AUSTRIAN JOURNAL OF STATISTICS
Volume 40 (2011), Number 1 & 2, 125-135

Active Learning in Black-Box Settings

Neil Rubens[1], Vera Sheinman[2], Ryota Tomioka[3] and Masashi Sugiyama[4]
[1] University of Electro-Communications, Tokyo, Japan
[2] Japanese Institute of Educational Measurement, Tokyo, Japan
[3] University of Tokyo, Tokyo, Japan
[4] Tokyo Institute of Technology, Tokyo, Japan

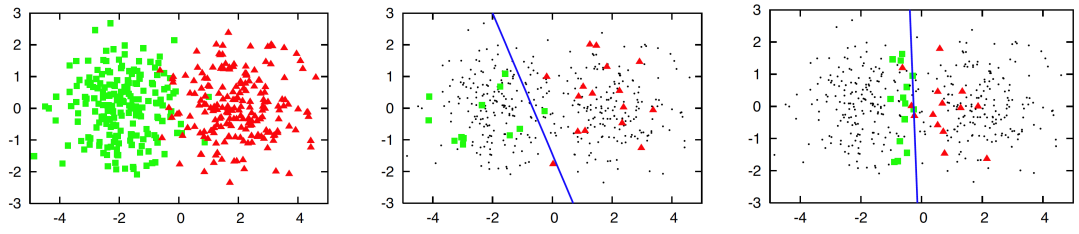
Abstract: Active learning refers to the settings in which a machine learning algorithm (learner) is able to select data from which it learns (selecting points and then obtaining their labels), and by doing so aims to achieve better accuracy (e.g., by avoiding obtaining training data that is redundant or unimportant). Active learning is particularly useful in cases where the labeling cost is high. A common assumption is that an active learning algorithm is aware of the details of the underlying learning algorithm for which it obtains the data. However, in many practical settings, obtaining precise details of the learning algorithm may not be feasible, making the underlying algorithm in essence a *black box* – no knowledge of the internal workings of the algorithm is available, and only the inputs and corresponding output estimates are accessible. This makes many of the traditional approaches not applicable, or at the least not effective. Hence our motivation is to use the only data that is accessible in black box settings – output estimates. We note that accuracy will improve only if the learner's output estimates change. Therefore we propose active learning criterion that utilizes the information contained within the changes of output estimates.

Keywords: Active Learning, Sampling, Experiment Design, Black Box Settings, Model Independent, Output Estimates.

1 Introduction

The goal of supervised learning is to learn a function that allows accurately predicting the output for previously unseen inputs. A function is learned from the training data consisting of inputs and outputs (labels) from the unknown target function. A popular phrase in computer science, "Garbage in, Garbage Out" summarizes well the importance of the training data in the learning process. Obtaining output values (labeling) often incurs a cost (in terms of money, effort, time, availability, etc.). While the cost of obtaining an output value is often the same, the degree to which a training point allows us to approximate the function varies (Figure 1). The goal of active learning (AL) is to select input points to label as to maximize the accuracy of the learned function. What makes the AL task challenging is that we have to predict the improvement in the accuracy of the learned function with regard to the input point before its output value (label) is obtained. This is because, once the output value is obtained, it incurs a cost.

A common assumption is that an active learning algorithm is aware of the details of the underlying learning algorithm for which it obtains the data (Figure 2a) (Settles, 2009).



(a) (b) (c)

Figure 1: Utilizing training points selected by an active learning method (1c), allows to more accurately predict the true values (1a), in comparison with selecting training points randomly (1b) (Settles, 2009).

(a) White Box Model. (b) Black Box Model.

Figure 2: For Black Box models (2b) only the inputs and outputs are accessible, internal workings are not accessible, unlike for White Box models (2a).

However, in many practical settings, obtaining details of the learning algorithm may not be feasible, e.g., learning algorithm could be very complex, consisting of many models and modules that are developed independently. For example, the most accurate predictive model of the NetFlix Challenge was an ensemble of over 100 models, developed by members in different parts of the world (Bell and Koren, 2007). Even if model details are available, developing active learning algorithms for complex models could be very difficult. In addition, active learning criterion may need to be reformulated each time the underlying model changes. Hence, in many situations, the underlying model is in practicality a *black box* – we can pass the inputs into the model and obtain the output estimates, but the inner workings of the model are unknown (Figure 2b).

Many active learning methods are inapplicable in black box settings, since they rely on the knowledge of at least some aspect of the model's workings, as indicated by recent surveys (Rubens, Kaplan, and Sugiyama, 2010; Settles, 2009). Variance-based active learning approaches are applicable, but are not effective for a number of reasons as outlined in Section 3.3. Since no information about the model is available, we propose to define an active learning criterion based on the indirect information available about the model – it's output estimates. We note that model's accuracy may improve only if its output estimates change (as a result of adding a new training point). In an attempt to speed up the improvements in accuracy of the model estimates, we propose to estimate the usefulness of labeling based on the magnitude of its impact on the estimates. We show that defining an active learning criterion by taking into account changes in the output estimates is a promising practical approach.

