Mitigating the "Curse of Dimensionality" in Machine Learning

Bob Durrant

School of Computer Science, University of Birmingham

Introduction

The expression "curse of dimensionality" was first coined by D. L. Donoho, a Stanford statistician, to describe the phenomenon that as data dimensionality increases it becomes more difficult to extract meaningful conclusions from a data set.

The reasons for this are many, but some typical problems are:

- The search space increases in size very rapidly with increasing dimensionality.
- Our geometric intuition regarding the structure of the data lets us down at high dimensions.
- The relative separation of points, as measured using some metric, becomes very low.

Focussing on the final point, it turns out that under very general conditions, all metrics and metric-like structures will suffer from the problem that if the dimensionality is high enough the relative separation between the data points is nearly zero. (Beyer et. al., 1999)

One method shown to be effective in mitigating this measure separation issue is the use of **fractional distance metrics**, as the relative separation of points shrinks more slowly using this approach than under other widely used metrics. (Aggarwal et. al., 2001)

Aim

An area where this issue concerning the relative separation of points is known to be problematic is when the data is high dimensional and the number of relevant features is comparatively small. Informally, the interesting features are "swamped" by the uninteresting ones.

I set out to show that in supervised learning (logistic regression with regularization), better results can be achieved if the regularization term employs a fractional distance metric as a measure of the size of the parameters than if the absolute value or the squared value of the parameters is used.

Ng's Paper on Machine Learning

In 2004 Andrew Ng, a computer scientist (also from Stanford), made a comparison of the properties of logistic regression with L_2-regularization (i.e. where the sum of the squared values of the parameters is used in the regularization term) and L_1-regularization (i.e. using absolute values) and demonstrated that L_1-regularization is better:

- L_2-regularization is rotationally invariant which has a negative impact on its effectiveness, especially when the number of relevant features or the training set size is small.

- L_1-regularization is not rotationally invariant.

- L_1-regularized logistic regression is capable of learning a problem even if the number of irrelevant features is exponentially larger than the training set size.

The following graphs, which are taken from Ng's paper, show the surprising extent to which the regularization method affects the final {0,1}-misclassification or log-loss error with respect to a hold-out cross-validation set.
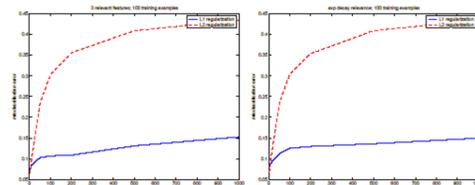
Misclassification error.
L_1- -vs- L_2-regularization. (Graph taken from Ng, 2004)
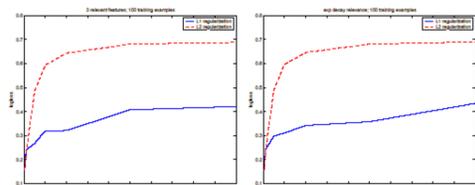
Log-loss error.
*L_1- -vs- L_2-regularization. (Graph taken from Ng, 2004)*

Results

Taking a similar approach to Ng, I used the method of **uniform covering number** bounds to prove some theoretical limits on how effective fractional distance metrics are in regularized logistic regression.

Covering numbers can be interpreted as an indication of how "complicated" the function to be learned is, and there are some quite general theorems that link the uniform covering number to the error achievable from a given training set size, or equivalently to the size of training set required in order to learn the function allowing some fixed margin of error.

Because of a technical difficulty (namely that fractional distance metrics do not obey the triangle inequality) many of the standard covering number results are not directly applicable to fractional distance metrics.

Resolving this issue proved to be hard, and ultimately upper-bounding the size of the training set required was achieved by bounding the fractional metric covering number between two values of the L_1-metric covering number. The picture below shows, for two dimensions, how the ball of the fractional $L_{(1/2)}$ distance metric can be bounded by two L_1-balls. The same trick can be carried out for higher dimensional data.
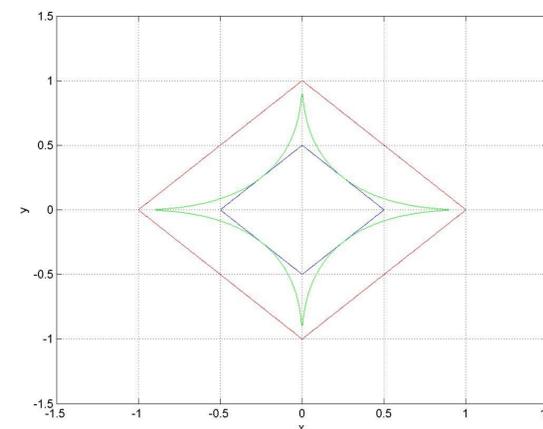
The $L_{(1/p)}$-metric ε-ball is bounded by the L_1-metric $\varepsilon/d^{(p-1)}$- and ε-balls, where d is the data dimensionality, and so the uniform covering numbers are likewise bounded.

Conclusion

By establishing the bounds on the uniform covering numbers we are able to conclude that logistic regression with fractional metric regularization is, at worst, big-Ω equivalent to logistic regression with L_1-regularization.

In fact it can also be established (by constructing some examples) that there are classes of functions for which the fractional uniform covering number must be strictly lower than that for the L_1-metric, which suggests that the performance should be better overall.

Some experimentation on both artificial and real-world data sets to validate this intuition is a natural next step to take as future work.

References

[1] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft. When is 'nearest neighbor' meaningful? In Proceedings 7th International Conference on Database Theory (ICDT'99), pages 217–235, 1999.

[2] C.C. Aggarwal, A. Hinneburg, and D.A. Keim. On the surprising behavior of distance metrics in high dimensional space. In Proceedings of the 8th International Conference on Database Theory ICDT 2001 (Lecture Notes in Computer Science Volume 1973/2001), pages 420–434, 2001.

[3] Andrew Y. Ng. Feature selection, L_1 vs. L_2 regularization, and rotational invariance. In Proceedings of the 21st International Conference on Machine Learning, 2004.

Acknowledgements