Regression Error Characteristic Curves

Jinbo Bi BIJ2@RPI.EDU
Kristin P. Bennett BENNEK@RPI.EDU
Department of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, NY 12180 USA

Abstract

Receiver Operating Characteristic (ROC) curves provide a powerful tool for visualizing and comparing classification results. Regression Error Characteristic (REC) curves generalize ROC curves to regression. REC curves plot the error tolerance on the x-axis versus the percentage of points predicted within the tolerance on the y-axis. The resulting curve estimates the cumulative distribution function of the error. The REC curve visually presents commonly-used statistics. The area-over-the-curve (AOC) is a biased estimate of the expected error. The R^2 value can be estimated using the ratio of the AOC for a given model to the AOC for the null model. Users can quickly assess the relative merits of many regression functions by examining the relative position of their REC curves. The shape of the curve reveals additional information that can be used to guide modeling.

1. Introduction

Receiver Operating Characteristic (ROC) curves have proven to be a valuable way to evaluate the quality of a discriminant function for classification problems (Egan, 1975; Swets et al., 2000; Fawcett, 2003). ROC curves address many of the limitations of comparing algorithms based on a single misclassification cost measure (Provost et al., 1998). An ROC curve characterizes the performance of a binary classification model across all possible trade-offs between the false negative and false positive classification rates. An ROC graph allows the performance of multiple classification functions to be visualized and compared simultaneously. ROC curves can be used to evaluate both expected accuracy and variance information. ROC curves are consistent for a given problem even if the distribution of positive and negative instances is highly skewed. The area under the ROC curve (AUC) represents the expected performance as a single scalar. The AUC has a known statistical meaning: it is equivalent to the Wilcoxon test of ranks. Fundamentals of interpreting ROC curves are easily grasped. ROC curves are effective tools for visualizing results for non-experts as well as experts and help them make more valid conclusions. For example, a non-expert can see that two functions have similar ROC curves and can conclude that there is no significant difference between the functions even though one may have a larger classification cost. Currently ROC curves are limited to classification problems.

The goal of this paper is to devise a methodology for regression problems with similar benefits to those of ROC curves. Our solution, the Regression Error Characteristic (REC) curve, plots the error tolerance on the x-axis versus the percentage of points predicted within the tolerance on the y-axis. The resulting curve estimates the cumulative distribution function (CDF) of the *error*. The *error* here is defined as the difference between the predicted value $f(\mathbf{x})$ and actual value y of response for any point $(\mathbf{x}, y)$. It could be the squared residual $(y - f(\mathbf{x}))^2$ or absolute deviation $|y - f(\mathbf{x})|$ depending on the error metric employed. Figure 1 provides an example of REC curves generated for industrial data. See Section 3 for more details.

REC curves behave much like ROC curves.

- REC curves facilitate visual comparison of regression functions with each other and the null model.
- The curve area provides a valid measure of the expected performance of the regression model. The REC curve estimates CDF of the error. The area **over** the curve (AOC) is a biased estimate of the expected error.
- The REC curve is largely qualitatively invariant to choices of error metrics and scaling of the resid-

Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington DC, 2003.

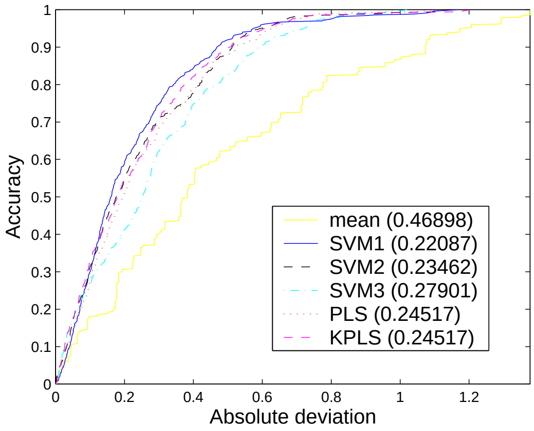

Figure 1. REC curve comparing results of five models with the null (mean) model on real-world data.

ual. Scaling the response does not change the graph other than the labeling of the x-axis. Using various error metrics, such as the absolute deviation or squared error, does change the REC curve, but the relative position of curves does not change qualitatively. A function that dominates another using the squared error will also dominate the alternative using the absolute deviation.

- REC curves provide an effective strategy for presenting results to non-experts. One can readily see when regression functions are alike and when they are quite different. In our industrial consulting practice, REC curves provide a much more compelling presentation of regression results than alternatives such as tables of mean squared errors.

- The information represented in REC curves can be used to guide the modeling process based on the goals of the modeler. For example, good choices of ϵ for the ϵ-insensitive loss function in the SVM regression can be found using the curves.