2 Problem Formulation

Supervised Learning Let us provide a brief formulation of supervised learning task – learning a function from training data. An input variable is considered to be a multi-dimensional data point and is denoted by a vector $\mathbf{x} \in \mathbb{R}^p$, where p is a number of attributes/features. The set of all points is denoted by $\mathcal{X}$. The target function that we are trying to approximate is denoted by f, and its output value (also referred to as label) is denoted as $f(\mathbf{x}) = y \in \mathbb{R}$, for simplicity we may also consider y to be a numerical label. The set of training input points is denoted by $\mathbf{X}$, and these points along with their corresponding output values are referred to as a *training set*, i.e. $\mathcal{T} = \{(\mathbf{x}_i, y_i\}_{\mathbf{x}_i \in \mathbf{X}}$. The task of supervised learning is, given a training set, to learn an estimate $\widehat{f}$ of the target function f. We measure how accurately the learned function predicts the true output values by the generalization error: $G(\widehat{f}) = \sum_{\mathbf{x} \in \mathcal{X}} \mathcal{L}\left(f(\mathbf{x}), \widehat{f}(\mathbf{x})\right)$. In practice, however, $f(\mathbf{x})$ is not available for all $\mathbf{x} \in \mathcal{X}$; it is therefore common to approximate the generalization error by the test error: $\widehat{G}(\widehat{f}) = \sum_{\mathbf{x} \in \mathbf{X}^*} \mathcal{L}\left(f(\mathbf{x}), \widehat{f}(\mathbf{x})\right) P(x)$, where $\mathbf{X}^*$ refers to the *test set*, and prediction errors are quantified by a loss function $\mathcal{L}$. For convenience we use the squared error (SE): $\mathcal{L}_{SE}\left(f(\mathbf{x}), \widehat{f}(\mathbf{x})\right) = \left(f(\mathbf{x}) - \widehat{f}(\mathbf{x})\right)^2$.

Active Learning We consider that we are allowed to sequentially select which points will be labeled. The active learning criterion aims to estimate the usefulness (in terms of generalization error) of labeling an input $\mathbf{x}_i$ (and adding it to the training set $\mathcal{T}$) as: $\widehat{G}(\mathbf{x}_i) = \widehat{G}(\widehat{f}_{\mathcal{T} \cup (\mathbf{x}_i, y_i)})$. For example, if we consider labeling a point $\mathbf{x}_j$ or a point $\mathbf{x}_k$, then we would estimate their usefulness by an active learning criterion, i.e. $\widehat{G}(\mathbf{x}_j)$ and $\widehat{G}(\mathbf{x}_k)$, and select the one that will result in a smaller generalization error. Note that we need to estimate the usefulness of labeling the point without knowing its actual label. The goal of active learning can then be stated as selecting an input point $\mathbf{x}$ to be labeled, so that after adding it to the training set the generalization error will be minimized: $argmin_{\mathbf{x}} \widehat{G}(\widehat{f}_{\mathcal{T} \cup (\mathbf{x}, y)})$.

Black-box Settings In black-box settings the details of learned function $\widehat{f}$ are not accessible, only its output estimates $\widehat{y} = \widehat{f}(\mathbf{x})$ are accessible.

3 Proposed Approach

Traditional model-based active learning approaches tend to aim at reducing the model error (i.e. the error of model parameters), which is hoped would result in the improvement of predictive error. However, in black box settings no information about the underlying model is assumed to be available. Therefore many of the traditional active learning methods are not effective or not even applicable in these settings. On the other hand, the output estimates are easily accessible. Motivated by this we aim at developing an active learning that utilizes the information contained within the output estimates.

3.1 Derivation

Let us provide the derivation and justifications of the proposed active learning criterion. The generalization error measures how well the estimated output values approximate the true output values. We note that in the calculation of the generalization error, the true output values are not affected by the addition of the new training point, while the estimates of the output values do change. Therefore, we propose to estimate the effect of a new training point on the value of the generalization error in terms of changes in the estimates of the output values.

First, let us reformulate the goal of minimizing the generalization error in terms of the changes in its value that adding a training point causes. Let us denote the generalization error when the number of training points is equal to t by G_t, the index of the next training point $\mathbf{x}_\delta$ by δ; and the generalization error after the output value y_δ is obtained by G_{t+1}. Let us express G_{t+1} as: $G_{t+1} = G_t - (G_t - G_{t+1})$. The value of G_t is fixed in advance (since we are considering a sequential scenario). The value of G_{t+1} depends on the choice of δ. In order for G_{t+1} to be minimized the difference between generalization errors G_t and G_{t+1} needs to be maximized i.e.: $\min_\delta G_{t+1} = G_t - \max_\delta(G_t - G_{t+1})$. So the original task of minimizing the generalization error could be reformulated as maximizing the difference between the generalization errors G_t and G_{t+1} i.e.: $\operatorname{argmin}_\delta G_{t+1} = \operatorname{argmax}_\delta(G_t - G_{t+1})$. Let us denote $\widehat{\mathbf{y}}_t$ as the estimates of output values when the number of training samples is equal to t; and $\widehat{\mathbf{y}}_{t+1}$ as the estimates of output values after the value of y_δ was obtained and added to the training set. Let us rewrite the difference between generalization errors G_t and G_{t+1} (also referred to as $\triangle G$) in terms of a difference between $\widehat{\mathbf{y}}_t$ and $\widehat{\mathbf{y}}_{t+1}$: $\triangle G = G_t - G_{t+1} = \left\|\widehat{\mathbf{y}}_t - \widehat{\mathbf{y}}_{t+1}\right\|^2 + 2\left\langle \widehat{\mathbf{y}}_{t+1} - \widehat{\mathbf{y}}_t, \mathbf{y} - \widehat{\mathbf{y}}_{t+1}\right\rangle$. Let us denote the first term as $T_1 = \left\|\widehat{\mathbf{y}}_t - \widehat{\mathbf{y}}_{t+1}\right\|^2$, and the second term as $T_2 = 2\left\langle \widehat{\mathbf{y}}_{t+1} - \widehat{\mathbf{y}}_t, \mathbf{y} - \widehat{\mathbf{y}}_{t+1}\right\rangle$. Note that this decomposition is different from the standard bias-variance decomposition.

