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

Woraphon Yamaka[1], Hamza Alkhatib[2],
Ingo Neumann[2], and Vladik Kreinovich[3]
[1]Faculty of Economics, Chiang Mai University
Chiang Mai, Thailand, woraphon.econ@gmail.com
[2]Geodesic Institute, Leibniz University of Hannover
Hannover, Germany, alkhatib@gih.uni-hannover.de
neumann@gih.uni-hannover.de
[3]Department of Computer Science, University of Texas at El Paso
El Paso, Texas 79968, USA, vladik@utep.edu

1. 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.
- The usual argument is that:
  - when x is small,
  - its square x^2 is so much smaller than x that it can safely be ignored.

2. Need for Regularization (cont-d)

- 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 this 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.

3. Need for Regularization (cont-d)

- Indeed, 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 estimate the values of physical quantities.
- This is common sense.
- Sometimes, 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.

4. Need for Regularization (cont-d)

- In simple cases, it is clear that 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.
- It is not realistic to expect to have a skilled physicist for each such problem.
- How to deal with situations when a professional physicist is not available?

5. Need for Regularization (cont-d)

- 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*.

6. Which Regularizations Are Currently Used

- Out of many possible regularizations, the following three techniques have been most empirically successful:
  - *LASSO* technique when we limit the sum of the absolute values $\sum\limits_{i=1}^{n} |a_i|$;
  - *ridge regression* method, in which we limit the sum of the squares $\sum\limits_{i=0}^{n} a_i^2$; and
  - the *Elastic Net* (EN) method, in which we limit a linear combination of the above two sums.
- Why?
- In this paper, we show that:
  - a natural formalization of commonsense intuition
  - indeed leads to these three regularization techniques.

7. Need for Degrees of Confidence

- Precise statements like "x is larger than 5" are either true or false.
- In contrast, 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.

8. Need for Degrees of Confidence (cont-d)

- 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.
- In the computer, "absolutely true" is usually described as 1, and "absolutely false" as 0.
- So, 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*.
- This technique was specifically developed to deal with such imprecision.

9. Need for Degrees of Confidence (cont-d)

- 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*.

10. Need for "And"- and "Or"-Operations

- Often, experts make complex statements.
- For example, 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.

11. "And"- and "Or"-Operations (cont-d)

- 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.
- However, this is not realistic.
- 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), except for empty set and 1-element sets.

12. “And”- and “Or”-Operations (cont-d)

- For $n = 30$, we have billions of such combinations.
- There is no way to ask that many questions to an expert.
- 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.

13. "And"- and "Or"-Operations (cont-d)

- 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*.

14. "And"- and "Or"-Operations (cont-d)

- 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: $f_\&(x, y) = f_\&(y, x)$.

15. "And"- and "Or"-Operations (cont-d)

- Similarly, since $A \,\&\, (B \,\&\, C)$ means the same as $(A \,\&\, B) \,\&\, C$, the "and"-operation must be associative.
- Similarly, the "or"-operation must be commutative and associative.
- Also, both operations should be monotonic is each of the variables, etc.

16. 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*.
- Let us take into account a known result that:
  - for every "and"-operation $f_\&(a,b)$ and 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 Archimedean.

17. 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.

18. General Analysis of the Problem (cont-d)

- So, to estimate the degree of confidence $\mu_0(a)$ in our complex statement:
  - we need to apply the corr. "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 $\mu_0(a) = f_\&(\mu(|a_0|), \ldots, \mu(|a_n|))$.
- So, the tuple of coefficient $a = (a_0, \ldots, a_n)$ is accepted if $\mu_0(a) = f_\&(\mu(|a_0|), \ldots, \mu(|a_n|)) \geq d_0$.

19. General Analysis of the Problem (cont-d)

- 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.
- So, $f_\&(x,y) = f^{-1}(f(x)\cdot f(y))$ for some strictly increasing function $f(x)$.
- Thus, the above condition 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:

$$F_0(a) = F(|a_0|) \cdot \ldots \cdot F(|a_n|) \geq D_0.$$

20. General Analysis of the Problem (cont-d)

- *Reminder:* $F_0(a) = F(|a_0|) \cdot \ldots \cdot F(|a_n|) \geq D_0$.
- Here we denoted $F_0(a) \stackrel{\text{def}}{=} f(\mu_0(a))$, $F(x) \stackrel{\text{def}}{=} f(\mu(x))$ and $D_0 \stackrel{\text{def}}{=} f(d_0)$.
- The function $f(x)$ is increasing and $\mu(x)$ is decreasing.
- Thus, 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.