This paper is organized as follows. We begin with a brief review of ROC curves. In Section 3, we define what an REC curve is and give an example of how it is interpreted based on real world data. We examine the area-over-the-curve statistic in Section 4 and show how it provides a biased estimate of the expectation of the error. In the next two sections, we explore the properties of REC curves and their potential use by investigating curves for synthetic data with known characteristics. We conclude with a summary and future work in the last section.

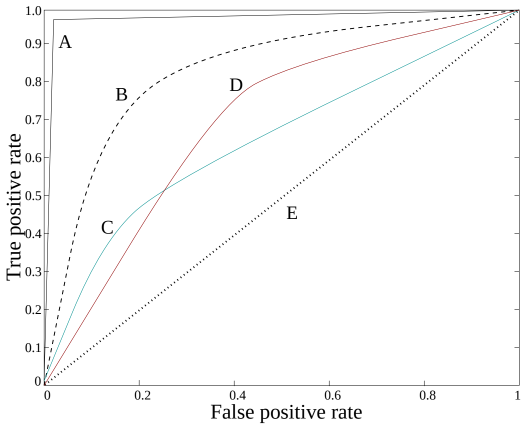

Figure 2. Sample ROC curves: A = almost perfect model. E = null model.

2. Review of ROC

We briefly review the properties of ROC curves. Readers should consult (Fawcett, 2003) for an excellent tutorial on ROC graphs. For two-class discrimination problems, the ROC curve for a discriminant function is constructed by varying the threshold or probability used to discriminate between classes for that function. The resulting pairs of false positive rates and true positive rates are plotted on the x and y axes respectively with lines interpolating between them.

Figure 2 illustrates ROC curves plotted for several classification models. A classifier performs well if the ROC curve climbs rapidly towards the upper left-hand corner. Function A in Figure 2 is an almost perfect classifier. The expected performance of a classifier can be characterized by the area under the ROC curve (AUC). The AUC for a perfect classifier is 1. Random guessing would yield the diagonal line (labelled as E) with the AUC equal to 0.5. No valid classifier should have an AUC below 0.5. A classifier is said to dominate an alternative classifier if the corresponding curve is always above that for the alternative. In Figure 2, function B dominates functions C and D, and thus would be preferable. Function B is better than functions C and D for any choice of the cost function. Correspondingly, the AUC for function B is larger than for functions C and D. Function D does not dominate function C, but it has a higher AUC and is preferable overall. However, function C may be preferable if a low false positive rate is desirable. Note that error bars can also be included in ROC curves to indicate variance information. We save presentation of ROC and REC curves with error bars for future work.

Our goal is to develop a curve that analogously characterizes the quality of regression models and maintains the benefits of ROC curves. In regression, the analogous concept to classification error is the residual $y - f(\mathbf{x})$. Like error rates in classification, existing measures of residuals such as mean squared error, mean absolute deviation, R^2 and Q^2, provide only a single snapshot of the performance of the regression model. In addition, performance metrics based on the absolute or squared residuals may produce different rankings of regression functions. Scatter plots of the predicted value versus the actual value allow users to visualize function performance across the range of data, but they are not practical for visualizing the performance of multiple functions simultaneously.

Typically, we prefer a regression function that fits most data within a desirable error tolerance. As in ROC curves, the graph should characterize the quality of the regression model for different levels of error tolerance. The best function depends on the goals of the modeler. We might want to choose a model that fits almost all data with a loose tolerance versus another model with a tight tolerance that suffers catastrophic failures on a few points. Ideally, a single statistic such as AUC should exist to provide a measure of expected performance with well-understood statistical meaning. Preferably this statistic should correspond to an easily assessed geometric quantity such as an area.

3. Definition of REC Curves

Regression Error Characteristic (REC) curves meet these goals. REC curves plot the error tolerance on the x-axis and the *accuracy* of a regression function on the y-axis. Accuracy is defined as the percentage of points that are fit within the tolerance. If we have zero tolerance, only those points that the function fits exactly would be considered accurate. If we choose a tolerance that exceeds the maximum error observed for the model on all of the data, then all points would be considered accurate. Thus there is a clear trade-off between the error tolerance and the accuracy of the regression function. The concept of error tolerance is appealing because most regression data are inherently inaccurate, e.g. due to experimental and measurement errors.

