6-2020

Why LASSO, Ridge Regression, and EN: Explanation Based on Soft Computing

Woraphon Yamaka

Hamza Alkhatib

Ingo Neumann

Vladik Kreinovich

Comments:
Technical Report: UTEP-CS-20-68

Why LASSO, Ridge Regression, and EN: Explanation Based on Soft Computing

Woraphon Yamaka, Hamza Alkhatib, Ingo Neumann, and Vladik Kreinovich

Abstract In many practical situations, observations and measurement results are consistent with many different models – i.e., the corresponding problem is ill-posed. In such situations, a reasonable idea is to take into account that the values of the corresponding parameters should not be too large; this idea is known as *regularization*. Several different regularization techniques have been proposed; empirically the most successful are LASSO method, when we bound the sum of absolute values of the parameters, ridge regression method, when we bound the sum of the squares, and a EN method in which these two approaches are combined. In this paper, we explain the empirical success of these methods by showing that these methods can be naturally derived from soft computing ideas.

1 Formulation of the Problem

Need for regularization. In practice, in addition to measurement results, we often use imprecise expert knowledge.

For example, physicists usually believe that when the value of a physical quantity x is small, we expand the dependence $y = f(x)$ of some other quantity y on x in Taylor series and ignore quadratic and higher order terms in this expansion; see, e.g., [3, 12]. The usual argument is that when x is small, its square x^2 is so much smaller

Woraphon Yamaka
Faculty of Economics, Chiang Mai University, Chiang Mai, Thailand
e-mail: woraphon.econ@gmail.com

Hamza Alkhatib and Ingo Neumann
Geodesic Institute, Leibniz University of Hannover, Nienburger Str. 1, 30167 Hannover, Germany
e-mail: alkhatib@gih.uni-hannover.de, neumann@gih.uni-hannover.de

Vladik Kreinovich
Department of Computer Science, University of Texas at El Paso, El Paso, Texas 79968, USA
e-mail: vladik@utep.edu

than x that it can safely be ignored. This is indeed true: if $x = 10\% = 0.1$, then $x^2 = 0.01 \ll 0.1$; if $x = 1\% = 0,01$, then we can say that $x^2 = 0.0001 \ll x = 0.01$ with even higher confidence.

However, from the purely mathematical viewpoint, this argument is not fully convincing: indeed, the quadratic term in the Taylor expansion is not x^2, but $a_2 \cdot x^2$ for some coefficient a_2. From the purely mathematical viewpoint, this coefficient a_2 can be huge – in which case the product $a_2 \cdot x^2$ will also be big, and we will not be able to ignore it. From the physicist's viewpoint, however, this argument is valid, since physicists usually assume that the coefficients cannot be too large, they must be reasonably small.

This imprecise additional assumption underlies many successes of physics. It can also be used as a supplement to measurements when we try to estimate the values of the physical quantities. This is common sense. If, after applying some mathematical techniques, we get too large values of some parameters, this usually means that something is not right either with our method or with some measurement results – they may be outliers. In simple cases, it is clear: if we have a record of temperature in some area, and we see 17, 18, 19, 18, 17, and then suddenly 42 degrees, we should get very suspicious – especially if the next day, we again have the high of 19.

Physicists' intuition is great, but we cannot always rely on this intuition: there are many problems that need solving, and it is not realistic to expect to have a skilled physicist for each such problem. To deal with situations when a professional physicist is not available, we need to have a precise description of what we mean when we say that the coefficients $a_0, \ldots, a_n$ describing a model must be reasonably small. Such descriptions are known as *regularization*; see, e.g., [15].

Which regularizations are currently used. Out of many possible regularizations, the following three techniques have been most empirically successful:

- *LASSO* technique (short of Least Absolute Shrinkage and Selection Operator), when we limit the sum of the absolute values $\sum_{i=0}^{n} |a_i|$; see, e.g., [13];
- *ridge regression* method, in which we limit the sum of the squares $\sum_{i=0}^{n} a_i^2$; see, e.g., [4, 14]; and
- the *Elastic Net* (EN) method, in which we limit a linear combination of the above two sums; see, e.g., [5, 17].

Why. In this paper, we show that a natural formalization of commonsense intuition indeed leads to these three regularization techniques.

2 How Can We Describe Imprecise Expert Knowledge: A Brief Reminder

Need for degrees of confidence. In contrast to precise statements like "x is larger than 5" – which are either true or false – imprecise statements like "x is reasonably

small" are not well-defined. For some values x, for example, for $x = 0.0001$, the expert is absolutely sure that x is small; for other values like $x = 10^7$, the expert is usually absolutely sure that this value is not reasonably small. However, for intermediate values x, the expert is usually not 100% sure whether this value is indeed reasonably small – he or she is only sure to some degree.

