A Model-Independent Measure of Regression Difficulty

Bin Zhang, Charles Elkan,
Umeshwar Dayal, Meichun Hsu
Software Technology Laboratory
HP Laboratories Palo Alto

HPL-2000-5
January, 2000

E-mail:bzhang@hpl.hp.com

data mining, machine learning, model fitting, regression, exploratory data analysis, error rate estimation, data modeling, data cleaning, data preparation, predictability

We prove an inequality bound for the variance of the error of a regression function plus its non-smoothness as quantified by the Uniform Lipschitz condition. The coefficients in the inequality are calculated based on training data with no assumptions about how the regression function is learned. This inequality, called the Unpredictability Inequality, allows us to evaluate the difficulty of the regression problem for a given dataset, before applying any regression method. The Inequality gives information on the tradeoff between prediction error and how sensitive predictions must be to predictor values. The Unpredictability Inequality can be applied to any convex subregion of the space ***X*** of predictors. We improve the effectiveness of the Inequality by partitioning ***X*** into multiple convex subregions via clustering, and then applying the Inequality on each subregion. Experimental results on genuine data from a manufacturing line show that, combined with clustering, the Unpredictability Inequality provides considerable insight and help in selecting a regression method.

1. Introduction

Given target function $T = \{(\boldsymbol{x},y) | \boldsymbol{x} \in \boldsymbol{X} \text{ and } y \in R\}$ where $\boldsymbol{X}$ is a compact region in R^d for some d and y is a scalar, a prediction is a function $F{:}\boldsymbol{X} \rightarrow R$ such that

$$y = F(x) + \varepsilon_x$$

where ε_x is the prediction error at $\boldsymbol{x}$. The prediction function F is unbiased at $\boldsymbol{x}$ if $E(\varepsilon_x | \boldsymbol{x}) = 0$. In this paper, we assume that F is unbiased for all $\boldsymbol{x}$.

Given a training set of data, the prediction function F could be drawn from many different classes of functions (e.g. polynomials or neural networks) and could be discovered (i.e. learned) by many different algorithms (e.g. least-squares regression or backpropagation). We make no assumption on what algorithm is used, or what function family F belongs to, except that F is continuous.

All widely used prediction methods are based on a smoothness assumption about the data (or system) to be modeled: that when $\boldsymbol{x}$ and $\boldsymbol{x'}$ are close, the corresponding y and y' are close. When such smoothness does not exist in some regions of $\boldsymbol{X}$, successful prediction is not possible there. For certain type of regressions, multiple linear regression as an example, the effect could go beyond these regions. Testing the smoothness of data first, helps the modeler in setting reasonable expectations about predictive accuracy and the minimum complexity of the model – the degree of a polynomial or the number of inner nodes of a neural network model as examples.

Intuitively, the smoothness of a target function can be measured by dividing $\boldsymbol{X}$ into small regions such that the variance of $\boldsymbol{x}$ in each region is small. Then if the y values in some regions still vary a lot, the target function is not smooth in those regions. For such data, either the values of *F* must vary sharply to match the variation of *y*, or the prediction error must have a big variance in those regions. This relationship is captured in the following inequality:

$$\sigma^2(\varepsilon_x) + L^2 * \sigma^2(x) \geq \sigma^2(y) \qquad (1)$$

Here *L* is the coefficient in the Lipschitz condition on *F*. This inequality holds in any convex region of ***X*.** The first term on the left is the prediction variance. $\sigma^2(\boldsymbol{x})$ and $\sigma^2(y)$ are treated as coefficients because they are estimated from the available training data. If $\sigma^2(y)$ is large compared with $\sigma^2(\boldsymbol{x})$, the inequality says that either *L* or $\sigma^2(\in_x)$ has to be large. We call this property *unpredictability*. If we consider $(L,\sigma^2(\varepsilon_x))$ as a point in the first quadrant of a 2D coordinate system, the inequality excludes the point from a region (see Figure 2 in Section 5) surrounding the point (0,0), which represents the easiest problem – the variable *y* to be predicted is a constant and there is no noise in the given training data.

When the target function is noise-free, the sensitivity of a predicted value, *y*, to the predictors, ***x***, is given by the gradient of *F*. The smaller the gradient, the less sensitive *y* is to the variation of ***x***. When the gradient is big, small changes of ***x*** will cause big changes of *y*, such a prediction function is not desirable in a noisy environment. The Lipschitz condition gives a bound on the gradient of *F* when the gradient exists.

Since the difficulty of a problem is invariance under scaling of the variables, we also present the scaling-invariant version of the inequality (1), which uses the (squared) coefficients of standard deviation instead of the variance, and “elasticity” of a function instead of its gradient.