REC curves were initially motivated by the ϵ-insensitive loss function used in SVM regression methods (Vapnik, 1995). The ϵ-insensitive loss ($\max\{0, |y - f(\mathbf{x})| - \epsilon\}$) has proven to be very robust, but there is no a priori way to pick ϵ. The REC curve based on absolute error metric considers an ϵ-insensitive loss for all possible values of ϵ – much the same as ROC curves plot the results for all possible misclassification costs. Further justification for this evaluation approach comes from statistics. The REC curve is an estimate of the CDF of the error. This enables us to estimate the expected errors of various models.

The basic idea of the ϵ-insensitive loss function is that residuals must be greater than a tolerance ϵ before they are considered as errors. Suppose we have a set of m data points, $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)$, where $\mathbf{x}_i \in R^n$ are the independent variables and $y \in R$ is the dependent or response variable. We define the accuracy at tolerance ϵ as:

$$\mathrm{acc}(\epsilon) := \frac{|\{(\mathbf{x}, y) : \mathrm{loss}(f(\mathbf{x}_i), y_i) \leq \epsilon,\ i = 1, \cdots, m\}|}{m}.$$

When using the squared error metric, the loss is defined as $\mathrm{loss}(f(\mathbf{x}), y) = (f(\mathbf{x}) - y)^2$. When the absolute deviation is adopted, the loss is defined as $\mathrm{loss}(f(\mathbf{x}), y) = |f(\mathbf{x}) - y|$. The REC curve is constructed by plotting ϵ versus $\mathrm{acc}(\epsilon)$. The following algorithm is used to plot an REC curve. Percentile plots can also be used to draw REC curves, which is not discussed in this article.

REC Plot Algorithm
Input: $\epsilon_i = \mathrm{loss}(f(\mathbf{x}_i), y_i),\ i = 1, \ldots, m$. We assume that the errors ϵ_i are sorted in ascending order and the "plot" command interpolates between the plotted points with a line.

```
1. $\epsilon_{prev} := 0$, $correct := 0$;
2. for $i = 1$ to $m$
3.     if $\epsilon_i > \epsilon_{prev}$ then
4.         plot($\epsilon_{prev}$, $correct/m$)
5.         $\epsilon_{prev} := \epsilon_i$
6.     end
7.     $correct := correct + 1$
8. end
9. plot($\epsilon_m$, $correct/m$)
```

As ϵ increases, the accuracy increases. The $\mathrm{acc}(\epsilon)$ eventually goes to 1 when ϵ becomes large enough. A practical problem is how to select the range of ϵ when drawing the REC plots for multiple models. The range is problem-specific even if the data are normalized. The range of ϵ adjusts the appearance of REC curves when we draw curves for different models in one box. For a very wrong model, to achieve accuracy 1, we have to use large ϵ. If the range is selected too large, the REC figure becomes difficult to read since the curves corresponding to better models will cluster in the upper left-hand corner. These are precisely the curves of interest since the corresponding models perform well.

To overcome this difficulty we scale the REC curve using a null model. In this paper the null model is a constant function with the constant equal to the mean of the response of the training data, $f(\mathbf{x}) = \bar{y}$. This mean model is the best constant model assuming Gaussian noise. If the noise follows a Laplacian distribution, a constant model equal to the median of the response of the training data may be preferable. In this paper we only show results for the null model based on the mean. However different null models can be defined if priori knowledge on the noise model is accessible for a problem at hand, and then REC curves can be extended to scale with different null models.

Reasonable regression approaches produce regression models that are better than the null model. An REC graph looks like a rectangle box that contains several monotonically increasing curves (REC curves) each corresponding to a regression model. The x-axis of the box usually starts with 0 since $\epsilon \geq 0$. We define the other end of the x-axis as the largest value of the errors, ϵ_i, obtained by the mean model on the sample data, denoted by $\hat{\epsilon}$. If on the sample data, a given model achieves accuracy of 1 for $\epsilon \leq \hat{\epsilon}$, then the full REC curve is plotted within the box for that model. If the smallest value of ϵ where the model achieves accuracy 1 is greater than $\hat{\epsilon}$, then the corresponding REC curve is truncated at $\hat{\epsilon}$. Hence it is possible that an REC curve may not reach accuracy 1 if the model performs worse than the mean model at high errors, which is a sign of a poor model. Since we focus on the analysis of good curves, we always plot from 0 to $\hat{\epsilon}$.