It is therefore reasonable to ask the expert to assign, to each value x, a degree $\mu(x)$ to which this expert believes that x is reasonably small. We can use different scales for such degrees. Since in the computer, "absolutely true" is usually described as 1, and "absolutely false" as 0, it is convenient to use a scale from 0 to 1 for such degrees. This assignment is one of the main ideas behind *fuzzy logic* – a technique specifically developed to deal with such imprecision; see, e.g., [2, 6, 7, 8, 10, 11, 16].

This way, we can assign, to each imprecise statement, a function $\mu(x)$ that describes to what degree this statement is satisfied for each value x. This function is known as a *membership function* or a *fuzzy set*.

Need for "and"- and "or"-operations. Often, experts make complex statements: e.g., they may say that x is reasonably small, but not very small. This statement is obtained from the basic statements "x is reasonably small" and "x is very small" by applying connectives "not" and "but" (which here means the same as "and").

In general, we can use connectives "and", "or", and "not" to combine elementary statements into a composite one. Since experts may make such statements, it is desirable to estimate not only the expert's degrees of confidence in elementary statements, but also the expert's degrees of confidence in different combined statements. An ideal solution would be to simply ask the expert to provide such an estimate for all possible combinations, but this is not realistic: e.g., even if we simply consider possible "and"-combinations of some of n statements, we have $2^n - 1 - n$ possible combinations (as many as there are subsets of the set $\{1,\ldots,n\}$ (2^n) with the exception of an empty set and n one-element sets). For $n = 30$, we have billions of such combinations – there is no way to ask that many questions to an expert.

Since we cannot directly ask the expert his/her degree of confidence in each combination, we therefore need to be able to estimate the degree of confidence in a complex statement based on whatever information we have – i.e., based on the expert's degree of confidence in each elementary statement. This means, in particular, that we need to estimate the expert's degree of confidence in an "and"-statement $A\,\&\,B$ based on the known expert's degrees of confidence x and y in each of the two statements A and B. We will denote this estimate by $f_{\&}(x,y)$. The operation that inputs the pair (x,y) and returns $f_{\&}(x,y)$ is known as an *"and"-operation* or, for historical reasons, a *t-norm*.

Similarly, a function that maps the pair (x,y) into an estimate for the expert's degree of confidence in $A \vee B$ is denoted by $f_{\vee}(x,y)$ and is known as an *"or"-operation* or a *t-conorm*.

These operations must satisfy several natural requirements. For example, since $A\,\&\,B$ means the same as $B\,\&\,A$, it is reasonable to require that the estimates for these two statements will be the same, i.e., that the "and"-operation must be commutative: $f_{\&}(x,y) = f_{\&}(y,x)$. Similarly, since $A\,\&\,(B\,\&\,C)$ means the same as $(A\,\&\,B)\,\&\,C$, the "and"-operation must be associative. Similarly, the "or"-operation must be commu-

tative and associative. Also, both operations should be monotonic is each of the variables, etc.

Need for Strictly Archimedean operations. With all these requirements, there are many different "and"- and "or"-operations. In particular, for each strictly increasing functions $f(x)$, the operation $f^{-1}(f(x)\cdot f(y))$ is an "and"-operation. Such "and"-operations are known as *strictly Archimedean*.

In this paper, we will take into account a result from [9] that for every "and"-operation $f_{\&}(a,b)$ and for every $\varepsilon > 0$, there exists a strictly Archimedean "and"-operation whose value is ε-close to $f_{\&}(x,y)$ for all x and y:

$$|f_{\&}(x,y) - f^{-1}(f(x)\cdot f(y))| \le \varepsilon.$$

From the practical viewpoint, very small differences in degree of confidence can be ignored. Thus, from the practical viewpoint, we can always assume that the "and"-operation is strictly Archimedean.

3 Let Us Apply Uncertainty Techniques to Our Problem: Why LASSO and Ridge Regression

General analysis of the problem. The main idea behind regularization is that a tuple $a = (a_0, \ldots, a_n)$ is accepted if the absolute values $|a_i|$ of all the coefficients are reasonably small. In other words, the value $|a_0|$ must be reasonably small *and* the value $|a_1|$ must be reasonably small, etc. We must select tuples a for which our degree of confidence $\mu_0(a)$ in this complex statement should be sufficiently large, i.e, larger than a certain threshold d_0.

