Non-trivial Questions

Danny Rogaar (s2393344)
Daan Opheikens (s3038416)
Panagiotis Giagkoulas (s3423883)
Saim Eser Comak (s3432548)
Carlos Huerta (s3743071)
Emile Muller (s3787915)

Group 14

October 7, 2018

1 How to detect if distance concentration is a problem for a data set?(Saim Eser Comak)

Nearest neighbor(NN) search is not feasible method when the number of dimensions goes to infinity. In this case, the variance of the division of the distance between two random points to the expected distance approaches to the zero. This means that the NN cannot be applied as all the data points surrounding the measurement of interest we want to cluster as all points will be equally distant. Let's first go over some formality that Bayer *et al.* [1] have put forward to make sense of this phenomenon. Within the measure theory of mathematics, for a finite sample $\chi = \{x_1, x_2, ..., x_m \dots\} \subset \mathbb{R}^m$, D^m_{max} and D^m_{min} denote the maximum and minimum distance points in hyperspace to the origin point with regards to the distance function used to calculate these distance norms. They are formally written as

$$D^m_{max} = max\{\|x^m_i\| = \rho\,(x^m_i, 0) : x^m_i \epsilon X^m\}$$
$$D^m_{min} = max\{\|x^m_i\| = \rho\,(x^m_i, 0) : x^m_i \epsilon X^m\}$$

Where $\|x\|$ is the norm of the vector and ρ is the distance function and m is the number of dimensions and the distance function p is used to calculate how dissimilar data points are. Using these two identities we can calculate relative contrast(RC) in which the distance function can be any of the Minkowski norms. RC is written as

$$\xi_p(m) = \frac{D^m_{max} - D^m_{min}}{D^m_{min}}$$

The value of $\xi_p(m)$ approaches to zero as the number of dimensions goes to infinity. Relative contrast is used to show degree of concentration. Seen in figure1, the x axis denotes the coefficient values that will be multiplied with m value. We see a fast decrease even after 2nd dimension however when we look at m=10x, the RC value begins already at 5. For m=100x, it is almost starting at zero.

2 How can we overcome the limitation of the underlying of linearity made by some filtering methods?

In general some authors recommend the use of wrapper or embedded methods that use a **linear** predictor as a filter, and after that train a **non-linear** model based on the variables selected by the linear filter. [2]

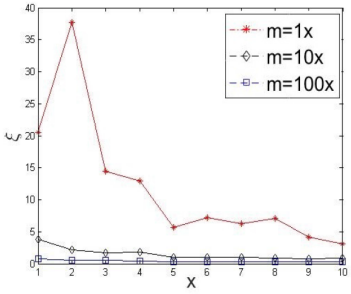

Figure 1: RC decreases as the number of dimensions increase in the x axis

3 Regarding the search used in the wrapper approach of feature selection, what is the main issue a hill-climbing search faces and what how could we avoid it?

Hill climbing is a greedy search technique that expands all possible alternatives of the current feature subset and and moves on to the one with the highest accuracy, until no further improvement takes place. However this approach makes the search vulnerable to local maxima, producing lees-than-optimal feature subsets. In order to avoid local maxima, a best-first search can be used [3]. The difference with hill-climbing is that this search terminates after a specified number of steps, over which no accuracy improvement has been detected, therefore allowing for the search to escape local maxima by furthering the search to the next possible subsets. It should be noted however that more search effort does not necessarily translate to better overall performance.

4 When is accuracy in embedded methods actually 'accurate'?

Embedded methods test their own performance to determine whether or not the algorithm has done its job the best it can. But this score might sometimes not be an accurate representation of the of the algorithm 'fitness'. Duval et al. [4] ran into this problem when using a Support Vector Machine algorithm in their project regarding genes. They state that "previous works have shown that cross-validation, and more specifically 10-fold cross-validation, provides an accuracy estimate with low variance.". By comparing different strategies of gene selection and the accuracy estimate obtained by the cross-validation, they are able to create reliable results. They state that it is necessary to include the gene selection process into the cross-validation schema. To do so, they split the data set before gene selection. If they did not do so, the accuracy results may be overestimated.

5 What does it mean for a variable to be in a dense region?

In the report, Unsupervised variable ranking is explained to use properties such as entropy and saliency of variables. The mentioned properties can be aptly understood as useful when selecting a model based on the data. For example, variables with high entropy do not give much information about data and should be avoided, similar to variables with very high saliency. However, the text did not go into detail on variable density, which refers to how much the variable correlates with other variables.

The correlation refers to the covariance divided by the individual variables variance. The correlation measure indicates whether variables grow in unison or whether one variable goes up as the other decreases. A variable in a high density region of data then implies that the variable depends on many other variables to raise and/or lower to explain its value. When depending on individual variables for a model, high density for variables, thus, may have a negative effect as it means violation of variable independence. In contrast, it is important to check up on the density of variables when explaining phenomenon that no model can when assuming variable independence.

In [5], Toloşi and Lengauer show that typical solutions will suffer performance when applied to common problems of genomics and transcriptomics. In these domains, large numbers of features combine with relatively small datasets leading to certain correlations between the features. The solutions (e.g. Lasso penalty [6]) then give lower relevance to large groups of correlated variables (i.e. dense) which biases the model towards simple features, even when the large groups have high total correlation to prediction. Although corrections are suggested, being aware of such phenomenon and high variable density guides the approach to building predictors on such datasets.

6 When does the Curse of Dimensionality occurs?

We can identify classically several areas in which curse of dimensionality appears. [7]

- In Optimization, Bellman's original usage. If we must approximately optimize a function of d variables and we know only that it is Lipschitz, say, then we need order $(1/\epsilon)^d$ evaluations on a grid in order to obtain an approximate minimizer within error ϵ.
- In Function Approximation. If we must approximate a function of d variables and we know only that it is Lipschitz, say, then we need order $(1/\epsilon)^d$ evaluations on a grid in order to obtain an approximation scheme with uniform approximation error ϵ.
- In Numerical Integration. If we must integrate a function of d variables and we know only that it is Lipschitz, say, then we need order $(1/\epsilon)^d$ evaluations on a grid in order to obtain an integration scheme with error ϵ.

References

[1] Kevin Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft. When is "nearest neighbor" meaningful? In *International conference on database theory*, pages 217–235. Springer, 1999.

[2] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *J. Mach. Learn. Res.*, 3:1157–1182, March 2003.

[3] Ron Kohavi and George H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97(1):273 – 324, 1997. Relevance.

[4] Duval B, Hao J.-K., and Hernandez J. C. H. A memetic algorithm for gene selection and molecular classification of an cancer. In *In Proceedings of the 11th Annual conference on Genetic and evolutionary computation.*, pages 201–208, New York, USA, 2009. ACM.

[5] Laura Toloşi and Thomas Lengauer. Classification with correlated features: unreliability of feature ranking and solutions. *Bioinformatics*, 27(14):1986–1994, 2011.

[6] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1994.

[7] David L. Donoho. High-dimensional data analysis: The curses and blessings of dimensionality. In *AMS CONFERENCE ON MATH CHALLENGES OF THE 21ST CENTURY*, 2000.