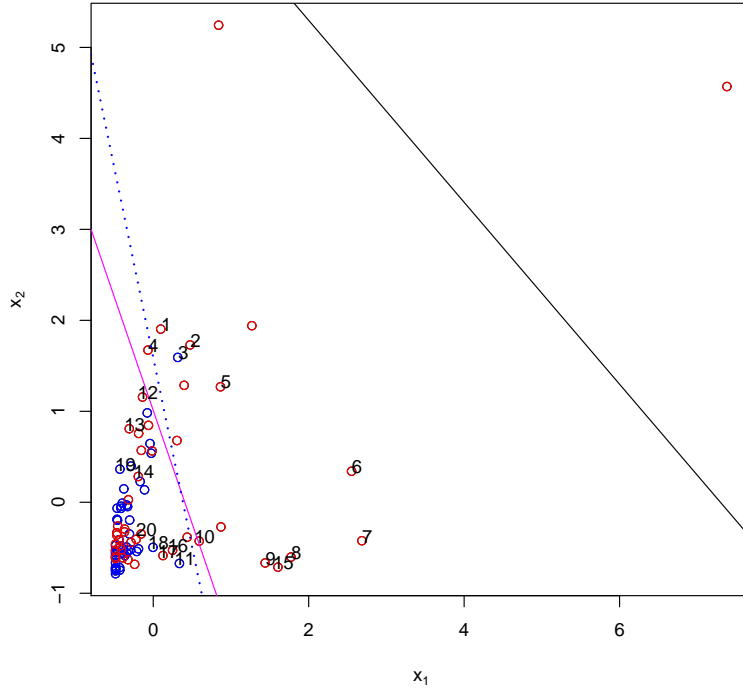


Figure 5.3: Comparison with the estimate based on full information. (Black line: the initial estimated threshold hyperplane by w_0, μ_0 and σ_0 . Pink line: the estimated threshold hyperplane after 20 points are sequentially selected. Blue dashed line: the estimated threshold hyperplane when all data are labelled.)

are only a few remaining accounts that need a thorough investigation, which clearly shows the efficiency of the proposed method.

To assess the accuracy of the proposed method, we asked the investigators at this financial institution to investigate all the 92 accounts carefully. Based on the obtained information for all the accounts, we estimated the threshold hyperplane, which is shown in Figure 5.3 as blue dashed line. We can see that it is very close to the estimated threshold hyperplane (i.e., pink line) by the active learning method. Thus, the proposed method can identify the true hyperplane by using only about 22% ($\approx 20/92$) of the data, which is a big saving for the financial institution.

To check the efficiency of the proposed method, we also compared the proposed method with a naive method. The naive method is to randomly select the next data point for getting the response. To gauge the performance of two methods, we measure the closeness between the estimated threshold hyperplane l_n and the true threshold hyperplane $l_{\mathbf{x}}$ when all data are labelled. The adopted measure is

$$\text{dist}(l_n, l_{\mathbf{x}}) \triangleq \sum_{\mathbf{t}_i \in \mathbf{T}} d_i^2, \quad (5.3)$$

where $\mathbf{T} = \{\mathbf{t}_i\}$ is a set of points which lie on the true threshold hyperplane $l_{\mathbf{x}}$, and d_i is the distance of \mathbf{t}_i to the estimated hyperplane l_n . Based on (5.3), a distance-based performance measure is defined as

$$\text{Dist_PM} = \frac{1}{M} \sum_{j=1}^M \text{dist}_j(l_n, l_{\mathbf{x}}), \quad (5.4)$$

where M is the number of simulations, and dist_j represents $\text{dist}(l_n, l_{\mathbf{x}})$ for the j -th simulation.