Estimating the value of term T_2 relies on the estimate of $\mathbf{y}$. In the current settings, the number of training samples is small, so the estimate of $\mathbf{y}$ is likely to be unreliable. However, estimating the value of term T_1 requires only the estimate of a single value $y^*_\delta \in \mathbf{y}$, so overall the estimate of T_1 is less likely to be error-prone than that of T_2.

Let us investigate if T_1 alone is a good predictor of $\triangle G$. Let us consider three possible cases of the location of $\widehat{y}_{t+1}$ (an element of $\widehat{\mathbf{y}}_{t+1}$) in relation to the corresponding elements $\widehat{y}_t$ and y, as illustrated in Figure 4. In case (b), adding a training point improves the estimate of the true output value. In this case, maximizing T_1 also maximizes $\triangle G$. In case (a), adding a training point deteriorates the estimate of the true output value. In case (c), adding a training point causes the estimate to overshoot the true output value. In both cases (a) and (c) maximizing T_1 does not maximize $\triangle G$. In Figure 3, we show the distribution of the location of $\widehat{y}_{t+1}$ relative to $\widehat{y}_t$ and y (plotted from the data from the numerical experiment described in Section 4). Case (b) is much more frequent than cases (a) and (c). Even when cases (a) and (c) do occur, the probability of the output estimate significantly deteriorating is low. Since T_1 is less prone to error and is more likely to be applicable, we use it as an estimator of $\triangle G$ and utilize it to define the proposed active learning criterion as:

$$\triangle \widehat{G}_{Proposed}(\widehat{f}_{\mathcal{T}\cup(\mathbf{x}_\delta, y_\delta)}) = \mathcal{L}(\widehat{\mathbf{y}}_t, \widehat{\mathbf{y}}_{t+1}) = \sum_{\mathbf{x}\in\mathcal{X}} \mathcal{L}(\widehat{f}_{\mathcal{T}}(\mathbf{x}), \widehat{f}_{\mathcal{T}\cup(\mathbf{x}_\delta, y_\delta)}(\mathbf{x}))\,.$$

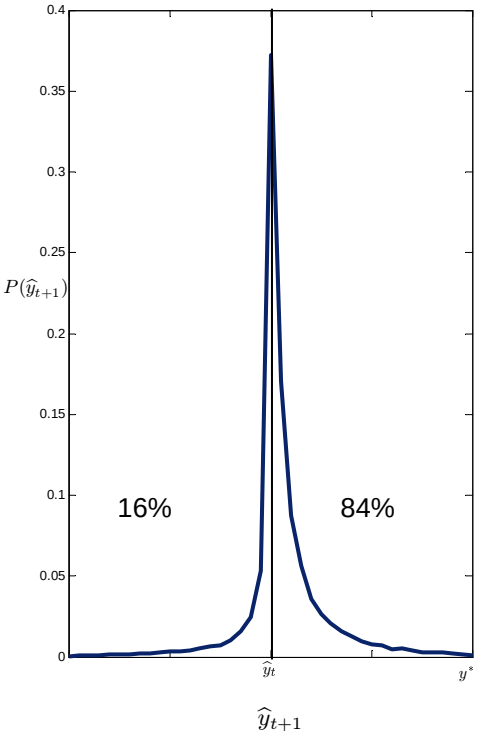


Figure 3: Distribution of the estimates $\widehat{y}_{t+1}$ in relation to the estimate $\widehat{y}_t$ and the true value y.

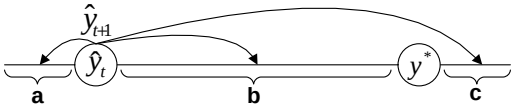


Figure 4: $\widehat{y}$ after the training point δ is added to the training set (making the number of training points equal to $t+1$).