Figure 1 illustrates the REC curves plotted for five different models and the null model for a proprietary industrial problem. The numbers in the legend give AOC values for each model. We present the results based on 50 random partitions of data into 90% training and 10% test. The results for the 50 trials are averaged by the simple strategy of including the results for every test point from every trial in the REC curve. This method is commonly used in ROC curves. However any alternative method like error bars for presenting cross-validated results developed for ROC curves could also be used. See (Fawcett, 2003) for a survey. The models were constructed using classic regression SVM trained with different values of ϵ (Vapnik, 1995), linear partial least squares (PLS) (Wold, 1966), and kernel partial least squares (KPLS) (Rosipal & Trejo, 2001)[1]. All methods produced reasonable results that outperformed the null model. The results show that the SVM1 model performs better than the rest. The KPLS and the SVM2 models perform very similarly. Their MSEs are not identical but there is no real difference in their performance. In this case, SVM1 was hand tuned based on testing information in order to estimate the best possible performance. Model parameters for SVM2 were selected using cross-validation. The KPLS parameter (number of latent variables) was selected by a fixed policy (Rosipal & Trejo, 2001). One could use this picture to argue with the manager that SVM2 and KPLS are doing almost as well as the best possible model SVM1. SVM3 illustrates what happens if a poor choice of the SVM parameters is made. Other experiments illustrated that KPLS was faster and less sensitive to tuning. Based on the REC curves, one might conclude that KPLS is a preferable method on these datasets.

4. Area Over the REC Curve

In ROC curves the area **under** the ROC curve provides an estimate of the expected accuracy. Correspondingly the area **over** the REC curve (AOC) is a measure of the expected error for a regression model. If we think of the error as a random variable, the REC curve is an estimate of the cumulative distribution function of this variable. We now show that calculating the expected value based on this estimate of the error probability function is equivalent to evaluating the area over the curve. We also show how AOC can be used to estimate other common sample or test statistics for regression models.

4.1. Area Calculates Expected Error

For a given model, the error calculated using the squared error or absolute deviation, can be regarded as a random variable $\varepsilon \geq 0$. The cumulative distribution of this random variable is defined as the probability of $\varepsilon \leq \epsilon$. Denote the CDF of the variable ε as $P(\epsilon)$ and the corresponding probability density function as $p(\epsilon)$. The REC curve represents the empirical CDF of ε with the probability P estimated by the frequency $\hat{P}$ for ϵ_i based on the training data. Assume the errors ϵ_i observed on the m training points are sorted in ascending order. Then ϵ_m is the maximum observed error, and the frequency $\hat{P}(\epsilon_i) = i/m$. By the Glivenko-Cantelli lemma, $\hat{P}(\epsilon)$ converges to $P(\epsilon)$ uniformly over all values of ϵ (DeGroot, 1986). We give a rough sketch of the argument about AOC assuming the full REC curve can be plotted in the figure, in other words, $\epsilon_m \leq \hat{\epsilon}$.

The expectation of ϵ is defined as $E(\epsilon) = \int \epsilon p(\epsilon) d\epsilon$. For any maximum observed error $\epsilon_m > 0$, we can break the integral into two parts:

$$E(\epsilon) = \int_0^{\epsilon_m} \epsilon p(\epsilon) d\epsilon + \int_{\epsilon_m}^{\infty} \epsilon p(\epsilon) d\epsilon. \qquad (1)$$

[1] The results were obtained using the Analyze^{TM} software created by Mark Embrechts at Rensselaer Polytechnic Institute and distributed at *www.drugmining.com*. Results are for REC demonstration only so modeling details are not provided.

The sample max ϵ_m converges to infinity or the maximum value of ϵ if it exists, so the second term converges to 0 as the sample size $m \to \infty$. Define $\Delta\epsilon_i = \epsilon_i - \epsilon_{i-1}$ and $\omega_i \in (\epsilon_{i-1}, \epsilon_i)$. Assume $\Delta\epsilon_i \to 0$ as $m \to \infty$. The expected value of ϵ is equal to the Riemann integral

$$E(\epsilon) = \lim_{m\to\infty} \sum_{i=1}^{m} \omega_i p(\omega_i) \Delta\epsilon_i. \tag{2}$$

By the Mean Value Theorem, we can always choose $\omega_i \in (\epsilon_{i-1}, \epsilon_i)$ such that

$$p(\omega_i) = \frac{P(\epsilon_i) - P(\epsilon_{i-1})}{\epsilon_i - \epsilon_{i-1}}.$$

The above equation can thus be rewritten as

$$\begin{aligned} E(\epsilon) &= \lim_{m\to\infty} \sum_{i=1}^{m} \omega_i \frac{P(\epsilon_i) - P(\epsilon_{i-1})}{\epsilon_i - \epsilon_{i-1}} \Delta\epsilon_i \\ &= \lim_{m\to\infty} \sum_{i=1}^{m} \omega_i (P(\epsilon_i) - P(\epsilon_{i-1})). \end{aligned} \tag{3}$$