According to the above general explanation, to estimate the degree of confidence $\mu_0(a)$ in our complex statement, we need to apply the corresponding "and"-operation $f_{\&}(x,y)$ to the degrees to which each $|a_i|$ is sufficiently small. These degrees, by definition of the membership function, can be obtained by applying the membership function $\mu(x)$ corresponding to "sufficiently small" to the values $|a_i|$. In other words, each of these degrees ie equal to $\mu(|a_i|)$. Thus, the degree of confidence that the above complex statement is true is equal to $\mu_0(a) = f_{\&}(\mu(|a_0|), \ldots, \mu(|a_n|))$. In these terms, the tuple of coefficient $a = (a_0, \ldots, a_n)$ is accepted if

$$\mu_0(a) = f_{\&}(\mu(|a_0|), \ldots, \mu(|a_n|)) \ge d_0. \quad (1)$$

Clearly, the larger the value x, the smaller the degree of confidence that this value is reasonably small. Thus, the membership function $\mu(x)$ that corresponds to "reasonably small" is a decreasing function of x.

We have agreed to assume that the "and"-operation is strictly Archimedean, i.e., that $f_{\&}(x,y) = f^{-1}(f(x)\cdot f(y))$ for some strictly increasing function $f(x)$. Thus, the condition (1) takes the form

$$\mu_0(a) = f^{-1}(f(\mu(|a_0|) \cdot \ldots \cdot f(\mu(|a_n|)) \geq d_0.$$

By applying the increasing function $f(x)$ to both sides of this inequality, we get an equivalent inequality

$$F_0(a) = F(|a_0|) \cdot \ldots \cdot F(|a_n|) \geq D_0, \tag{2}$$

where we denoted $F_0(a) \stackrel{\text{def}}{=} f(\mu_0(a))$, $F(x) \stackrel{\text{def}}{=} f(\mu(x))$ and $D_0 \stackrel{\text{def}}{=} f(d_0)$.

Since the function $f(x)$ is increasing and $\mu(x)$ is decreasing, the composition $F(x) = f(\mu(x))$ of these two functions is a decreasing function of x.

To further analyze this situation, we need to make some additional assumptions reflecting commonsense. In this paper, we will describe two such natural assumptions, and we will show that they lead, correspondingly, to LASSO and to the ridge regression.

Why LASSO. A reasonable idea is that if x and y are reasonably small, then their sum $x+y$ is also reasonable small. Thus, it is reasonable to conclude that for the membership function $\mu(x)$ that corresponds to "reasonable small", the degree to which $x+y$ is reasonably small is equal to the degree that x is reasonably small and y is reasonably small, i.e., that

$$\mu(x+y) = f_{\&}(\mu(x), \mu(y)). \tag{3}$$

What we can deduce from this idea? We have assumed that the "and"-operation is strictly Archimedean, so the equality (3) has the form

$$\mu(x+y) = f^{-1}(f(\mu(x)) \cdot f(\mu(y)).$$

By applying the function $f(x)$ to both sides of this equality, we conclude that $f(\mu(x+y)) = f(\mu(x)) \cdot f(\mu(y))$, i.e., that $F(x+y) = F(x) \cdot F(y)$. It is known (see, e.g., [1]) that every decreasing solution to this functional equation has the form $F(x) = \exp(-k \cdot x)$ for some $k > 0$. Thus, the inequality (2) takes the form

$$F_0(a) = \exp(-k \cdot |a_0|) \cdot \ldots \cdot \exp(-k \cdot |a_n|) \geq D_0,$$

i.e., equivalently, the form

$$F_0(a) = \exp\left(-k \cdot \sum_{i=0}^{n} |a_i|\right) \geq D_0.$$

By taking the logarithm of both sides and dividing both sides of the resulting inequality by $-k$, we get an equivalent inequality

$$|a_0| + \ldots + |a_n| \leq c_0,$$

where we denoted $c_0 \stackrel{\text{def}}{=} -\dfrac{\ln(D_0)}{k}$. This is exactly the LASSO approach, so we indeed justified the use of LASSO regularization.