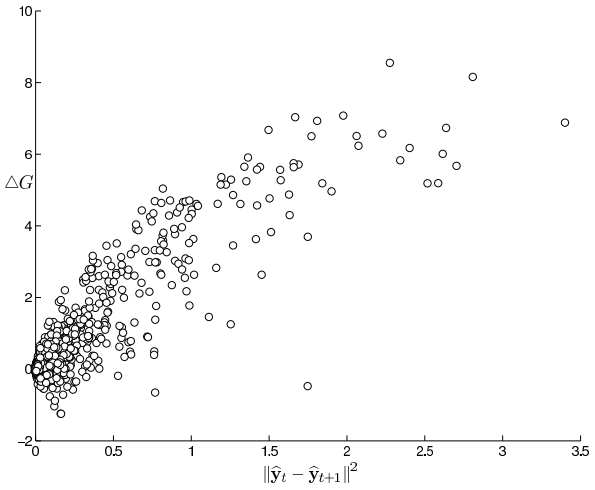


Figure 5: $T_1 = \|\widehat{\mathbf{y}}_t - \widehat{\mathbf{y}}_{t+1}\|$ and the value that it tries to approximate $\triangle G$ (Section 3.1). Most importantly, high values of $\|\widehat{\mathbf{y}}_t - \widehat{\mathbf{y}}_{t+1}\|^2$ should correspond to high values of $\triangle G$, since those are the points that are likely to be chosen.

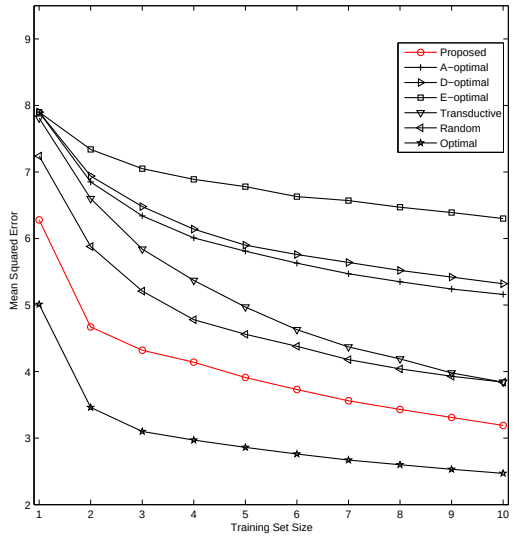


Figure 6: Evaluation of active learning criterions (Mean Squared Error: lower values are better, values are different at the statistical significance level of 95%).

Algorithm 1 Pseudocode of Proposed Method.

$\triangle\widehat{G}$ estimates changes in predictive error that labeling an item $\mathbf{x}_\delta$ would allow to achieve function $\triangle\widehat{G}(\widehat{f}_{\mathcal{T}\cup(\mathbf{x}_\delta,y_\delta)})$
  # learn a preference approximation function $\widehat{f}$ based on the current training set $\mathcal{T}$
  $\widehat{f}_{\mathcal{T}}$ =learn($\mathcal{T}$)
    # for each possible output of an item $\mathbf{x}_\delta$ e.g. $\{1, 2, \ldots, 5\}$
    for y_δ in $\mathcal{Y}$
      # add a hypothetical training point $(\mathbf{x}_\delta, y_\delta)$
      $\mathcal{T}^{(\delta)} = \mathcal{T} \cup (\mathbf{x}_\delta, y_\delta)$
      # learn a new approximation function $\widehat{f}$ based on the new training set $\mathcal{T}^{(\delta)}$
      $\widehat{f}_{\mathcal{T}^{(\delta)}}$ =learn($\mathcal{T}^{(\delta)}$)
      # for each unlabeled point
      for $\mathbf{x}$ in $\mathbf{X}^*$
        # record the differences between outputs estimates
        # before and after a hypothetical training point $(\mathbf{x}_\delta, y_\delta)$ was added to the training set $\mathcal{T}$
        $\triangle\widehat{G} = \triangle\widehat{G} - \left(\widehat{f}_{\mathcal{T}}(\mathbf{x}) - \widehat{f}_{\mathcal{T}^{(\delta)}}(\mathbf{x})\right)^2$
  return $\triangle\widehat{G}$

We are not able to calculate the actual value of the above criterion since the output value y_δ is not known. However, we can approximate criterion by obtaining its expected value as: $\triangle\widehat{G}_{Proposed}(\mathbf{x}_\delta) \approx \sum_{\mathbf{x}\in\mathcal{X}} \mathbb{E}_{y\in\mathcal{Y}}\mathcal{L}(\widehat{f}_{\mathcal{T}}(\mathbf{x}), \widehat{f}_{\mathcal{T}\cup(\mathbf{x}_\delta,y)}(\mathbf{x}))$. By assuming no prior knowledge about the label of the candidate point, we can represent it by a non-informative uniform distribution. Utilizing the mean squared loss function the above criterion could be written as:

$$\triangle\widehat{G}_{Proposed}(\mathbf{x}_\delta) \approx \sum_{\mathbf{x}\in\mathcal{X}} \frac{1}{|\mathcal{Y}|} \sum_{y\in\mathcal{Y}} \left(\widehat{f}_{\mathcal{T}}(\mathbf{x}) - \widehat{f}_{\mathcal{T}\cup(\mathbf{x}_\delta,y)}(\mathbf{x})\right)^2,$$

where $1/|\mathcal{Y}|$ is a normalizing constant since we consider all equiprobable outputs $y \in \mathcal{Y}$ of an item $\mathbf{x}_\delta$. The point to be labeled is then selected as $\operatorname{argmax}_{\mathbf{x}\in\mathcal{X}} \triangle\widehat{G}_{Proposed}(\widehat{f}_{\mathcal{T}\cup(\mathbf{x}_\delta,y)})$. The effectiveness of the proposed criterion is evaluated empirically in Figure 5.