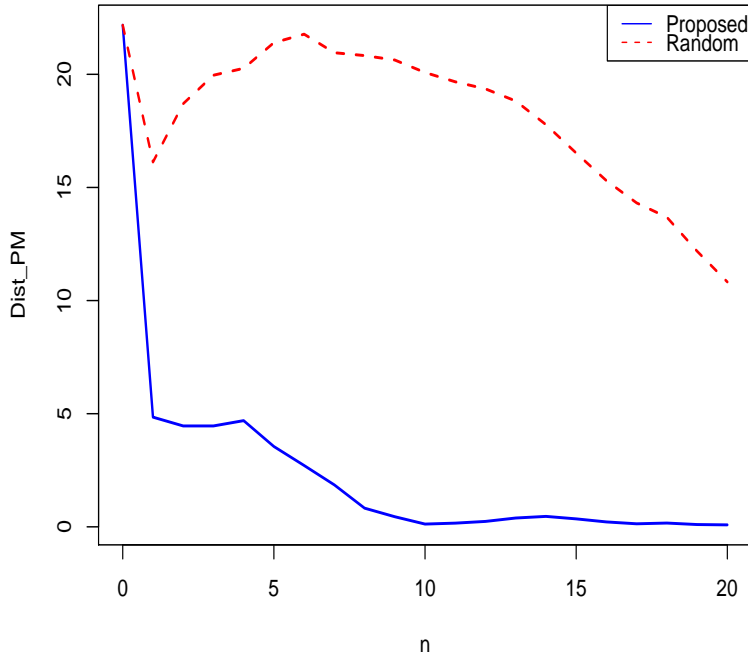


Figure 5.4: Learning Curves of Two Methods

Figure 5.4 shows the learning curves for the two methods. It is clear that the proposed method is much more efficient than the naive method. The estimated threshold hyperplane by the proposed method also moves towards the true threshold hyperplane quickly and consistently. The proposed method converges in about 10 steps for this problem.

6 Simulations

6.1 Numerical Examples

As stated before, the proposed method is expected to be more flexible and robust to model assumptions. Some experiments were conducted to study its performance. The simulated data were based on different models of $F(\mathbf{x})$. Four models were used in the study:

$$\text{Logistic distribution: } F(\mathbf{x}) = \frac{\exp(\frac{z-\mu}{\sigma})}{1 + \exp(\frac{z-\mu}{\sigma})},$$

$$\text{Uniform distribution: } F(\mathbf{x}) = \frac{\frac{z-\mu}{\sigma} - (-2)}{2 - (-2)},$$

$$\text{Normal distribution: } F(\mathbf{x}) = \Phi\left(\frac{z-\mu}{\sigma}\right),$$

$$\text{Cauchy distribution: } F(\mathbf{x}) = \frac{1}{2} + \frac{1}{\pi} \tan^{-1}\left(\frac{z-\mu}{\sigma}\right),$$

where $z = wx_1 + (1-w)x_2$ and Φ is the standard normal distribution function. The true values of parameters were set as $\mu = 0.5$, $\sigma = 1$ and $w = 0.7$. The response outcome at each point was generated according to $F(x)$.

In this simulation we chose $\alpha = 0.5$ and $\alpha = 0.8$ for illustration. The same performance measure in (5.4) is used here. Let $k_0 = 15$ in (4.5). The specification of hyper-parameters is done by using the heuristic procedure discussed in the previous section. 100 simulations were performed and $n = 30$ points were sequentially selected in each simulation.

Several methods are considered for comparison. We denote the fully sequential version of the proposed method as Method I. We denote by Method II the proposed method whose iterative scheme is preceded by choosing a fixed initial sample. To get a baseline comparison we used Method III, where the points are selected randomly, i.e., without using any active

learning method. For Method II, we used stratified random sampling to choose eight initial points. It is implemented as follows. With the initial guess on the parameters μ_0, σ_0 and w_0 , we can get $z = w_0x_1 + (1 - w_0)x_2$. Then we divide the range of z into four strata as $(-\infty, \mu_0 - 1.6\sigma_0)$, $[\mu_0 - 1.6\sigma_0, \mu_0)$, $[\mu_0, \mu_0 + 1.6\sigma_0)$ and $[\mu_0 + 1.6\sigma_0, +\infty)$. Since each point \mathbf{x} can be mapped into the z value, we randomly choose two $\mathbf{x}'s$ in each stratum according to the corresponding z value. The choice of the constant ± 1.6 is based on the asymptotic

