PROJECT FOR 9.520:
FEATURE SELECTION ANALYSIS

Nathan Eagle

MIT Media Lab
nathan@media.mit.edu

ABSTRACT

Several underlying behaviors of feature selection techniques are analyzed in this paper. A bound relating sample size and dimensionality is derived and verified empirically. The 's-curve' relationship between test error and amount of training data is shown not to be a generalized behavior of all feature selection techniques.

1. INTRODUCTION

Feature selection techniques have recently proven themselves invaluable in many statistical learning domains. Pearson correlation coefficients, Fisher criterion scores, and the Kolmogorov-Smirnov test are methods useful for filtering large amounts of irrelevant data down to only relevant features. Other, more complex, feature selection techniques have also been developed recently for use with SVM classifiers.

Problems in fields, such as bioinformatics, involve a large number of irrelevant dimensions. Attempting to classify some types of highly dimensional data sets may be intractable, both computationally and economically. Establishing the few relevant features is a critical component in the classification process. In order to ascertain which features are indeed relevant, a number of "training samples" are required.

This paper is laid out into two main sections. The first section attempts to determine the relationship between the dimensionality of a data set and the number of samples required to select the relevant features. The second section discusses the 's-curve' appearance with certain feature selection techniques and establishes that it is not a generalized property of all feature selection methods.

2. SAMPLE SIZE AND DIMENSIONALITY

Given the prevalence of feature selection techniques, it may be beneficial to establish a better idea of how they operate. In section 2.1 the relationship between sample size and dimensionality is shown empirically with simulated data sets. The proof in section 2.2 constitutes an initial attempt at developing a theoretical understanding of generalized feature selection behavior. Although such a relationship is highly dependent on the specific problem as well as the particular classifier, we show how to derive a "soft" upper bound on the number of samples required to identify the relevant features.

2.1. Empirical Results on Dimensionality vs. Samples

Many toy data sets were generated in order to test the relationship between dimensionality and training sample size. The data had a range of relevant and irrelevant features and was analyzed using a variety of feature selection techniques and classifiers.

2.1.1. The "Toy" Data

The generated data was drawn from three Gaussian distributions with different means and variances. The majority of the features were drawn from a distribution with 0 mean and represented the 'irrelevant' features which were shared equally between both classes. The number of these features ranged from ten, up to 600 and shared a constant variance. The means of the remaining two distributions were equally offset on either side of zero and shared similar variances. Typically there were anywhere between two and ten of these features correlated with the class labels. In some instances, the features were created as linear combinations of other relevant features.

2.1.2. The Feature Section and Classification

The two main feature selection techniques were a SVM feature selection technique, as described in [1], and a variation of the Fisher Score. The feature selection methods selected between two and ten relevant features from the data set and used only the selected features to train the classifier. The classifier then took as training inputs only the selected features and attempted to classify the entire test set. This classifier was typically either an SVM[1] or the Fisher Linear Discriminant[2].

2.1.3. Results

The results from these tests appeared independent of the classification techniques and the feature selection methods described in the previous section. Figure 1 depicts the relationship that appeared to be generalized across the tests.

Although changing the parameters of the relevant features shifted the slope of the line, an empirically linear relationship was seen on all data analyzed. This verifies that for at least the two feature selection techniques and the two classifiers used in this work, the theoretical results derived in section 2.2 do hold true. Whether the linear relationship between dimensionality and sample size is a generalized property for all feature selection techniques and classifiers remains a subject for further investigation.

---

[1] Using a linear kernel

[2] The separating hyperplane is defined in equation 1.11

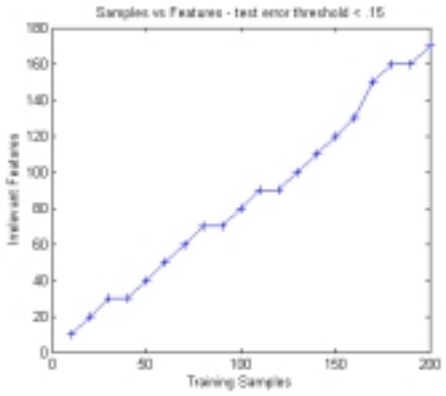

Figure 1: Empirical Dimensionality vs. Sample Size[3]

2.2. Deriving An Upper Bound on Amount of Training Data

It is important that this linear relationship between sample size and dimensionality have theoretical as well as empirical verification. Below is a proof combining several well-known theorems to derive the probability that no irrelevant features in the data set are outliers and could possibly be mistaken as relevant features.