The probability distribution P at ϵ is unknown, but we can approximate this quantity based on the finite sample. The infinite series (3) can be truncated at m equal to the sample size. But this approximation underestimates $E(\epsilon)$ since it does not estimate the portion of integral (1) corresponding to $\epsilon > \epsilon_m$ for the finite sample. We use the empirical distribution $\hat{P}$ estimated on the sample data to approximate P and take $\omega_i = \epsilon_i$,

$$\begin{aligned} E(\epsilon) &\approx \sum_{i=1}^{m} \epsilon_i (\hat{P}(\epsilon_i) - \hat{P}(\epsilon_{i-1})) \\ &= \epsilon_m \hat{P}(\epsilon_m) + \sum_{i=1}^{m-1} (\epsilon_i - \epsilon_{i+1}) \hat{P}(\epsilon_i) - \epsilon_1 \hat{P}(\epsilon_0) \\ &= \epsilon_m \hat{P}(\epsilon_m) - \left(\sum_{i=1}^{m-1} (\epsilon_{i+1} - \epsilon_i) \hat{P}(\epsilon_i) + \epsilon_1 \hat{P}(\epsilon_0) \right) \end{aligned}$$

where $\epsilon_0 = 0$. The first term computes the area of the entire region in the box corresponding to this REC curve since $\hat{P}(\epsilon_m) = 1$. This is shown in Figure 3 as the area in the left shadowed rectangle. The terms in the parentheses evaluate the area under the curve.[2] Therefore $E(\epsilon)$ can be approximated by the area over the curve (AOC) within the box as shown in Figure 3. This analysis assumes that the entire REC curve can be plotted. If a model has $\epsilon_m > \hat{\epsilon}$, the REC curve will be truncated. The area observed in the plot does not correspond to the full area over the curve. However the AOC statistic can always be calculated using the sample ϵ_i even though they may not be plotted in the figure. Note AOC is a biased estimate since it always underestimates the actual expectation due to the drop of the second term of the integral (1). For good models (better than the mean model in our illustration), the AOC is a reasonable estimate of the expected error. As the amount of data tends to infinity, AOC converges to the expectation of ϵ. Therefore we include AOC information in REC plots. In all REC figures, the numbers in the legend represent the AOC values corresponding to each model.

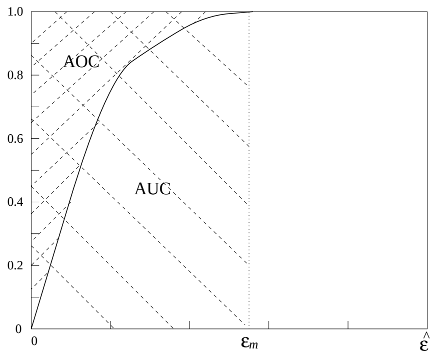

Figure 3. A REC curve with $\epsilon_m \le \hat{\epsilon}$. The area over the curve (AOC) estimates the $E(\epsilon)$.

Table 1. Comparison of AOC estimates and Mean estimates of the expected error $E(\epsilon)$.

	SE		AD	
Model	AOC	MSE	AOC	MAD
SVM1	0.0875	0.0888	0.2209	0.2215
PLS	0.0991	0.1004	0.2452	0.2463
MEAN	0.3548	0.3570	0.4690	0.4704

Certainly there are other more common estimates of $E(\epsilon)$. For example, the sample mean, $\frac{1}{m}\sum_{i=1}^{m} \epsilon_i$, also estimates the expected error. If ϵ is calculated using the absolute deviation (AD), then the AOC is close to the mean absolute deviation (MAD). If ϵ is based on the squared error (SE), the AOC approaches the mean squared error (MSE). Table 1 illustrates the difference between AOC and mean estimates of the SE and AD on the industrial data presented in Figure 1. We can clearly see that the AOC estimate is biased because it always produces a lower estimate of the expected error than the MSE or MAD. The REC curve offers an alternative error estimate that can be visualized directly and that has potential of providing additional information.

[2] We actually calculate the area under the curve by the trapezoidal rule. The same analysis holds but is slightly more complicated for that case.

4.2. Estimating Other Statistics

More statistics can be represented by REC curve graphs. For REC curves with the SE, the AOC for the mean model is an estimate of the variance of the response y since it evaluates the expectation $E[(y-\bar{y})^2]$. Once we have an estimate of the response variance, we can estimate other commonly-used statistics such as Q^2 and R^2. Widely used in chemometrics, Q^2 is the ratio of the MSE over the sample response variance, so Q^2 can also be estimated using AOC by

$$Q^2 = \frac{\sum_{i=1}^{m}(y_i - f(\mathbf{x}_i))^2}{\sum_{i=1}^{m}(y_i - \bar{y})^2} \approx \frac{AOC_{model}}{AOC_{mean}}. \qquad (4)$$

This yields an estimate of R^2 as well because R^2 is often defined as $1-Q^2$. One could also evaluate equivalent statistics defined based on the AD.