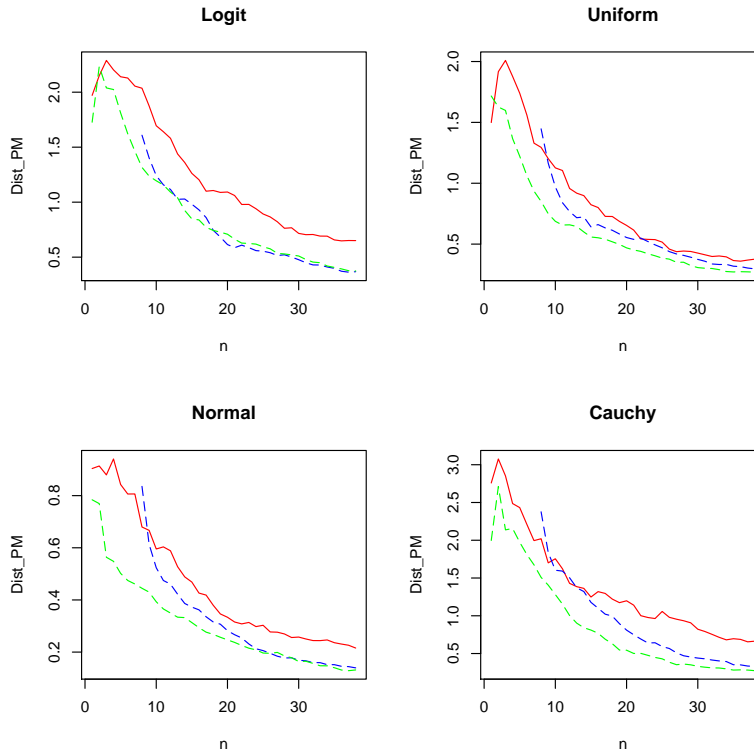


Figure 6.1: Dist_PM for four models with $\alpha = 0.5$. (Green line: method I. Blue line: method II. Red line: method III.)

optimality of the estimators under logistic distribution (see, e.g., Neyer 1994). Performance of the three methods for two chosen values of α are shown in Figures 6.1 and 6.2.

Clearly the proposed active learning methods (I and II) perform much better than Method III. Between I and II, Method I outperforms Method II. This is expected because Method I starts the active learning from the first point, whereas Method II starts active learning only

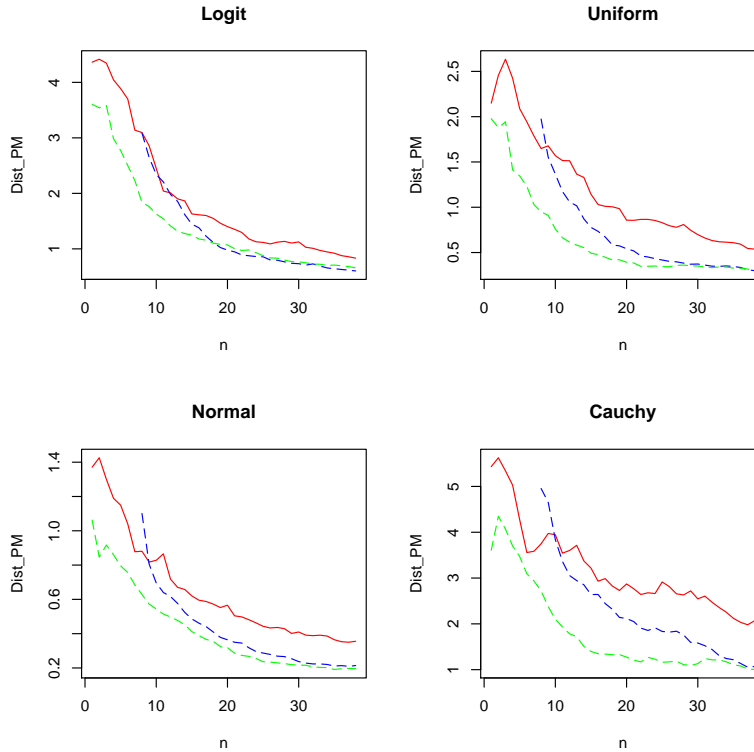


Figure 6.2: Dist_PM for four models with $\alpha = 0.8$. (Green line: method I. Blue line: method II. Red line: method III.)

after the selection of eight initial points. Unless stratified samples can be properly chosen, Method II will give inferior results. However, as n increases to about 30, its performance is comparable to that of Method I.