3.2 Considerations

Outlier Robustness Outlier is a point that is numerically distant from the rest of the data. In current settings, were the number of training samples is assumed to be very small, selecting an outlier for labeling provides little benefit (limited amount of information, and potentially negative effect on the accuracy of the learned function). Many AL methods select points to label based on their uncertainty, and therefore tend to select outliers due to inherent uncertainty of their labels. Proposed method is only slightly affected by this, since uncertainty component is much smaller than the influence component (e.g. labeling less certain, but more influential point (b), is by far preferable to rating an uncertain outlier point). If an outlier has a strong effect on the learned function resulting in changes of many estimates, proposed active learning method will be affected strongly, and will then tend to favor outliers. It is therefore recommended to use learning methods that

are robust to outliers; or identify and remove outliers from a set of candidate points for labeling (Andersen, 2008).

Both the proposed active learning criterion and many of the outlier detection methods (Hodge and Austin, 2004; Cook, 1977) tend to select points that are expected to exert large influence on the output estimates. While outliers are considered detrimental to the accuracy, proposed criterion considers them useful; which seems contradictory, but is not. Problem settings have a strong effect on whether influential point is useful or not. In the outlier detection settings, it is often assumed that a lot of labeled data has already been acquired and an outlier is a point that does not fit well with the rest of the data and the underlying patterns which have been learned, and therefore may exert large potentially detrimental influence on the output estimates. On the other hand, in active learning settings, the number of labeled data points is very small and only a few underlying predictive patterns have been yet discovered. Hence an influential point may allow to discover a new pattern which could be used to predict labels of many not yet labeled points; and is therefore influential.

3.3 Relation with Variance-based Active Learning

Let us show that the proposed criterion could be interpreted in traditional active learning settings and its advantages. In traditional active learning settings predictive error is decomposed into bias and variance components. Typical approach is to assume that the bias component becomes negligible (e.g. by assuming that asymptotically unbiased methods are used), and to concentrate on minimize the variance component of the error by utilizing various properties of information matrix (Boyd and Vandenberghe, 2004). Limitations of the traditional active learning methods are as follows. Since the main objective is to minimize the variance of model's parameters, these AL methods are not applicable to model free approaches, or may not be practical for models where number of parameters is very large. In addition, reducing variance of model's parameters does not necessarily result in significant reduction of variance of output estimates, e.g. where many input attributes are missing or are zero (which occurs frequently in many domains). These methods are especially ineffective in the early stages of learning (when the number of training samples is very small), which is often the most practically important stage (since if the model appears to be inaccurate, it may not be possible to justify obtaining more data for it). This is caused by unreliability of parameter variance estimate that is dependent on number training data (which is very small in the early stage). In addition, in the early stage, bias component is often much larger than variance; therefore focusing on reducing the variance may not be effective. The proposed active learning method addresses these limitations by aiming at directly improving output estimates (rather than improving model's parameter which may have little effect on achieving the objective of improving output estimates), considers both bias and variance error components, and is applicable to any of the learning models.

As to compare the proposed criterion with variance-based AL methods let us formulate the proposed criterion in the linear regression settings as: $J(\delta) = \|\widehat{\mathbf{y}}_t - \widehat{\mathbf{y}}_{t+1}\|^2 = \left\|\mathbf{X}^* \left(\widehat{\boldsymbol{\beta}}_t - \widehat{\boldsymbol{\beta}}_{t+1}\right)\right\|^2$, where $\widehat{\boldsymbol{\beta}}_t$, $\widehat{\boldsymbol{\beta}}_{t+1}$ are the least-squares estimators of the parameter

values. Let $\mathbf{A} = \mathbf{X}^\top\mathbf{X} + \alpha\mathbf{I}$, where $\alpha\mathbf{I}$ is a regularization parameter (where $0 < \alpha \ll 1$ ensures that matrix $\mathbf{A}$ is invertible). By using the Woodbury expansion (Hager, 1989) we express the difference between the parameter estimates as

$$\widehat{\boldsymbol{\beta}}_{t+1} - \widehat{\boldsymbol{\beta}}_t = \mathbf{A}^{-1}\mathbf{x}_\delta y_\delta - \frac{\mathbf{A}^{-1}\mathbf{x}_\delta\mathbf{x}_\delta^\top\mathbf{A}^{-1}\mathbf{x}_\delta y_\delta}{1+\mathbf{x}_\delta^\top\mathbf{A}^{-1}\mathbf{x}_\delta} = \frac{\mathbf{A}^{-1}\mathbf{x}_\delta(y_\delta - \mathbf{x}_\delta^\top\widehat{\boldsymbol{\beta}}_t)}{1+\mathbf{x}_\delta^\top\mathbf{A}^{-1}\mathbf{x}_\delta}\,.$$