The significance of difference between two REC curves can be assessed by examining the maximum deviation between the two curves across all values of ϵ. This corresponds to the Kolmogorov-Smirnov (KS) two-sample test (DeGroot, 1986) for judging the hypothesis that the error ε generated by two models f and g follows the same distribution. The KS-test requires no assumptions about the underlying CDF of ε. Since REC curves represent the sample CDF $\hat{P}(\epsilon)$, the statistics used in the KS-test such as $D^+ = \sup_\epsilon(\hat{P}_f(\epsilon) - \hat{P}_g(\epsilon))$, $D^- = \sup_\epsilon(\hat{P}_g(\epsilon) - \hat{P}_f(\epsilon))$ and $D = \max\{D^+, D^-\}$, can be estimated and visualized by the maximum vertical distance between the two REC curves corresponding respectively to the models f and g.

Hence the REC curve facilitates the visualization of all these statistics simultaneously for many regression functions in a single graph. To deliver the same information, one would have to examine large tables of results and do mental calculations. Moreover, REC curves have the benefit of providing additional information besides all individual statistics.

5. Noise and Loss Models

We explore the relationship between the noise model and error metrics by investigating REC curves on synthetic data with known noise distribution. The goals here are to examine if the model evaluation based on REC curves is sensitive to choices of error metrics, and if REC curves can help identify characteristics of regression models. We draw the REC curves based on both SE and AD to illustrate the effect of error metrics on the appearance of REC curves.

All experiments were conducted using 200 points randomly generated in a 20-dimensional space from a uniform distribution on $[-1,1]^{20}$. The first 10 dimensions were used to construct y and the remaining 10 dimensions are just noise. The goal was to examine how REC curves vary when y is disturbed by additive noise. The response y is constructed via $y = 0.5\sum_{i=1}^{10} x_i + \xi$ where ξ is the additive noise. Several noise models were considered: Gaussian, Laplacian, and Weibull (DeGroot, 1986). To save space, Weibull noise data will be analyzed in the next section not here. Intuitively, the distribution of the residual depends on the regression model f and the noise variable ξ. Figures 4 illustrates REC curves produced for the data with Gaussian and Laplacian additive noise of mean 0 and standard deviation 0.8. Each plot considers four functions: the true model $0.5\sum_{i=1}^{10} x_i$, the mean (null) model, a random model $\sum_{i=1}^{10} w_i x_i$ where w_i are randomly generated from $[0,1]$, and a biased model $0.5\sum_{i=1}^{10} x_i + 1.5$. The AOC values corresponding to each curve are also presented beside the curves or in the legends.

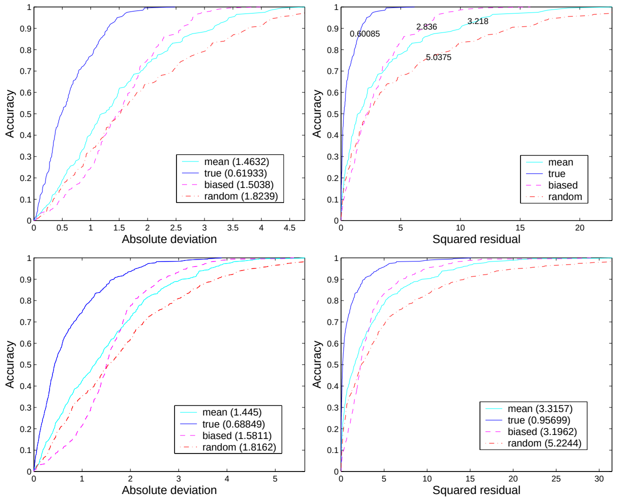

Figure 4. REC curves with Gaussian noise (above) and Laplacian noise (below). Left: AD, right: SE.

As expected, the true model dominates the other models. Similar to ROC curves, an REC curve *dominates* another if it is always above the other. The mean model dominates the random model indicating that the random model is poor. Analogous to the ROC curve, a random model can provide the bottom line for comparison although we use the mean model $\bar{y}$ as the worst case in comparison rather than a random guess. The biased model initially performs the worst, but once the error tolerance is large enough, it outperforms the mean model. Laplacian noise generates more outliers than Gaussian noise. When outliers are present, the top of the REC curve will be flat and not reach 1 until the error tolerance is high. This can be observed in Figure 4(below) for Laplacian noise data.