The remainder of this paper is organized as follows. In Section 1(a) related research is reviewed. Then in Section 2 we briefly review the definitions related to the Lipschitz conditions, and in Section 3 we derive the Unpredictability Inequality. Sections 4 discuss the invariance of the inequality under scaling of data. Section 5 gives a geometric view of the Inequality. In Sections 6, the Inequality is combined with data clustering techniques (*k*-means in particular) for higher effectiveness. Section 7 gives a visualization of the results from large number of clusters. Section 8 presents a real-world example of how the Inequality can be applied to high-dimensionality data. Section 9 concludes the paper.

1(a). Related Research

The aim of our work is to estimate the best achievable error rate for a regression problem, given a fixed model complexity, where the complexity of a model is defined to be its lack of smoothness, quantified as a high Lipschitz coefficient *L*.

There is a range of related research, but we are not aware of any other work that attempts to estimate the best achievable error rate as a function of a measure of model complexity.

For a classification problem (as opposed to a regression problem), the minimum error rate that *any* classifier can achieve is called the Bayes' error rate. The Bayes' error rate is a measure of the degree of randomness in a categorical target function, rather than a measure of its lack of smoothness.

For a review of several methods for estimating the Bayes' error rate from a training dataset, see [TG99]. The most useful methods in practice for estimating Bayes' error rates are based on nearest neighbor classification, and reviewed in [RH95]

Extending the ideas of this paper from regression to classification is an area for future work, as is exploring the connection between our ideas and existing methods for estimating Bayes' error rates. Intuitively, a connection does exist. The nearest neighbor Bayes' error estimation procedure says essentially that a target function is predictable over a small subregion of the space $\boldsymbol{X}$ if and only if all training points in the subregion have the same label y. The target function is intrinsically unpredictable if training points that are nearest neighbors to each other have different labels.

Our work can be viewed as a test of best possible goodness of fit. Therefore, it only has a weak connection to traditional methods for testing goodness of fit, which evaluate the goodness of fit of a specific regression model. For a review of goodness of fit tests developed in the last century, see D'Agostino and Stephens 1986, and Rayner and Best 1989.

Even non-parametric tests of goodness of fit, such the X^2 test, assume that a fixed model has been chosen. One can imagine trying to estimate the intrinsic difficulty of fitting a dataset by combining the results of X^2 tests of many different models, combining p-values obtained from many different tests on the same data is a difficult mathematical topic [see Benjamini's web site "http://www.math.tau.ac.il/~ benja/"].

Another perspective on our work is that it is an attempt to assess the intrinsic quality of a dataset. Some other work addresses similar issues but in quite different ways. In particular, the aim of outlier detection methods is to identify data points that are poor examples of all the classes in a classification problem. These methods do not solve the problem that we address in this paper. Often a target function is non-smooth, but none of its ($\boldsymbol{x}$,y) points can be called an outlier. For example, missing variables in X may cause the non-smoothness of the function. If $y = F(x_1, x_2, x_3)$ is a (deterministic) function of three variables, but only the first two variables are included in $\boldsymbol{X}$, a finite training dataset $D =\{(x_1, x_2,, y)|\ y = F(x_1, x_2,,*)\}$, formed by randomly sampling (x_1, x_2, x_3), is likely to be non-smooth everywhere.

2. Lipschitz Conditions

Lipschitz condition is a commonly known condition in *The Theory of Functions.* The definitions and basic properties of Lipschitz conditions are given in this section without proof, more details can be found in [P82] [RN55] (or other books on *Functional Analysis*).

The function class that satisfies Lipschitz condition is very large. For example, all differentiable functions is a subset of Lipschitz functions, which includes linear functions and the non-linear functions generated by neural networks with finite number of nodes, or generated by many other popular modeling tools.

There are two versions of Lipschitz Condition: 1) Point-wise Lipschitz Condition and 2) Uniform Lipschitz Condition.

Point-wise Lipschitz Condition: Function F satisfies the Point-wise Lipschitz Condition at $\boldsymbol{x_0}$ if there exists a δ>0 and a constant L such that for any $\boldsymbol{x}$ satisfying $||\mathbf{x}- \boldsymbol{x_0}||<\delta$, the following holds

$$|F(x)-F(x_0)|\le L\,\|x-x_0\|.$$

It is “Point-wise” because δ and L depend on $\boldsymbol{x_0}$**.** Even when F satisfies the point-wise Lipschitz Condition at every point, there may not be a common pair of δ and L for all points. The Uniform Lipschitz Condition given next is a sronger condition on F, which fixes the problem.