Why ridge regression. Another reasonable idea isthat if all the coordinates of a point are reasonably small, then the distance from this point to the origin of the coordinate system is also small. In the 2-D case, the distance between the point with coordinates (x,y) and the origin $(0,0)$ of the coordinate system is equal to $\sqrt{x^2+y^2}$. Thus, we conclude that if x and y are reasonably small, then the value $\sqrt{x^2+y^2}$ is also reasonably small. So, it is reasonable to conclude that for the membership function $\mu(x)$ that corresponds to "reasonable small", the degree to which $\sqrt{x^2+y^2}$ is reasonably small is equal to the degree that x is reasonably small and y is reasonably small, i.e., that

$$\mu\left(\sqrt{x^2+y^2}\right) = f_{\&}(\mu(x),\mu(y)). \tag{4}$$

What we can deduce from this idea? We have assumed that the "and"-operation is strictly Archimedean, so the equality (4) has the form

$$\mu\left(\sqrt{x^2+y^2}\right) = f^{-1}(f(\mu(x))\cdot f(\mu(y)).$$

By applying the function $f(x)$ to both sides of this equality, we conclude that $f\left(\mu\left(\sqrt{x^2+y^2}\right)\right) = f(\mu(x))\cdot f(\mu(y))$, i.e., that

$$F\left(\sqrt{x^2+y^2}\right) = F(x)\cdot F(y).$$

Thus, for an auxiliary function $G(x) \stackrel{\text{def}}{=} F(\sqrt{x})$ for which $F(x) = G(x^2)$, we get $G(x^2+y^2) = G(x^2)\cdot G(y^2)$. This is true for all possible non-negative values x and y. Every non-negative number X can be represented as a square: namely, as $X = x^2$ for $x = \sqrt{X}$. Thus, for all possible non-negative numbers X and Y, we have $G(X+Y) = G(X)\cdot G(Y)$. As we have mentioned in our derivation of LASSO, for a monotonic function $G(X)$, this implies that $G(X) = \exp(-k\cdot X)$ for some $k > 0$. Thus, we conclude that $F(x) = G(x^2) = \exp(-k\cdot x^2)$.

So, the inequality (2) takes the form

$$F_0(a) = \exp(-k\cdot a_0^2)\cdot\ldots\cdot\exp(-k\cdot a_n^2) \geq D_0,$$

or, equivalently,

$$F_0(a) = \exp\left(-k\cdot\sum_{i=0}^{n} a_i^2\right) \geq D_0.$$

By taking the logarithm of both sides and dividing both sides of the resulting inequality by $-k$, we get an equivalent inequality

$$a_0^2+\ldots+a_n^2 \leq c_0,$$

where we denoted $c_0 \stackrel{\text{def}}{=} -\dfrac{\ln(D_0)}{k}$. This is exactly the ridge regression approach, so we indeed justified the use of ridge regression.

4 Why EN

Idea. In the previous section, we considered the case when we have a *single* expert. In practice, we often have *several* different experts corresponding to different areas of expertise. Each expert can dismiss some of the possible models since they are not realistic according to his or her area of expertise. It is therefore reasonable to conclude that a tuple $a = (a_0, \ldots, a_n)$ of possible values of parameters is reasonable if all the experts consider it reasonable.

Let us formalize and explore this idea. Let E denote the number of experts, and let $\mu_j(a)$ ($j = 1, \ldots, E$) denote the degree to which the tuple a is reasonable according to the j-th expert. The overall degree that all the experts consider this tuple to be reasonable is thus equal to $f_\&(\mu_1(a), \ldots, \mu_E(a))$. So, we accept this tuple if this overall degree is greater than or equal to some threshold d_0:

$$f_\&(\mu_1(a), \ldots, \mu_E(a)) \geq d_0.$$

For the strictly Archimedean "and"-operation, this inequality takes the form

$$f^{-1}(f(\mu_1(a)) \cdot \ldots \cdot f(\mu_E(a)) \geq d_0.$$

By applying the function $f(x)$ to both sides, we get an equivalent inequality $f(\mu_1(a)) \cdot \ldots \cdot f(\mu_E(a)) \geq D_0$, i.e., $F_1(a) \cdot \ldots \cdot F_E(a) \geq D_0$, where $D_0 \stackrel{\text{def}}{=} f(d_0)$.

From the previous section, we know that for each expert j, the function $F_j(a) = f(\mu_j(a))$ takes either the form $F_j(a) = \exp\left(-k_j \cdot \sum\limits_{i=0}^{n} |a_i|\right)$ or the form $F_j(a) = \exp\left(-k_j \cdot \sum\limits_{i=0}^{n} a_i^2\right)$. By grouping together experts with these types of functions, we conclude that the acceptance criterion takes the form

$$\left(\prod_{j \in E_1} \exp\left(-k_j \cdot \sum_{i=0}^{n} |a_i|\right)\right) \cdot \left(\prod_{j \in E_2} \exp\left(-k_j \cdot \sum_{i=0}^{n} a_i^2\right)\right) \geq D_0,$$

where E_1 is the set of all experts whose functions $F_j(a)$ take the LASSO form and E_2 is the set of all experts whose functions $F_j(a)$ take the ridge regression form. The above inequality can be represented in the equivalent form

$$\exp\left(-K_1 \cdot \sum_{i=0}^{n} |a_i| - K_2 \cdot \sum_{i=0}^{n} a_i^2\right) \geq D_0,$$

where $K_1 \stackrel{\text{def}}{=} \sum_{j \in E_1} k_j$ and $K_2 \stackrel{\text{def}}{=} \sum_{j \in E_2} k_j$.