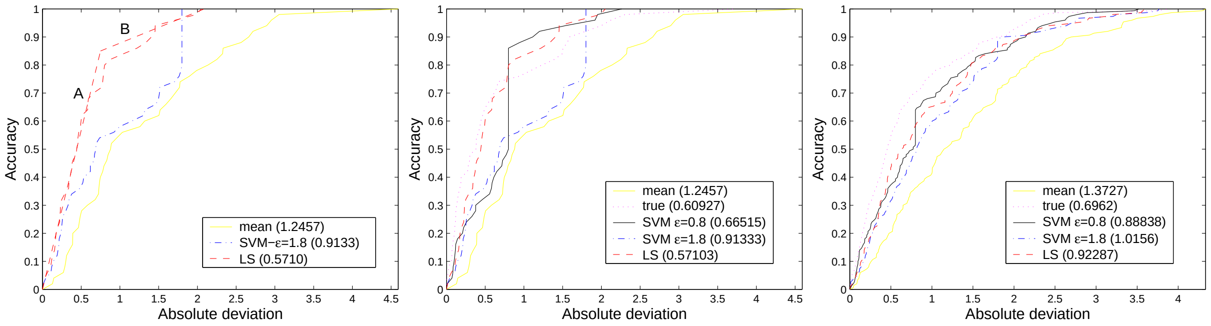

Figure 5. REC curves with Laplacian noise for trained models. *Left:* REC curves based on training data for LS and SVM ($\epsilon = 1.8$) models. *Middle:* Training REC curves including the new SVM model with $\epsilon = 0.8$. *Right:* REC curves generated based on test data for various models.

The curves for the biased model are of particular interest. Note how quickly the curve for the biased model climbs in the middle on contrary to the flat part at the lower end. This characteristic nonconvex behavior of the REC curve is caused by the position of the model relative to data. For example, the regression model is about in the middle of the data, but the data is far from the regression function on both sides. This case is rare in practice. More commonly, this kind of curve results from a biased model for which the data largely lies on one side of the regression model rather than the other side. The biased model will exhibit poor accuracy when the tolerance ϵ is small, but once ϵ exceeds the bias the accuracy rapidly increases. Hence this characteristic shape in the REC curve indicates that the model is likely biased. If the model should be unbiased, this could be a sign of potential problems. In SVMs, a bias can be introduced intentionally via the ϵ-insensitive loss function. As observed in Figure 6, the REC curve for the SVM model ($\epsilon = 0.1$) has a slight nonconvexity at the low end.

The same conclusion would be drawn for each noise model according to REC plots based on both AD and SE. Hence the evaluation of regression models using REC curves is qualitatively invariant to the choices of error metrics. We prefer the AD-based REC curves because the plots tend to be more spaced-out and are convenient to distinguish the trend of different curves. In the next section, we only show AD-based plots.

6. Potential Use of REC Curves

REC curves can help determine an appropriate ϵ value for the ϵ-insensitive loss in SVMs. If the data is largely distributed within a band of width ϵ about the true model with only a few points lying outside, the ϵ-insensitive loss is preferred in modeling. On this data, a good regression model can fit the majority of data within a relatively small tolerance ϵ. We expect a sharp jump in the REC curve around the ϵ value used in the insensitive loss. The curve becomes flat after it hits that ϵ value. Hence we can see an abrupt bend in the trend of the REC curve, especially in SVM training REC curves as shown in Figure 5(middle).

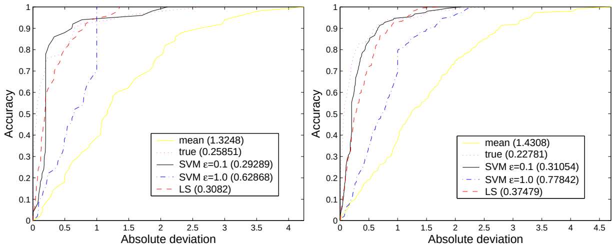

Figure 6. REC curves on Weibull noised data for trained models. Left: training, right: test.

We examine how REC plots could assist in finding a proper ϵ value using synthetic data with Laplacian noise. Figure 5 presents REC curves for least squares (LS) and SVM models trained using 50 points (left and middle) and tested on 150 points (right). Initially the LS model performed better than SVM with parameter $\epsilon = 1.8$. Inappropriate choices of ϵ in the insensitive loss can hurt the performance of SVM. Figure 5(left) shows that the LS REC curve exhibits a dramatic change of slope at about $\epsilon = 0.8$. To show this we draw two straight lines roughly tangent to the lower part of the curve (the line A) and the high end of the curve (the line B), respectively. The two lines cross at about $\epsilon = 0.8$. The similar behavior can be also observed on the REC curve for SVM ($\epsilon = 1.8$). Based on this information observed on training REC curves, we trained a SVM with $\epsilon = 0.8$ in the loss function.