Comparing Figures 6.1 and 6.2, we can see that the performance of the methods is better when $\alpha = .5$. This is a well known fact in the literature that the estimation of extreme quantiles is much more difficult than with $\alpha = .5$ (see, e.g., Joseph 2004). It is also clear from the figures that the proposed methods are quite robust to model assumptions.

In the proposed active learning approach in (4.5), one selects k_0 candidate points which are closest to the estimated hyperplane. Here k_0 is considered as a tuning parameter but its optimal value has not yet been addressed. An additional experiment was conducted regarding the choice of k_0 . Setting $\alpha = 0.6$, the proposed active learning in a fully sequential

version (i.e., Method I) is performed for different k_0 , i.e., $k_0 = 1, 5, 10, 15$ and $k_0 = N$, where N is the total number of data points in the data set. $k_0 = 1$ means active learning using stochastic approximation, whereas $k_0 = N$ means active learning using a fully D -optimal-based sequential design. 100 simulations were generated for each k_0 and each model. The hyper-parameters were chosen as in Section 4. Figure 6.3 shows the simulation results .

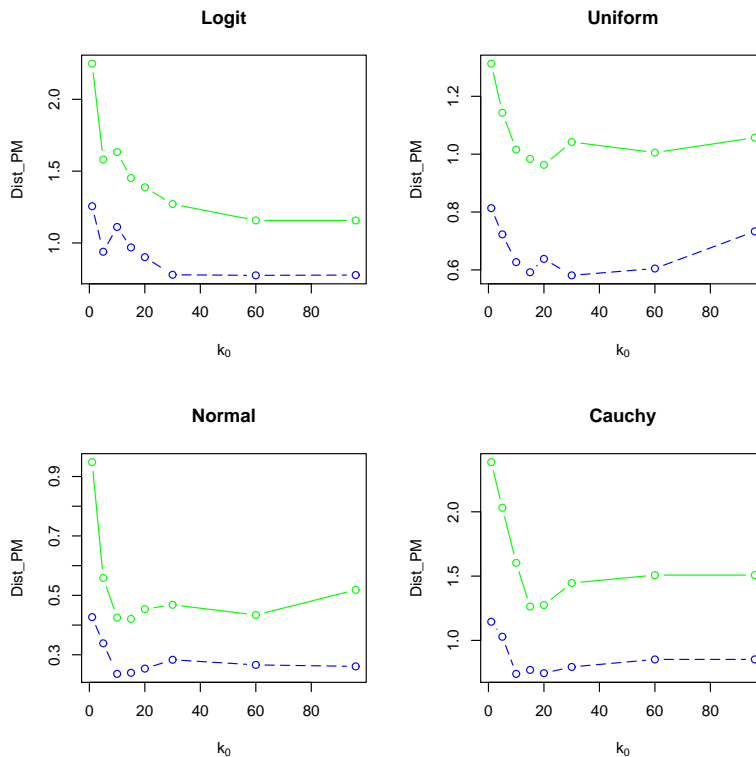


Figure 6.3: Performance with different k_0 . (Green line: $n = 10$; Blue line: $n = 20$.)

As can be seen in Figure 6.3, except for the logistic distribution the Dist_PM decreases up to some value of k_0 and then increases. This agrees with our initial intuition that choosing a large value of k_0 may not be good if the assumed model is not correct. Our procedure assumes the logistic model. Thus, when the model is changed to uniform, normal, or Cauchy, the method did not do well with a large k_0 . As expected, the performance did not deteriorate with k_0 when the true model is logistic. It is also clear that $k_0 = 1$ is a bad choice as the Dist_PM is the largest in all cases. Thus, using a purely stochastic approximation method

for active learning is not good in this particular problem. It is not clear what is the best value of k_0 . The simulation results suggest choosing k_0 to be 20%-50% of N .