The difference between the output estimates could now be expressed as

$$\widehat{\mathbf{y}}_{t+1} - \widehat{\mathbf{y}}_t = \mathbf{X}^*\frac{\mathbf{A}^{-1}\mathbf{x}_\delta(y_\delta - \mathbf{x}_\delta^\top\widehat{\boldsymbol{\beta}}_t)}{1+\mathbf{x}_\delta^\top\mathbf{A}^{-1}\mathbf{x}_\delta}\,.$$

The proposed criterion is then formulated as

$$\triangle\widehat{G} = \|\widehat{\mathbf{y}}_{t+1} - \widehat{\mathbf{y}}_t\|^2 = \left(\frac{y_\delta - \mathbf{x}_\delta^\top\widehat{\boldsymbol{\beta}}_t}{1+\mathbf{x}_\delta^\top\mathbf{A}^{-1}\mathbf{x}_\delta}\right)^2 \mathbf{x}_\delta^\top\mathbf{A}^{-1}\mathbf{X}^{*\top}\mathbf{X}^*\mathbf{A}^{-1}\mathbf{x}_\delta\,.$$

3.3.1 Interpretation

Let us look at a possible interpretation of the proposed criterion in linear regression settings and its relation to existing active learning criterions. Let us rewrite the criterion as

$$\triangle\widehat{G} = (y_\delta - \mathbf{x}_\delta^\top\widehat{\boldsymbol{\beta}}_t)^2\frac{\mathbf{x}_\delta^\top\mathbf{A}^{-1}\mathbf{X}^{*\top}\mathbf{X}^*\mathbf{A}^{-1}\mathbf{x}_\delta}{(1+\mathbf{x}_\delta^\top\mathbf{A}^{-1}\mathbf{x}_\delta)^2} = J_R\frac{J_S}{J_P}\,,$$

where $J_R = (y_\delta - \mathbf{x}_\delta^\top\widehat{\boldsymbol{\beta}}_t)^2$, $J_S = \mathbf{x}_\delta^\top\mathbf{A}^{-1}\mathbf{X}^{*\top}\mathbf{X}^*\mathbf{A}^{-1}\mathbf{x}_\delta$, $J_P = (1+\mathbf{x}_\delta^\top\mathbf{A}^{-1}\mathbf{x}_\delta)^2$.

The term $J_R = (y_\delta - \mathbf{x}_\delta^\top\widehat{\boldsymbol{\beta}}_t)^2$ represents the residual value, i.e. the squared error between the actual output value y_δ and its estimate $\mathbf{x}_\delta^\top\widehat{\boldsymbol{\beta}}_t$. The $\mathbf{x}_\delta$ with larger residual value is then favored by the term J_R. Taking the residual value into account corresponds to the residual-based active learning methods (Romano and Kinnaert, 2005).

The next term J_S can be rewritten further as $J_S = \sum_{\mathbf{x}_t\in\boldsymbol{X}^*}\left(\mathbf{x}_\delta^\top\mathbf{A}^{-1}\mathbf{x}_t\right)^2$. By noticing that $\mathbf{x}_\delta \in \boldsymbol{X}^*$, we can further rewrite the above term as: $J_S = (\mathbf{x}_\delta^\top\mathbf{A}^{-1}\mathbf{x}_\delta)^2 + \sum_{\mathbf{x}_t\in\boldsymbol{X}^*\setminus\mathbf{x}_\delta}\left(\mathbf{x}_\delta^\top\mathbf{A}^{-1}\mathbf{x}_t\right)^2$. The part J_S/J_P of the proposed criterion could now be rewritten as

$$\begin{aligned}\frac{J_S}{J_P} &= \frac{(\mathbf{x}_\delta^\top\mathbf{A}^{-1}\mathbf{x}_\delta)^2 + \sum\limits_{\mathbf{x}_t\in\boldsymbol{X}^*\setminus\mathbf{x}_\delta}\left(\mathbf{x}_\delta^\top\mathbf{A}^{-1}\mathbf{x}_t\right)^2}{(1+\mathbf{x}_\delta^\top\mathbf{A}^{-1}\mathbf{x}_\delta)^2} \\ &= \frac{(\mathbf{x}_\delta^\top\mathbf{A}^{-1}\mathbf{x}_\delta)^2}{(1+\mathbf{x}_\delta^\top\mathbf{A}^{-1}\mathbf{x}_\delta)^2} + \frac{\sum\limits_{\mathbf{x}_t\in\boldsymbol{X}^*\setminus\mathbf{x}_\delta}\left(\mathbf{x}_\delta^\top\mathbf{A}^{-1}\mathbf{x}_t\right)^2}{(1+\mathbf{x}_\delta^\top\mathbf{A}^{-1}\mathbf{x}_\delta)^2} = J_O + J_T\,.\end{aligned}$$

In order to interpret the meaning of the terms J_O and J_T, let us eigen-decompose the matrix $\mathbf{X}^\top\mathbf{X}$ into its eigenvalues and eigenvectors as $\mathbf{X}^\top\mathbf{X} = \sum_{i=1}^p \lambda_i\boldsymbol{\varphi}_i\boldsymbol{\varphi}_i^\top$, where λ_i are eigenvalues such that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m > \lambda_{m+1} = \cdots = \lambda_d = 0$ and associated eigenvectors $\boldsymbol{\varphi}_i$, and m is the rank of the matrix $\mathbf{X}^\top\mathbf{X}$. We can also rewrite the matrix $\mathbf{A}$ as $\mathbf{A} = \mathbf{X}^\top\mathbf{X} + \alpha\mathbf{I} = \sum_{i=1}^p \lambda_i\boldsymbol{\varphi}_i\boldsymbol{\varphi}_i^\top + \alpha\mathbf{I}$.