If there is a common pair of δ and L for all $\boldsymbol{x}$ in a compact region, the point-wise Lipschitz condition implies the following.

Uniform Lipschitz Condition**:** Function F satisfies the Uniform Lipschitz Condition if there exists a δ>0 and a constant L such that for any $\boldsymbol{x}$ and $\boldsymbol{x'}$ satisfying $\|\mathbf{x}-\boldsymbol{x'}\|<\delta$, the following holds

$$|F(x)-F(x')|\le L\,\|x-x'\|. \tag{2}$$

The requirement, $\|\mathbf{x}-\boldsymbol{x'}\|<\delta$, is not necessary in a convex region and can be dropped because for any pair of points, $\boldsymbol{x}$ and $\boldsymbol{x'}$, the line segment between them, which falls within the region, can be equally divided into a finite number of segments $<\boldsymbol{x},\boldsymbol{x_1}>,<\boldsymbol{x_1},\boldsymbol{x_2}>,\ldots\ldots,<\boldsymbol{x_n},\boldsymbol{x'}>$, each shorter than δ. Applying Lipschitz condition on each one of them, we have

$$|F(x)-F(x_1)|\le L\,\|x-x_1\|, |F(x_1)-F(x_2)|\le L\,\|x_1-x_2\|,\ldots, |F(x_n)-F(x')|\le L\,\|x_n-x'\|.$$

Adding them up, we have

$$\begin{aligned}&|F(x)-F(x')|\le|F(x)-F(x_1)|+|F(x_1)-F(x_2)|+\ldots+|F(x_n)-F(x')|\\&\le L[\|x-x_1\|+\|x_1-x_2\|+\ldots+\|x_n-x'\|]\le L^*\|x-x'\|\end{aligned}$$

The last step follows the fact that all (n+2) points are sequentially ordered on a straight line.

Using the same argument backwards, we can show that if Uniform Lipschitz condition is violated by a pair of points, $\boldsymbol{x}$**,** $\boldsymbol{x'}$**,** in a convex region, then it is violated by a pair of points arbitrarily close (to prove: divide the line segment <***x*,*x*'**>). If F is differentiable, the length of the gradient of F is greater than L somewhere along the line segment <***x*,*x*'**>.

We will use only Uniform Lipschitz condition on convex sets from now on. The word "Uniform" will be dropped.

3. Unpredictability Inequality

We treat $\boldsymbol{X}$ as a probability space. The probability corresponds to the distribution of $\boldsymbol{x}$ in the data. We do not make any assumption on the distribution because the calculations of the coefficients in the Inequality is based on real data (non-parametric).

We separate the integration over the probability space $\boldsymbol{X}$ and the probability spaces of the random noise terms ϵ_x when it is more comprehensible. Figure 1 gives an illustration of the whole probability space for the case dimension($\boldsymbol{X}$) =2. Each vertical dash line gives a probability space for a particular ε_x .

Let $\boldsymbol{S} \subseteq \boldsymbol{X}$ be a convex compact region (closure of an open subset) in $\boldsymbol{X}$. The inequality *(1)* is derived over $\boldsymbol{S}$ under the conditional probability $p(\boldsymbol{A}|\boldsymbol{S})=p(\boldsymbol{A})/p(\boldsymbol{S})$ where $\boldsymbol{A}\subseteq\boldsymbol{S}$. We use $E_{\boldsymbol{S}}$ for calculating the mean over $\boldsymbol{S}$, and E_ϵ for calculating the mean over a random noise at any one point $\boldsymbol{x}$. Similar notations, $\sigma_{\boldsymbol{S}}()$ and $\sigma_\epsilon()$, are used for the variances. $E()$ and $\sigma()$ without any subscript means calculating the average over the compound space.

Since continuity (or discontinuity) of a function is an infinitesimal property, testing the “continuity” of data has to be carried out in small regions of ***x*** variables. This is achieved by combining with clustering techniques to derive inequality (1) in multiple regions that aligned with the natural clustering in ***X***. The details are presented in a later section.

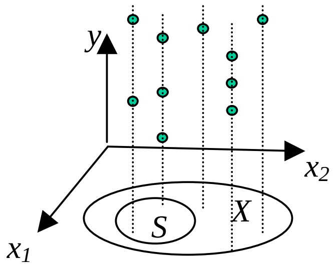


Figure 1. The Probability Space.