The new curve for SVM ($\epsilon = 0.8$) in Figure 5(middle) shows a significant enhancement in the SVM training accuracy. The generalization performance was also improved for SVM ($\epsilon = 0.8$) as shown in Figure 5(right). The same trick was also applied to synthetic data generated with Weibull-distributed noise where the sign of the noise was selected using a coin toss. The SVM with $\epsilon = 0.1$ shown in Figure 6 chosen in the above way outperforms the LS model and the other SVM model. We leave automating and validating this heuristic for selecting the ϵ parameter in SVMs to future research.

Single statistics such as MSE and MAD may produce different rankings when comparing two models. In this case, the corresponding two REC curves must cross. Let the best model selected by MSE differ from that of MAD. Suppose the curves based on the AD do not cross, in other words, curve 1 is always above curve 2. Let $\mathbf{r}^1$ and $\mathbf{r}^2$ be the two vectors of absolute residuals corresponding to the two models sorted in ascending order. Since curve 1 dominates curve 2, it implies $\mathbf{r}_i^1 \leq \mathbf{r}_i^2,\ 1 \leq i \leq m$, which in turns implies $(\mathbf{r}_i^1)^2 \leq (\mathbf{r}_i^2)^2$. Thus REC curves based on the SE also do not cross. In addition, we have $\sum \mathbf{r}_i^1 \leq \sum \mathbf{r}_i^2$ and $\sum (\mathbf{r}_i^1)^2 \leq \sum (\mathbf{r}_i^2)^2$, so the MAD and MSE provide the same ranking. By contradiction, the curves must cross if models are ranked differently by MSE and MAD.

If two REC curves cross, how should we rank the two models? One solution is to evaluate the AOC values for the two models. If AOC values based on AD and SE both show preferences for the same model, we can trust this ranking. If the preferences are different, then the MSE and MAD rankings will differ as well. Thus the ranking has to rely on more information revealed by REC visualization. Moreover, there may be no significant difference in the curves. The maximum vertical distance between two REC curves represents the KS-test statistic used to assess the significance of the difference between the two error distributions. We leave investigation of the KS-test in REC curves to future research.

7. Conclusions

We proposed a new technique for evaluation and comparison of regression models. REC curves depict the trade-off between error tolerance versus the accuracy of the functions. The REC curve is an estimate of the error cumulative distribution function. Commonly used measures of the distribution can be estimated using the geometry of the figure. For example the area over the curve is a biased estimate of the expected error. Using REC curves, non-experts can quickly evaluate the relative merits of regression functions without consulting tables or other graphs. Experienced modelers can exploit the additional information contained in the REC curve beyond simple statistics. These properties can be used to help diagnose problems with models and to guide model selection. For example, we use REC curves to select values of the ϵ-insensitive loss function in regression. REC curves can potentially profit from the innovations and variations developed for ROC curves. Different methods for averaging models and incorporating error bars to represent variance can be used. The convex hull of the set of points in the REC space may be useful for constructing ensembles of regression functions.

Acknowledgements

This work was supported by NSF grant IIS-9979860. Many thanks to the reviewers for their valuable suggestions.

References

DeGroot, M. H. (1986). *Probability and statistics*. MA: Addison-Wesley.

Egan, J. P. (1975). Signal detection theory and ROC analysis. In *Series in cognition and perception*. New York: Scientific Press.

Fawcett, T. (2003). *ROC graphs: Notes and practical considerations for data mining researchers* (Technical Report HPL-2003-4). Hewlett Packard, Palo Alto, CA.

Provost, F., Fawcett, T., & Kohavi, R. (1998). The case against accuracy estimation for comparing induction algorithms. *Proceedings of the Fifteenth International Conference on Machine Learning* (pp. 445–453). San Francisco: Morgan Kaufmann.

Rosipal, R., & Trejo, L. (2001). Kernel partial least squares regression in reproducing kernel hilbert space. *Journal of Machine Learning Research*, *2*, 97–123.

Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Better decisions through science. *Scientific America*, *283*, 82–87.

Vapnik, V. N. (1995). *The nature of statistical learning theory*. New York: Springer.

Wold, H. (1966). Estimation of principle components and related models by iterative least squares. *Multivariate Analysis* (pp. 391–420). New York: Academic Press.