Let us examine the conditions under which the value of J_O increases. Let $\mathbf{x}_\delta^\top \mathbf{A}^{-1}\mathbf{x}_\delta = a$, then we can rewrite J_O as $J_O = a^2/(a+1)^2$. The value of J_O is non-negative and is monotone increasing with respect to a. So let us examine under which conditions the value of a is large. We can rewrite a as

$$a = \sum_{i=1}^{p} \left(\mathbf{x}_\delta^\top \boldsymbol{\varphi}_i\right)^2 \frac{1}{\lambda_i + \alpha} = \sum_{i=1}^{m} \left(\mathbf{x}_\delta^\top \boldsymbol{\varphi}_i\right)^2 \frac{1}{\lambda_i + \alpha} + \frac{1}{\alpha} \sum_{i=m+1}^{p} \left(\mathbf{x}_\delta^\top \boldsymbol{\varphi}_i\right)^2 .$$

Since α is set to a value close to zero, the $\frac{1}{\alpha}\sum_{i=m+1}^{p}\left(\mathbf{x}_\delta^\top \boldsymbol{\varphi}_i\right)^2$ part dominates in the above equation. We may then approximate a as $a \approx \frac{1}{\alpha}\sum_{i=m+1}^{p}\left(\mathbf{x}_\delta^\top \boldsymbol{\varphi}_i\right)^2$. The value of $\frac{1}{\alpha}\sum_{i=m+1}^{p}\left(\mathbf{x}_\delta^\top \boldsymbol{\varphi}_i\right)^2$ is large when we choose $\mathbf{x}_\delta$ that belongs to the null space of $\mathbf{X}^\top\mathbf{X}$ spanned by $\{\boldsymbol{\varphi}_i\}_{i=m+1}^{p}$. This is equivalent to $\mathbf{x}_\delta$ being orthogonal to the *training space* (the range of $\mathbf{X}^\top\mathbf{X}$ spanned by $\{\boldsymbol{\varphi}_i\}_{i=1}^{m}$). So the J_O part of the criterion favors $\mathbf{x}_\delta$ that is "not close" to the training space. This would be related to variance-based AL methods (John and Draper, 1975; Chan, 1981; Dette and Studden, 1993; Sugiyama and Ogawa, 2000).

Let us examine the conditions under which the value of J_T increases. First, let us simplify the formulation of the J_T. By using the fact that the denominator $(1+\mathbf{x}_\delta^\top \mathbf{A}^{-1}\mathbf{x}_\delta)^2 \geq 1$, we may obtain the lower bound of J_T as $J_T \geq \sum_{\mathbf{x}_t \in \boldsymbol{X}^* \setminus \mathbf{x}_\delta}\left(\mathbf{x}_\delta^\top \mathbf{A}^{-1}\mathbf{x}_t\right)^2$. The $\mathbf{x}_\delta^\top \mathbf{A}^{-1}\mathbf{x}_t$ part of the above equation can be rewritten as

$$\begin{aligned}\mathbf{x}_\delta^\top \mathbf{A}^{-1}\mathbf{x}_t &= \sum_{i=1}^{p} \left(\mathbf{x}_\delta^\top \boldsymbol{\varphi}_i\right)\left(\mathbf{x}_t^\top \boldsymbol{\varphi}_i\right) \frac{1}{\lambda_i + \alpha} \\ &= \sum_{i=1}^{m} \left(\mathbf{x}_\delta^\top \boldsymbol{\varphi}_i\right)\left(\mathbf{x}_t^\top \boldsymbol{\varphi}_i\right) \frac{1}{\lambda_i + \alpha} + \frac{1}{\alpha} \sum_{i=m+1}^{p} \left(\mathbf{x}_\delta^\top \boldsymbol{\varphi}_i\right)\left(\mathbf{x}_t^\top \boldsymbol{\varphi}_i\right) .\end{aligned}$$

Since α is set to a value close to zero, the $\frac{1}{\alpha}\sum_{i=m+1}^{p}\left(\mathbf{x}_\delta^\top \boldsymbol{\varphi}_i\right)\left(\mathbf{x}_t^\top \boldsymbol{\varphi}_i\right)$ part dominates in the above equation. We may then approximate the above equation as

$$\mathbf{x}_\delta^\top \mathbf{A}^{-1}\mathbf{x}_t \approx \frac{1}{\alpha} \sum_{i=m+1}^{p} \left(\mathbf{x}_\delta^\top \boldsymbol{\varphi}_i\right)\left(\mathbf{x}_t^\top \boldsymbol{\varphi}_i\right) ,$$

and may now approximate the lower bound of the term J_T as

$$J_T \geq \sum_{\mathbf{x}_t \in \boldsymbol{X}^* \setminus \mathbf{x}_\delta} \left(\mathbf{x}_\delta^\top \mathbf{A}^{-1}\mathbf{x}_t\right)^2 \approx \sum_{\mathbf{x}_t \in \boldsymbol{X}^* \setminus \mathbf{x}_\delta} \left(\frac{1}{\alpha} \sum_{i=m+1}^{d} \left(\mathbf{x}_\delta^\top \boldsymbol{\varphi}_i\right)\left(\mathbf{x}_t^\top \boldsymbol{\varphi}_i\right) \right)^2 .$$