To derive the Unpredictability Inequality, we calculate the variance of *y*,

$$\begin{aligned}\sigma^2(y) &= E([y-E(y)]^2) = E([F(x)+\varepsilon_x - E(F(x)+\varepsilon_x)]^2) = E([F(x)-E(F(x))+\varepsilon_x]^2) \\ &= E([F(x)-E(F(x))]^2) + E(\varepsilon_x^2) + 2*E([F(x)-E(F(x))]*\varepsilon_x)\end{aligned} \quad (3)$$

The last term (cross term) in (3) is zero because

$$E([F(x)-E(F(x))]*\varepsilon_x) = E_S([F(x)-E(F(x))]*E_\varepsilon(\varepsilon_x)) = 0.$$

The first term in (3) is the variance of the prediction function. Formula (3) shows that the variance in *y* is split into the variance of the prediction function and the variance of the prediction error.. Using Lipschitz condition on the first term on the right in (3), we get

$$E([F(x)-E(F(x))]^2) \le E([F(x)-F(E(x))]^2) \le L^2 E([x-E(x)]^2) \le L^2\sigma^2(x) \quad (4)$$

The first step in (4) follows the fact that *E([F(x) - z]^2)* reaches its minimum at z=*E(F(x));* and the second step follows the Lipschitz condition.

From (3) and (4), we get the Unpredictability Inequality,

$$\sigma^2(y) \geq L^2 * \sigma^2(x) + E_S(\sigma_\varepsilon^2(\varepsilon_x)) \qquad (5)$$

the last term in (5) and $\sigma^2(\varepsilon_x)$ are both equal to $E(\varepsilon_x{}^2)$ because $E_\in(\varepsilon_x) = 0$ and $E(\varepsilon_x) = 0$. The final form of the Unpredictability Inequality is

$$\sigma^2(y) \geq L^2 * \sigma^2(x) + \sigma^2(\varepsilon_x) \qquad (6)$$

$\sigma^2(\mathbf{x})$ and $\sigma^2(y)$ are estimated from the data using standard techniques [NWK90],

$$\bar{x} = \frac{\sum_{(x,y)\in D} x}{|D|}\ , \quad \bar{y} = \frac{\sum_{(x,y)\in D} y}{|D|}\ , \quad s^2(y) = \frac{\sum_{(x,y)\in D}(y-\bar{y})^2}{|D|-1}\ , \quad s^2(x) = \frac{\sum_{(x,y)\in D}(x-\bar{x})^2}{|D|-1}$$

where $\boldsymbol{D} = \{\ (\mathbf{x},y) \mid \mathbf{x} \in \boldsymbol{S}\ \}$. Using these estimates, the inequality (6) is a quadratic inequality with "unknowns" L and $\sigma^2(\varepsilon_x)$. If the total variance of y is large and the total variance of $\mathbf{x}$ is small, either L has to be large or the average variance of prediction has to be large. The inequality marks a region away from the origin $<L=0,\ \sigma^2(\varepsilon_x)=0>$. This geometric view is given next.

When $\sigma^2(\mathbf{x}) > 0$ (i.e. $\boldsymbol{S}$ is not a single point), another way of looking at (6) is $L^2 \geq [\sigma^2(y) - \sigma^2(\varepsilon_x)]/\ \sigma^2(\mathbf{x})$. The quantity in "[ ]" is always greater or equal to zero.

4. Invariant Version of the Inequality under Scaling of Data – Using Coefficients of Standard Deviation and Elasticity

The difficulty of a prediction problem is invariant under scaling of the variables. Accordingly, the unpredictability inequality should also be invariant. This is achieved by using the (squared) coefficients of standard deviation rather than variance. Intuitively, the absolute prediction error changes under scaling of data but the percentage of error does not. Using the coefficients of standard deviation, everything is put into percentages. To convert the variances in the inequality into the coefficients of standard deviation, we divide both sides of the inequality by $E^2(y)$,

$$\sigma^2(y)/E^2(y) \ge (L*E(x)/E(y))^2*\sigma^2(x)/E^2(x)+\sigma^2(\varepsilon_x)/E^2(y) \tag{6.1}$$

or

$$\theta^2(y) \ge (L*E(x)/E(y))^2*\theta^2(x)+\theta^2(\varepsilon_x) \tag{6.2}$$

where $\theta^2(\mathbf{x}) = \sigma^2(\mathbf{x})/E^2(\mathbf{x})$ and $\theta^2(y) = \sigma^2(y)/E^2(y)$ are the (squared) coefficients of standard deviation of $\mathbf{x}$ and y, $\theta^2(\varepsilon_x)= \sigma^2(\varepsilon_x)/E^2(y)$ measures the variance of the percentage of error and can be called the "coefficient" (percentage) of standard prediction deviation. If we temporarily replace L by the length of the gradient of F (reminder: we used L, the Lipschitz coefficient, to bound the gradient), and calculate the gradient along a special direction of $\delta\mathbf{x}$, then