21. Why LASSO

- A reasonable idea is that if x and y are reasonably small, then their sum $x+y$ is also reasonable small.
- So, it's reasonable to conclude that for the membership function $\mu(x)$ corresponding 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

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

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

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

22. Why LASSO (cont-d)

- By applying the function $f(x)$ to both sides of this equality, we conclude that:

$$f(\mu(x+y)) = f(\mu(x))\cdot f(\mu(y)), \text{ i.e., } F(x+y) = F(x)\cdot F(y).$$

- It's known that every decreasing solution to this equation has the form: $F(x) = \exp(-k \cdot x)$ for some $k > 0$.
- Thus, the above inequality takes the form

$$F_0(a) = \exp(-k \cdot |a_0|) \cdot \ldots \cdot \exp(-k \cdot |a_n|) \geq D_0, \text{ i.e.}$$

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

23. Why LASSO (cont-d)

- *Reminder:* $F_0(a) = \exp\left(-k \cdot \sum_{i=0}^{n} |a_i|\right) \ge D_0.$
- By taking the logarithm of both sides and dividing both sides of the resulting inequality by $-k$, we get:

$$|a_0| + \ldots + |a_n| \le c_0, \text{ where } c_0 \stackrel{\text{def}}{=} -\frac{\ln(D_0)}{k}.$$

- This is exactly the LASSO approach, so we indeed justified the use of LASSO regularization.

24. Why Ridge Regression

- Another reasonable idea is that:
  - 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 (x, y) and the origin $(0, 0)$ of the coordinate system is $\sqrt{x^2 + y^2}$.
- Thus, we conclude that if x and y are reasonably small, then the value $\sqrt{x^2 + y^2}$ is also reasonably small.

25. Why Ridge Regression (cont-d)

- 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

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

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

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

26. Why Ridge Regression (cont-d)

- By applying the function $f(x)$ to both sides of this equality, we conclude that

$$f\left(\mu\left(\sqrt{x^2+y^2}\right)\right) = f(\mu(x)) \cdot f(\mu(y)), \text{ i.e., that}$$

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

- 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 $G(X+Y) = G(X) \cdot G(Y)$.

Need for Regularization
Which Regularizations . . .
Need for Degrees of . . .
Need for "And"- and . . .
Need for Strictly . . .
General Analysis of the . . .
Why LASSO
Why Ridge Regression
Why EN: Idea
Home Page
Title Page
Page 27 of 34
Go Back
Full Screen
Close
Quit

27. Why Ridge Regression (cont-d)

- As we have mentioned in our derivation of LASSO, for a monotonic function $G(X)$, this implies that

$$G(X) = \exp(-k \cdot X) \text{ for some } k > 0.$$

- Thus, we conclude that $F(x) = G(x^2) = \exp(-k \cdot x^2)$.
- So, the above inequality takes the form

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

- This is equivalent to

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

28. Why Ridge Regression (cont-d)

- *Reminder:* $F_0(a) = \exp\left(-k \cdot \sum\limits_{i=0}^{n} a_i^2\right) \geq D_0.$
- By taking the logarithm of both sides and dividing both sides of the resulting inequality by $-k$, we get:

$$a_0^2 + \ldots + a_n^2 \leq c_0, \text{ where } c_0 \stackrel{\text{def}}{=} -\frac{\ln(D_0)}{k}.$$

- This is exactly the ridge regression approach, so we indeed justified the use of ridge regression.

29. Why EN: Idea

- In the previous sections, 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 models if they are not realistic according to his/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.

30. Let Us Formalize and Explore This Idea

- Let E denote the number of experts.
- 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.,

$$F_1(a) \cdot \ldots \cdot F_E(a) \geq D_0, \text{ where } D_0 \stackrel{\text{def}}{=} f(d_0).$$

31. Let Us Explore This Idea (cont-d)

- From the previous sections, 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 get:

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

- Here, E_1 is the set of all LASSO experts and E_2 the set of all ridge regression experts.

32. Let Us Explore This Idea (cont-d)

- The above inequality can be represented in the equivalent form:

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

- Here $K_1 \stackrel{\text{def}}{=} \sum\limits_{j \in E_1} k_j$ and $K_2 \stackrel{\text{def}}{=} \sum\limits_{j \in E_2} k_j$.
- By taking logarithms of both sides and dividing the resulting inequality by $-K_1$, we get:

$$\sum_{i=0}^{n} |a_i| + c \cdot \sum_{i=1}^{n} a_i^2 \leq c_0, \text{ where } c \stackrel{\text{def}}{=} K_2/K_1 \text{ and } c_0 \stackrel{\text{def}}{=} -\frac{\ln(D_0)}{K_1}.$$

- This is exactly EN approach – thus EN regularization is also justified.

33. Acknowledgments

This work was supported:

- by the Center of Excellence in Econometrics, Chiang Mai University, Thailand.
- by the Institute of Geodesy, Leibniz University of Hannover, and
- by the US National Science Foundation grants 1623190 and HRD-1242122.

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