6.2 Comparison with Support Vector Machine

Active learning using support vector machine (SVM) for classification has been proposed with several versions (Schohn & Cohn, 2001; Campbell et al., 2001; Tong & Koller, 2001). The basic idea is to label points that lie closest to the SVM's dividing hyperplane. It is known that the hyperplane in SVM converges to the Bayes rule $P(Y = 1|\mathbf{x}) = \alpha$, where $\alpha = 0.5$. The proposed active learning via sequential design can also converge to the threshold hyperplane when $\alpha = 0.5$. To start the active learning with SVM, some initial sample of points are needed. Therefore, to have a fair comparison, we used eight points as the initial sample chosen based on the stratified random sampling discussed in Section 5.1. The hyperparameters were chosen as before. 100 simulations were generated for comparison.

The Dist_PM values are plotted in Figure 6.4. We can see that the Dist_PM values of the proposed active learning (Method II) are much smaller than that of the active learning with SVM. Moreover, the proposed active learning is quite stable, whereas the SVM is quite unstable for small n . The SVM is not robust because adding one more point into the training set can cause big changes in the SVM's dividing hyperplane. Thanks to the use of the Bayesian approach, the estimation in the proposed active learning is stable.

The proposed active learning seems to converge within 20 steps, while the active learning with SVM needs at least 10 more steps to achieve similar performance. The improvement is even more pronounced with heavy tail distributions like Cauchy. Thus in this particular problem, the proposed active learning outperformed active learning with SVM in all aspects including accuracy, stability, and robustness.

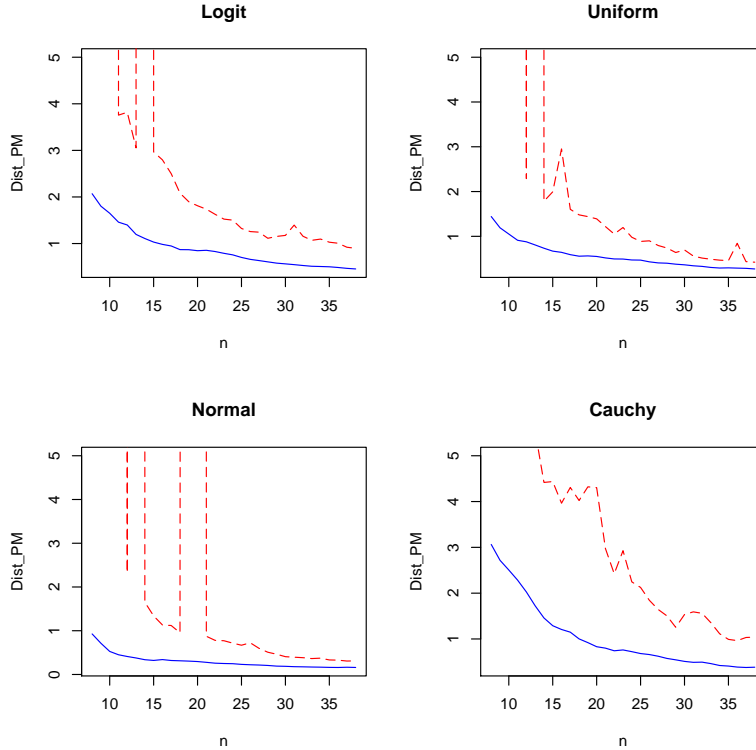


Figure 6.4: Comparison of Active Learning with SVM. Solid blue line: the proposed active learning (method II). Dashed red line: active learning with SVM.

7 Discussions and Conclusions

In this article, we propose an active learning via sequential design and report its application to a real world problem in money laundering detection. Due to the large amount of transactions and various business categories, it is crucial to find an efficient way to get the threshold hyperplane for prioritization. The proposed method is efficient and accurate for estimating the threshold hyperplane, and its performance is robust to model assumptions. It can help investigation to put more effort on those accounts with great importance. Therefore, this approach can significantly improve the productivity of money laundering detection.

The propose active learning method uses a combination of stochastic approximation and optimal design methods. From the sequential design perspective, we have shown that the proposed method works better than both stochastic approximation and optimal design.