$$|\Delta F| = \frac{|\delta F|}{|\delta x|}$$

and

$$|\Delta F|*\frac{E(x)}{E(y)} = \frac{|\delta F|}{|\delta x|}\frac{E(x)}{E(y)} = \frac{|\delta F|/E(y)}{|\delta x|/E(x)}$$

which remind us of the elasticity of F, defined as

$$\frac{|\delta F|/F(x)}{|\delta x|/|x|}.$$

The elasticity of a function gives the percentage of change of its value over the percentage of change of its parameters. In the inequality (6.2), all quantities are put into "percentages", which is invariant under re-scaling of the variables $\mathbf{x}$ and y; and is usually more informative than the absolute values. To put (6.2) in a similar form as (6), we use L_θ for $L*E(\mathbf{x})/E(y)$. (6.2) becomes,

$$\theta^2(y) \ge L_\theta^2*\theta^2(x)+\theta^2(\varepsilon_x) \tag{6.3}$$

This is the Unpredictability Inequality invariant under scaling of data (or variables).

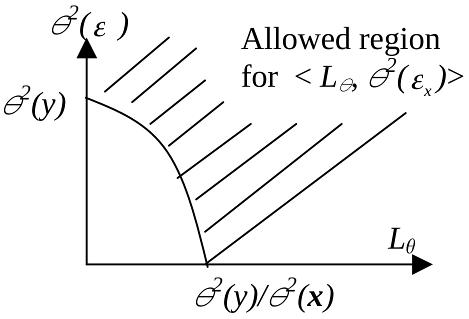
5. Geometric View of the Unpredictability Inequality

The boundary of the region defined by the inequality (6.3) is a parabola curve plotted in Figure 2, in which $\theta^2(\varepsilon_x)$ is considered as a function of L_θ,

$$\theta^2(y) \geq L_\theta^2 * \theta^2(x) + \theta^2(\varepsilon_x) .$$

L_θ and $\theta^2(\varepsilon_x)$ can not both be small because they (together as a point) are kept away from the origin by the parabolic curve. The intersection of the curve with the vertical axis is at $\theta^2(y)$ and the intersection with the horizontal axis at $\theta^2(y)/\theta^2(x)$.

In the next section, we show that combining with clustering techniques, the space ***X*** is divided into many small convex regions, the center of which are aligned with the local density of data in ***X*,** the unpredictability inequality is derived over all of these convex regions and a collective view of the smoothness of data can be built.



$\theta^2(\varepsilon\ )$
Allowed region
for $< L_\theta, \theta^2(\varepsilon_x)>$
$\theta^2(y)$
L_θ
$\theta^2(y)/\theta^2(\mathbf{x})$

Figure 2. L_θ and $\theta^2(\varepsilon_x)$ pair is kept away from the origin by a parabolic curve.

They can not both be small. This is called unpredictability.

6. Combining with Data Clustering

The Unpredictability Inequality was presented in previous sections over a (small) convex region ***S*** of ***X*** to detect discontinuity of data in that region. The purpose of the inequality is to detect discontinuity in data: a small (percentage of) change in ***x*** causing a big (percentage of) change in *y*. Since continuity or discontinuity of a function is a infinitesimal property, we really intend to let the diameter of ***S*** approach zero. But a given data set in practice is much more likely to be a discrete set, the diameter of ***S*** can not be too small. For example, if ***S*** is so small that there is only one data point in ***S***, we will not be able to tell anything about the quality of data near that point. The Nearest Neighbor method [RH95] chooses ***S*** to have two points.

It is desirable to have a collective view of the discontinuity of data over all regions of ***X***. These are achieved by partitioning ***X*** into multiple convex regions (that are aligned with the local densities in ***X***). It is known that center based clustering algorithms (like *K*-Means [M67][GG91] or other center based clustering algorithms [MK97][ZH99]) partitioning the space into convex regions (Voronoi partition). The clusters tend to have much smaller variances than an arbitrary partition.