By taking logarithms of both sides and dividing the resulting inequality by $-K_1$, we get an equivalent inequality

$$\sum_{i=0}^{n} |a_i| + c \cdot \sum_{i=1}^{n} a_i^2 \le c_0,$$

where $c \stackrel{\text{def}}{=} K_2/K_1$ and $c_0 \stackrel{\text{def}}{=} -\dfrac{\ln(D_0)}{K_1}$. This is exactly EN approach – thus EN regularization is also justified.

Acknowledgments

The first author is grateful for the financial support of the Center of Excellence in Econometrics, Chiang Mai University, Thailand.

This work was also supported by the Institute of Geodesy, Leibniz University of Hannover, and by the US National Science Foundation grants 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science) and HRD-1242122 (Cyber-ShARE Center of Excellence).

This paper was written when V. Kreinovich was visiting Leibniz University of Hannover.

References

1. J. Aczel and J. Dhombres, *Functional Equations in Several Variables*, Cambridge University Press, Cambridge, UK, 1989.
2. R. Belohlavek, J. W. Dauben, and G. J. Klir, *Fuzzy Logic and Mathematics: A Historical Perspective*, Oxford University Press, New York, 2017.
3. R. Feynman, R. Leighton, and M. Sands, *The Feynman Lectures on Physics*, Addison Wesley, Boston, Massachusetts, 2005.
4. A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems", *Technometrics*, 1970, Vol. 12, No. 1, pp. 55–67.
5. B. Kargoll, M. Omidalizarandi, I. Loth, J.-A. Paffenholz, and H. Alkhatib, "An iteratively reweighted least-squares approach to adaptive robust adjustment of parameters in linear regression models with autoregressive and t-distributed deviations", *Journal of Geodesy*, 2018, Vol. 92, No. 3, pp. 271–297.
6. G. Klir and B. Yuan, *Fuzzy Sets and Fuzzy Logic*, Prentice Hall, Upper Saddle River, New Jersey, 1995.
7. J. M. Mendel, *Uncertain Rule-Based Fuzzy Systems: Introduction and New Directions*, Springer, Cham, Switzerland, 2017.
8. H. T. Nguyen and V. Kreinovich, "Nested intervals and sets: concepts, relations to fuzzy sets, and applications", In: R. B. Kearfott and V. Kreinovich (eds.), *Applications of Interval Computations*, Kluwer, Dordrecht, 1996, pp. 245–290.

9. H. T. Nguyen, V. Kreinovich, and P. Wojciechowski, "Strict Archimedean t-Norms and t-Conorms as Universal Approximators", *International Journal of Approximate Reasoning*, 1998, Vol. 18, Nos. 3–4, pp. 239–249.
10. H. T. Nguyen, C. Walker, and E. A. Walker, *A First Course in Fuzzy Logic*, Chapman and Hall/CRC, Boca Raton, Florida, 2019.
11. V. Novák, I. Perfilieva, and J. Močkoř, *Mathematical Principles of Fuzzy Logic*, Kluwer, Boston, Dordrecht, 1999.
12. K. S. Thorne and R. D. Blandford, *Modern Classical Physics: Optics, Fluids, Plasmas, Elasticity, Relativity, and Statistical Physics*, Princeton University Press, Princeton, New Jersey, 2017.
13. R. Tibshirani, "Regression shrinkage and selection via the lasso", *Journal of the Royal Statistical Society*, 1996, Vol. 58, No. 1.
14. A. N. Tikhonov, "On the stability of inverse problems", *Doklady Akademii Nauk SSSR*, 1943, Vol. 39, No. 5, pp. 195–198.
15. A. N. Tikhonov and V. Y. Arsenin, *Solutions of Ill-Posed Problems*, Winston and Sons, Washington, DC, 1977.
16. L. A. Zadeh, "Fuzzy sets", *Information and Control*, 1965, Vol. 8, pp. 338–353.
17. H. Zou and T. Hastie, "Regularization and variable selection via the elastic net', *Journal of the Royal Statistical Society B*, 2005, Vol. 67, pp. 301–320.