Through simulations we have also shown that the proposed method outperforms active learning methods using SVM. With proper prior information, the fully sequential version of the proposed method (Method I) performs better than the one which starts with an initial sample (Method II). Regarding the choice of k_0 (i.e., the number of candidate points in (4.5)), the simulation study suggests choosing k_0 to be 20%-50% of N .

The proposed method is described for two variables $\mathbf{x} = (x_1, x_2)^T$. It can be easily extended to high dimensions. In multivariate situations where $\mathbf{x} = (x_1, x_2, \dots, x_p)^T$, we can define a synthetic variable z as a convex combination of the feature variables, i.e., $z = \sum_{i=1}^p w_i x_i$, where $w_i \geq 0$ and $\sum_{i=1}^p w_i = 1$. Then the active learning procedure is the same as the one described in Section 4.1. The prior for $w = (w_1, w_2, \dots, w_p)^T$ can be chosen to be a Dirichlet distribution.

The proposed active learning via sequential design is flexible in estimating threshold hyperplane for different α . On the other hand, the standard support vector machine is mainly for classification problem with $\alpha = 0.5$. Lin et al. (2002) proposed a modified support vector machine to account for α different from 0.5. It will be interesting to compare the proposed method with active learning using the modified support vector machine. Note, however, the absence of active learning method using the modified support vector machine in the literature.

Although the proposed method was motivated by the problem of detecting money laundering, the approach is quite general and can be applied to different classes of problems. For example, it can be used in sensitivity experiments (Neyer, 1994) or in bioassay experiments (McLeish and Tosh, 1990). Another advantage of the proposed method is that it can be applied to multivariate problems. For this to work, we need to assume the direction of the effect for each of the variables. This assumption seems to be reasonable in problems we have encountered so far.

Acknowledgements

The research of Deng, Joseph and Wu was supported in part by a grant from the U. S. Army Research Laboratory and the U. S. Army Research Office under contract number W911NF-05-1-0264.

Appendix

Equivalence between (4.7) and (4.8).

From (4.6), $I(\hat{\boldsymbol{\theta}}_n, \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n, \mathbf{x}) = I(\hat{\boldsymbol{\theta}}_n, \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n) + \kappa_x \boldsymbol{\eta}_x \boldsymbol{\eta}_x^T$, where $\kappa_x = e^{g(\mathbf{x})} / (1 + e^{g(\mathbf{x})})^2$ and $\boldsymbol{\eta}_x = \frac{\partial g(\mathbf{x})}{\partial \boldsymbol{\theta}}$. Under mild regularity conditions, the Fisher information matrix $I(\hat{\boldsymbol{\theta}}_n, \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)$ is positive semi-definite and nonsingular. Therefore, applying the identity $\det(A + c\mathbf{x}\mathbf{x}^T) = \det(A)(1 + c\mathbf{x}^T A^{-1}\mathbf{x})$, we obtain

$$\begin{aligned} \det(I(\hat{\boldsymbol{\theta}}_n, \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n, \mathbf{x})) &= \det(I(\hat{\boldsymbol{\theta}}_n, \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n) + \kappa_x \boldsymbol{\eta}_x \boldsymbol{\eta}_x^T) \\ &= \det(I(\hat{\boldsymbol{\theta}}_n, \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n))(1 + \kappa_x \boldsymbol{\eta}_x^T I^{-1}(\hat{\boldsymbol{\theta}}_n, \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n) \boldsymbol{\eta}_x). \end{aligned}$$

Thus $\min_{\mathbf{x}} \det(I(\hat{\boldsymbol{\theta}}_n, \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n, \mathbf{x}))$ is the same as $\min_{\mathbf{x}} \kappa_x \boldsymbol{\eta}_x^T I^{-1}(\hat{\boldsymbol{\theta}}_n, \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n) \boldsymbol{\eta}_x$. Now under the constraint in (4.7), $\kappa_x = \alpha(1 - \alpha)$ is a constant. Thus we get (4.8). Note that $\boldsymbol{\eta}_x = (-1/\sigma, -\log(\alpha/(1 - \alpha))/\sigma, (x_1 - x_2)/\sigma)^T$ under constraint in (4.7). \square