Let ***X*** = *{* ***S_k*** | *k = 1,……,K }* be a partition of ***X*** from clustering and ***M*** = *{ m_k* | *k = 1,……,K }* the centroids of the clusters, which we do not use directly. Using (6), the Unpredictability Inequality on each cluster is,

$$\sigma^2_{S_k}(\varepsilon_x) + L_k^2\sigma^2_{S_k}(x) \geq \sigma^2_{S_k}(y). \qquad (7)$$

When the number of clusters (convex regions) is large, aggregated views of the inequalities results are more comprehensible. From (7), many different versions of aggregation could be derived. Multiplying both sides of (7) by the probability *p(S_k)* and sum over *k*, we have

$$\sum_{k=1}^{K} p(S_k)\sigma^2_{S_k}(\varepsilon_x) + \sum_{k=1}^{K} p(S_k)L_k^2\sigma^2_{S_k}(x) \geq \sum_{k=1}^{K} p(S_k)\sigma^2_{S_k}(y). \qquad (8)$$

From (8), two versions of aggregation are:

Unpredictability Inequality A: Replacing L_k^2 in (8) by $max_k(L_k^2)$, we have

$$\sigma^2(\varepsilon_x) + \max_k(L_k^2) * W_M^2(x) \geq W_M^2(y), \qquad (9)$$

where $W_M()$ is the with-in cluster variance [DH72] of the clusters defined by center-set ***M***. This view gives a stronger bound but only on the *maximum* of *L*.

Unpredictability Inequality B: Replacing $\sigma_{Xk}^2(\mathbf{x})$ by $max_k(\sigma_{Xk}^2(\mathbf{x}))$, we have another view (10) that gives a bound on the average of L_k but it may not be as strong because $max(\sigma_{Xk}^2(\mathbf{x})) \geq W_M^2(x)$.

$$\sigma^2(\varepsilon_x) + [\sum_{k=1}^{K} p(S_k) L_k^2] * \max_k(\sigma_{S_k}^2(x)) \geq W_M^2(y). \qquad (10)$$

(9) and (10*)* are different versions of Unpredictability Inequality after combining with clustering techniques. They provide different information on the difficulty of the prediction problem.

To achieve invariance under scaling, the coefficient of standard deviation is used in the previous section. When the Inequality is applied to multiple regions (clusters), invariance can be achieved by dividing the variances in each cluster by the square of the global means of *y*, $E_X(y)$, instead of its own cluster's mean value, because only the global scaling of data (all clusters share the same scaling coefficients) is concerned. Dividing by the global mean tend not to discriminate the errors from different clusters. Dividing by each clusters own mean tend to discriminate against the errors from the clusters with a smaller mean *y* value and be more tolerant to the errors from clusters with larger mean *y* value. After dividing by the global mean of *y* (and ***x***), the scaling-invariant form of (9) and (10) are:

$$\sigma^2(\varepsilon_x)/E^2(y) + \max_k(L_k^2 * E^2(x)/E^2(y)) * W_M^2(x)/E^2(x) \geq W_M^2(y)/E^2(y), \qquad (9^*)$$

and

$$\sigma^2(\varepsilon_x)/E^2(y)+[(\sum_{k=1}^{K} p(S_k)L_k^2)*E^2(x)/E^2(y)]*\max_k(\sigma_{S_k}^2(x)/E^2(x)) \geq W_M^2(y)/E^2(y). \quad (10^*)$$

When the partitions is very fine, for example, down to individual points in ***X***, the inequality degenerates to the following equality: $\sigma^2(y|x)$ = $\sigma^2(\varepsilon_x|x)$ at each ***x***. Another method of aggregation is to plot the results from all clusters in one XY-plot.

7. Using XY-Plots on Pairs $<L\ ,\ \sigma^2(\varepsilon_x)>$

Some information is lost due to aggregation and the "*max()*" operation in (9) and (10). As another option of summarizing the results from all clusters, plotting all results,

$$<\sigma_{S_k}^2(y)/\sigma_{S_k}^2(x),\sigma_{S_k}^2(y)> \qquad (11)$$

for *k=1,…,K,* in an XY-plot provides bird's-eye view. When the resolution of the plot is good enough (not too crowded) for counting the dots in the plot, this method gives more information than the aggregations in the previous section.

The *k*th point in (11) uniquely determines the parabola (See Figure 3.) for the *k*th cluster. Instead of drawing the whole parabola, a single point is sufficient to represent it.

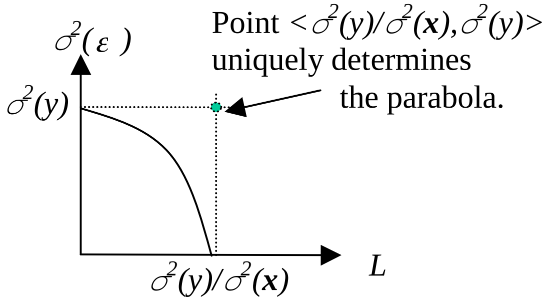

Point $<\sigma^2(y)/\sigma^2(\mathbf{x}),\sigma^2(y)>$
uniquely determines
the parabola.
$\sigma^2(\varepsilon\ )$
$\sigma^2(y)$
L
$\sigma^2(y)/\sigma^2(\mathbf{x})$

Figure 3. One point in the plane uniquely determines the parabola.