Starting with the standard definitions of mean and variance:

$$E\left[\frac{X_1+\ldots+X_n}{n}\right]=\mu \tag{1.1}$$

$$Var\left(\frac{X_1+\ldots+X_n}{n}\right)=\frac{\sigma^2}{n} \tag{1.2}$$

and the Weak Law of Large Numbers:

$$as: n\to\infty$$

$$P\left\{\left|\frac{X_1+\ldots+X_n}{n}\right|\geq\varepsilon\right\}\to 0 \tag{1.3}$$

From Chebyshev's Inequality:

$$P\left\{\left|X-\mu\right|\geq\varepsilon\right\}\leq\frac{\sigma^2}{\varepsilon^2} \tag{1.4}$$

Combining the Weak Law of Large Numbers and Chebyshev's Inequality yields equation 1.5.

$$P\left\{\left|\frac{X_1+\ldots+X_n}{n}-\mu\right|\geq\varepsilon\right\}\leq\frac{\sigma^2}{n\varepsilon^2} \tag{1.5}$$

Simplifying 1.5:

$$P\left\{\left|\mu_m-\mu\right|\geq\varepsilon\right\}\leq\frac{\sigma^2}{n\varepsilon^2} \tag{1.6}$$

where μ_m is the empirical mean.

Equation 1.6 can be viewed as the probability of a single feature being an 'outlier' - being beyond a certain distance from its empirical mean.

Inversing this probability yields equation 1.7:

$$P\left\{\left|\mu_m-\mu\right|\leq\varepsilon\right\}\leq 1-\frac{\sigma^2}{n\varepsilon^2} \tag{1.7}$$

Equation 1.7 is simply the opposite of equation 1.6: the probability that a specific feature's empirical mean is within a certain bound from its true mean. Finally, equation 1.8 generalizes this bound for *m* features.

$$P\left\{\left|\mu_{mi}-\mu_i\right|\leq\varepsilon\right\}\leq\left(1-\frac{\sigma^2}{n\varepsilon^2}\right)^m \tag{1.8}$$

The results for a specific data set with variance of $\sigma^2=1.5$, a error threshold of .15, and $\varepsilon=2$ are in equation 1.8 and Figure 2 below.

$$\begin{aligned} &error\_threshold=.15\\ &\sigma^2=1.5\\ &\varepsilon=\frac{\left|\mu 1-\mu 2\right|}{2}=2\\ &1-.15\leq\left(1-\frac{.25}{2^2 n}\right)^m \end{aligned} \tag{1.9}$$

As evident in Figure 2, equation 1.9 yields a graph that appears to be linear. This result agrees with the results generated empirically in section 2.1.

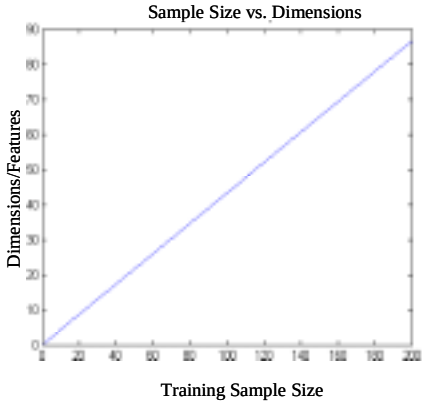

Figure 2: Theoretical Dimensionality vs. Sample Size

[3] This plot was generated using a data set described in section 2.1.1 with two relevant features, with variance of 1.5 and offset of 2. It employed a feature selection technique described in [1] that selected ten relevant features. The final classification was done using SVMfu with a linear kernel.

3. VERIFYING THE 'S-CURVE'

It was shown empirically in [1] that several feature selection techniques display an interesting property when their respective test error is plotted with sample size. As depicted Figure 3, a plateau is formed until the correct features are selected, at which point the error quickly drops to its lowest level. It was hypothesized that all feature selection techniques need to go through a number of iterations before they would be able to correctly identify the appropriate features. Upon reaching the critical number of samples, test error should quickly drop monotonically as features begin to be correctly identified.

To show that the 's-curve' phenomenon is a generalized property of all feature selection techniques, it must be validated empirically and theoretically. Hypothesis testing and deriving the number of samples required to develop a correct estimate of the relevant features' priors was one possibility at the theoretical solution. However, the generalization error for two Gaussians was calculated in section 3.1 instead. Empirically, the 's-curve' theory was tested using a Fisher Linear Discriminant.

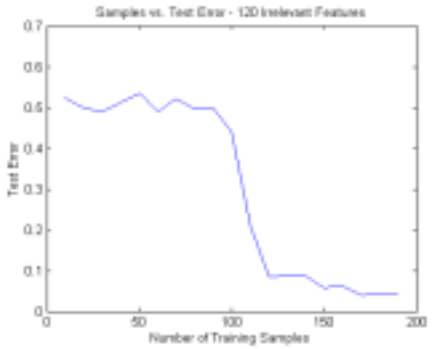

Figure 3: The Feature Selection "S-Curve"

3.1. Theoretical Validation of the 'S-Curve'

The equation for generalization error with two classes drawn from Gaussian distributions is shown in [2] and in equation 1.10.