References

- Albert, A. and Anderson, J. A. (1984), “On the Existence of Maximum Likelihood Estimates in Logistic Regression Models,” *Biometrika*, 71, 1–10.
- Campbell, C., Cristianini, N. and Smola, A. (2000), “Query Learning with Large Margin Classifiers,” In *Proceedings of 17th International Conference on Machine Learning*, pages 111–118.
- Cohn, D. A., Ghahramani, Z. and Jordan, M. I. (1996), “Active Learning with Statistical Models,” *Journal of Artificial Intelligence Research*, 4, 129–145.

- Duda, R. O., Hart, P. E. and Stork, D. G. (2000), *Pattern Classification*, New York: John Wiley & Sons.
- Fedorov, V. V. (1972), *Theory of Optimal Experiments*, Academic Press, New York.
- Fukumizu, K. (2000), “Statistical Active Learning in Multilayer Perceptrons,” *IEEE Transactions on Neural Networks*, 11(1), 17–26.
- Joseph, V. R. (2004), “Efficient Robbins-Monro Procedure for Binary Data,” *Biometrika*, 91, 461–470.
- Joseph, V. R., Tian, Y. and Wu, C. F. J. (2007), “Adaptive Designs for Stochastic Root-Finding,” *Statistica Sinica*, 17, 1549–1565.
- Kiefer, J. (1959), “Optimum Experimental Designs,” *Journal of the Royal Statistical Society, Series B*, 21, 272–304.
- Lewis, D. and Gale, W. (1994), “A Sequential Algorithm for Training Text Classifiers,” In *Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 3–12, Springer-Verlag.
- Lin, Y., Lee, Y. and Wahba, G. (2002), “Support Vector Machines for Classification in Nonstandard Situations,” *Machine Learning*, 46, 191–202.
- MacKay, D. J. C. (1992), “Information-Based Objective Functions for Active Data Selection,” *Neural Computation*, 4(4), 590–604.
- McLeish, D. L. and Tosh, D. (1990), “Sequential Designs in Bioassay,” *Biometrics*, 46, 103–116.
- Neyer, B. T. (1994), “D-Optimality-Based Sensitivity Test,” *Technometrics*, 36, 61–70.
- Pukelsheim, F. (1993), *Optimal Design of Experiments*, New York: John Wiley & Sons.
- Robbins, H. and Monro, S. (1951), “A Stochastic Approximation Method,” *The Annals of Mathematical Statistics*, 22, 400–407.

- Santner, T. J. and Duffy, D. E. (1986), “A Note on A. Albert and J. A. Andersons Conditions for the Existence of Maximum Likelihood Estimates in Logistic Regression Models,” *Biometrika*, 73, 755–758.
- Schohn, G. and Cohn, D. (2000), “Less is More: Active Learning with Support Vector Machines,” In *Proceedings of the Seventeenth International Conference on Machine Learning*.
- Silvapulle, M. J. (1981), “On the Existence of Maximum Likelihood Estimators of the Binomial Response Model,” *Journal of the Royal Statistical Society, Series B*, 43, 310–313.
- Silvey, S. D. (1980), *Optimal Design*. London: Chapman and Hall.
- Tong, S. and Koller, D. (2001), “Support Vector Machine Active Learning with Applications to Text Classification,” *Journal of Machine Learning Research*, 2, 45–66.
- Vapnik, V. N. (1998), *Statistical Learning Theory*, New York: John Wiley & Sons.
- Wu, C. F. J. (1985), “Efficient Sequential Designs with Binary Data,” *Journal of the American Statistical Association*, 80, 974–984.
- Ying, Z. and Wu, C. F. J. (1997), “An Asymptotic Theory of Sequential Designs Based on Maximum Likelihood Recursions,” *Statistica Sinica*, 7, 75–91.
- Young, L. J. and Easterling, R. G. (1994), “Estimation of Extreme Quantiles Based on Sensitivity Tests: A Comparative Study,” *Technometrics*, 36, 48–60.