Two cases we like to explain more:

Case 1: $\sigma_{Xk}^{2}(y) >> \sigma_{Xk}^{2}(y)/ \sigma_{Xk}^{2}(\boldsymbol{x})$. It is not conclusive in this situation because $\sigma_{Xk}^{2}(\boldsymbol{x})$ is still large. Further partition of the *k*th cluster is recommended. $\sigma_{Xk}^{4}(y)/\sigma_{Xk}^{2}(\boldsymbol{x})$ can be used as an indicator to further subdivide clusters (by voronoi partitions).

Case 2: $\sigma_{Xk}^{2}(y) << \sigma_{Xk}^{2}(y)/ \sigma_{Xk}^{2}(\boldsymbol{x})$. Local (in a small neighborhood of ***x***) sharp changes of *y* are found. This shows local ripples (including the extreme case that many different values of *y* in the data correspond to the same vector ***x***).

The XY-plot applies to the scaling-invariant version of the Inequality also.

8. An Example

This example is from a real modeling problem of a manufacturing line. It is also the problem that motivated this work.. There are 12 quality control variables, ***y***, and 40 monitoring variables, ***X***. All of them have a large number of values recorded. We want to predict the 12 quality control variables from the 40 monitoring variables. If the prediction is good, quality control can be done in earlier in the production at the monitoring locations, which means savings. We regressed (linear and non-linear) the 12 variables on the 40 predictors with very limited results, which lead to the model-independent examination of data. About *150,000* data points are available (size of ***X***).

$$\sigma^2(x) = 256.84, \qquad W_M^2(x) = 0.55, \qquad \max_k(\sigma_{X_k}^2(x)) \approx 7.802$$

The Unpredictability Inequalities are calculated on both the whole data set and partitioned data sets. The *k*-means algorithm is directed to find 100 clusters and *77* of them turned out to be non-empty. The following quantities are estimated from the data. The variances of the whole data set and the with-in cluster variances of the partitioned data set for the 12 different *y* variables are listed in Table 1.

From Table 1, Unpredictability Inequality A is most effective for this data set. Inequality B is the 2nd and the global inequality on the whole ***X*** is the least. It is identified by the Inequality that the 5th, 6th,7th,10th, and 11th *y*-variables are not predictable in certain regions. Comparing the 2nd and the 3rd columns in the Table, the within cluster variance of *y* decreased much less than that of *x*, which decreased by a factor of 500, except for the 8th variable, which turned out to be very easy to predict (confirmed from multiple linear regression).

	Global Inequality without clustering		With clustering, 77 clusters are aggregated		
y#	σ^2(y)	σ^2(y)/ σ^2(x)	$W_M^2(y)$	Inequality A: $W_M^2(y)/W_M^2(x)$	Inequality B: $W_M^2(y)/max(\sigma_{Sk}^2(x))$
1	2.09	0.008	1.63	2.97	0.209
2	0.89	0.003	0.56	1.02	0.072
3	5.66	0.022	4.81	8.77	0.617
4	1.91	0.007	1.77	3.22	0.226
5	67.97	0.265	25.03	45.63	3.208
6	13.61	0.053	12.06	21.99	1.546
7	9.57	0.037	9.04	16.48	1.159
8	119.1	0.464	1.44	2.62	0.185
9	5.94	0.023	4.04	7.36	0.517
10	63.28	0.246	55.57	101.3	7.123
11	76.22	0.297	72.79	132.7	9.329
12	0.02	0.000	0.02	0.04	0.003

Table 1. The coefficients for Unpredictability Inequality (6), Unpredictability Inequalities A (9) and B (10).

Figure 4 has the XY-Plots for the first twelve *y*-variables in the Examples. Due to the large spread in the values of $\sigma_{Xk}^2(y)/\sigma_{Xk}^{\ 2}(\mathbf{x})$, $log_{10}(\sigma_{Xk}^{\ 2}(y)/\sigma_{Xk}^{\ 2}(\mathbf{x}))$ is plotted as the horizontal coordinate and $\sigma_{Xk}^2(y)$ as the vertical coordinate. The results from 77 clusters are plotted as 77 points, each represent a parabola, in the plots. The disadvantage of using an XY-Plot here is that the clusters are not necessarily equal in probability, or in size or diameter).