$$A = \frac{1}{2}\left[erf\left(\frac{w \cdot \mu_1}{\left\| \sigma_1 w_1 \right\|} \right) + erf\left(-\frac{w \cdot \mu_2}{\left\| \sigma_2 w_2 \right\|} \right) \right] \qquad (1.10)$$

The separating hyperplane, w, is from the Fisher Linear Discriminant and is defined as:

$$w = \frac{\sigma_1 \mu_{m1} - \sigma_2 \mu_{m2}}{\sigma_1 + \sigma_2} \qquad (1.11)$$

Using a data set described in section 2.2.1, the generalization error given equation 1.10 was calculated for a variety of sample sizes. Only four features were selected. As Figures 3 and 4 show, no clear 's-curve' was ever established. After many iterations, the curve resembled an inverse power law. Modifying all of the model's parameters[4] never resulted in the desired 's-curve'.

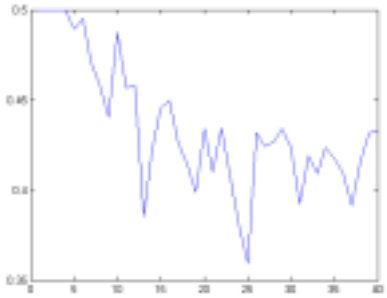

Figure 4: Theoretical Results after 1 Iteration

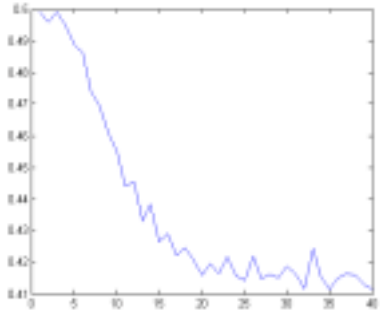

Figure 5: Theoretical Results after 50 Iterations

3.2. Empirical Validation of the 'S-Curve'

To attempt to verify the 's-curve' empirically, a data set was developed that seemed to be most likely to recreate it. Just as in section 3.1, only four features were selected. Fisher scores and the Fisher Linear Discriminant were used to select these four features and classify the test data.

As can be seen in Figure 6, after a single iteration the algorithm seemed to find a relevant feature, but then lost it when the number of samples was increased. However, by connecting the peaks, an definite 's-curve' does emerge. Fundamentally, this suggests that there is a critical number of samples above which, guarantees the correct features are selected given the current model parameters.

After running several of the tests, it was soon clear that the probability of the feature selection technique correctly selecting the appropriate feature was a function of the current sample size. The algorithm was unlikely to select the correct features with only a few samples, but the possibility remained. Thus, when 50 iterations were averaged, the error vs. sample size curve in Figure 7 looks very similar to an inverse power law curve - following the same shape as the theoretical results.

It is left to further research to establish whether this result is dependent on the type of distribution from which the data is drawn.[5]

[4] Including variances, means, number of relevant features, number of features selected

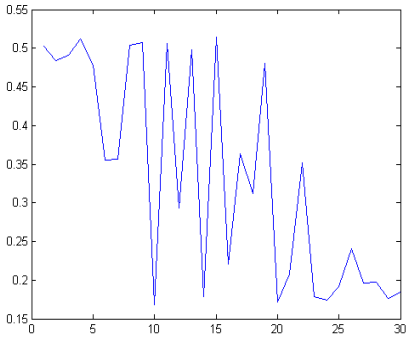

Figure 6: Empirical Results after 1 Iteration

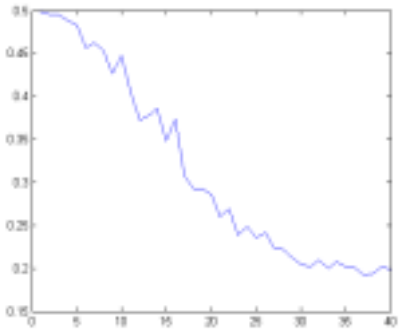

Figure 7: Empirical Results after 50 Iterations

4. CONCLUSIONS

This paper attempts to uncover generalized behavior common in features selection techniques. One such behavior seemed to be the linearity bound between dimensionality and sample size. This relationship was shown both theoretically and demonstrated empirically. Another behavior believed to be common to feature selection techniques was the 's-curve' generated when test error and samples size are plotted. Although it occurs using some feature selection techniques, the theory was easily disproved. While this initial work is inadequate to make broad generalizations about all feature selection techniques, it does provide some insight into their underlying behavior.

5. REFERENCES

[1] J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, and V. Vapnik. Feature selection for SVMs. In Sara A Solla, Todd K Leen, and Klaus-Robert Muller, editors, *Advances in Neural Information Processing Systems 13*. MIT Press, 2001.

[2] Mukherjee, S., PhD Thesis, *MIT*.

[5] As described in section 2.2.1, the data used for all the experiments in this paper consisted of Gaussian distributions.