The term J_T favors $\mathbf{x}_\delta$ whose projection onto the null space of $\mathbf{X}^\top\mathbf{X}$ is "close" to the projections of the vectors $\boldsymbol{X}^* \setminus \mathbf{x}_\delta$ onto the null space. Taking the test points into account is related to the transductive active learning method (Yu, Bi, and Tresp, 2006).

4 Experimental Settings

Let us describe settings of the experiments described in Section 3.1. We have selected the MovieLens dataset (Riedl and Konstan, 1998) for the numerical experiments. The

MovieLens dataset consists of approximately 1 million ratings (outputs) for 3900 movies by 6040 users (treated as features/attributes). We randomly select 100 users that have each rated at least 100 items. For each user, we randomly select 50 points (items) as potential training points and use the rest of the points as a test set. All of the users' output values (outputs) are withheld. For each user, training points are selected in a sequential manner by an active learning algorithm. For the random active learning method, training points are selection following the uniform distribution. After the training point is selected, its output value is revealed and the point is added to the training set. Linear regression is used as underlying learning model. To emphasize the effectiveness of the proposed active learning method, we compare it with traditional active learning methods that utilize knowledge about the underlying model, while the proposed method does not.

5 Conclusion

Black box settings are very common in practice, but have not been explicitly addressed in the domain of active learning. By not explicitly considering these settings, existing methods tend suffer from a number of limitations (Section 3.3). The proposed active learning approach is designed specifically for black box settings. It utilizes only on the estimates of output values, available from any learning method, which in turn provides significant advantages for practical deployment.

References

Andersen, R. (2008). *Modern methods for robust regression* (No. 152). Thousand Oaks, CA, USA: Sage Publications.

Bell, R. M., and Koren, Y. (2007). Lessons from the netflix prize challenge. *SIGKDD Explorations Newsletter*, *9*, 75–79.

Boyd, S., and Vandenberghe, L. (2004). *Convex optimization*. Cambridge: Cambridge University Press.

Chan, N. (1981). *A-optimality for regression designs* (Tech. Rep.). Stanford, CA, USA: Stanford University, Department of Statistics.

Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics*, *19*(1), 15–18.

Dette, H., and Studden, W. J. (1993). Geometry of e-optimality. *The Annals of Statistics*, *21*(1), 416-443.

Hager, W. (1989). Updating the inverse of a matrix. *SIAM review*, *31*(2), 221–239.

Hodge, V., and Austin, J. (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review*, *22*(2), 85–126.

John, R. C. S., and Draper, N. R. (1975, Feb.). D-optimality for regression designs: A review. *Technometrics*, *17*(1), 15-23.

Riedl, J., and Konstan, J. (1998). *Movielens data set.* http://movielens.umn.edu.

Romano, D., and Kinnaert, M. (2005). An experiment-based methodology for robust design of optimal residual generators. In *IEEE conference on decision and control* (p. 6286 - 6291). Seville, Spain: IEEE.

Rubens, N., Kaplan, D., and Sugiyama, M. (2010). Recommender systems handbook. In (chap. Active Learning for Recommender Systems). New York, NY: Springer.

Settles, B. (2009). *Active learning literature survey* (Computer Sciences Technical Report No. 1648). Madison, Wisconsin, USA: University of Wisconsin–Madison.

Sugiyama, M., and Ogawa, H. (2000). Incremental active learning for optimal generalization. *Neural Computation*, *12*(12), 2909–2940.

Yu, K., Bi, J., and Tresp, V. (2006). Active learning via transductive experimental design. In *Proceedings of the 23rd int. conference on machine learning icml '06* (pp. 1081–1088). New York, NY, USA: ACM.

Authors' addresses:

Neil Rubens
Department of Information Systems
University of Electro-Communications
1-5-1 Chofugaoka, Chofu-City, Tokyo, 182-8585
Japan

E-mail: rubens@hrstc.org

Vera Sheinman
Japanese Institute of Educational Measurement
3-2-4 Aoyama, Minato-ku Tokyo, 107-0061
Japan

E-mail: vera46@gmail.com

Ryota Tomioka
Graduate School of Information Science and Technology
University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656
Japan

E-mail: tomioka@mist.i.u-tokyo.ac.jp

Masashi Sugiyama
Graduate School of Information Science and Engineering
Tokyo Institute of Technology
2-12-1-W8-74, O-okayama, Meguro-ku, Tokyo, 152-8552
Japan

E-mail: sugiyama@cs.titech.ac.jp