9. Conclusion

We developed an inequality to access the Unpredictability of data independent of the regression model that are going to be used. With such an up-front model independent exploration of the data, we will have reasonable expectations, confidence and be more focused during the modeling phrase of the data mining tasks. This inequality can also be used to test the residue errors from regression to see if further improvement of prediction is possible.

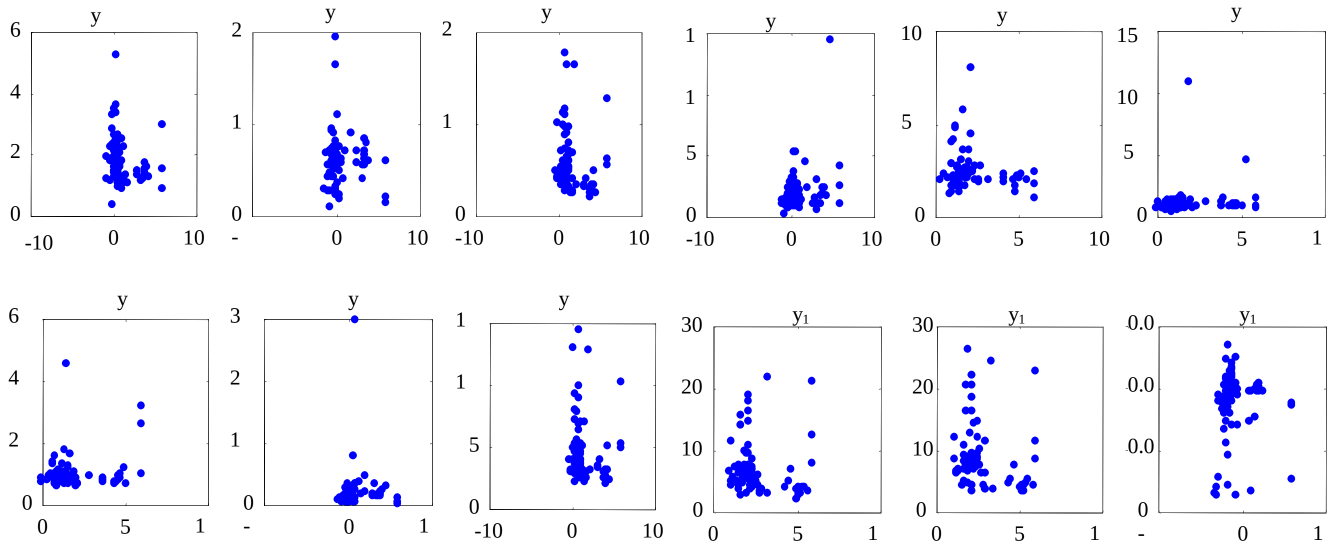


Figure 5. Using XY-Plots on Pairs <$\boldsymbol{log_{10}(\sigma_{Xk}^2(y)/\sigma_{Xk}^2(x)),\sigma_{Xk}^2(y)}$>.

References

[DS86] D'Agostino, R. B. and Stephens, M. A., "Goodness-of-fit Techniques.", New York: Marcel Dekker, 1986.

[DH72] Duda, R., Hart, P., "Pattern Classification and Scene Analysis", John Wiley & Sons, 1972.

[GG91] Gersho, A. and Gray, R, *Vector Quantization and Signal Compression*, Kluwer Academic Publishers 1991.

[K99] Kanji, G.K., "100 Statistical Tests", SAGE, 1999.

[M67] MacQueen, J. 1967. Some methods for classification and analysis of multivariate observations. Pp. 281-297 *in*: L. M. Le Cam & J. Neyman [eds.] Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, Vol. 1. University of California Press, Berkeley. xvii + 666 p.

[MK97] McLachlan, G. J. and Krishnan, T., "The EM Algorithm and Extensions.", John Wiley & Sons, Inc., 1997

[NWK90] Neter, J., Wasserman, W., Kutner, M. H., "Applied Linear Statistical Models", 1990.

[P82] Parzynski, W. R., "Introduction to Mathematical Analysis", 1982.

[RH95] Brian D. Ripley and N. L. Hjort, "Pattern Recognition and Neural Networks," Cambridge Univ Press, 1995.

[RB89] Rayner, J.C.W. and Best, D.J., Smooth Test of Goodness of Fit", Oxford University Press, 1989.

[RN55] Riesz, F. and Nagy, B. SZ., "Lecons D'Analyse Fonctionnelle", Budapest 1955, 3rd Edition.

[TG99] Kagan Tumer and Joydeep Ghosh, "A Mutual Information based ensemble method to estimate Bayes error,".