Review of "Understanding Black-box Predictions via Influence Functions"

Eole Cervenka [1] **Geovani Rizk** [1]

Abstract

Pang Wei Koh and Percy Liang introduce a new way to analyze a model predictions via its training data. () Pang Wei Koh & Percy Liang, 2017) We explain how influence functions can be used to efficiently track the impact of each training points on any test point prediction, even when using a non-convex and/or non-differentiable loss function. We reproduce use cases of tracking influence in understanding model behavior, learning adversarial training, debugging a model and checking the labels efficiently. We show the usefulness of using influence functions on other domain like NLP, etc...

1. Introduction

State-of-the-art learning models such as deep neural networks are often black boxes (Krizhevsky et al., 2012) and understanding model predictions pose a challenge that limits our ability to interpret, debug existing models or create new models. Prior efforts in understanding model predictions have always started with the assumption that model where fixed functions; taking an input and producing an output.This paper takes a new approach by measuring the influence of training data in the learning process and using this information to measure the influence of training data on the prediction process. The brute-force method to measure how changing the training data can affect the model predictions is prohibitively expensive as it requires retraining for each modified training set. Influence functions offer a closed-form solution to approximate new model parameters based on the previous models after training set perturbation without retraining from scratch. Using influence function requires an expensive second derivative calculations (Hessian matrix). Well see how Hessian Vector Product can be used to make influence functions tractable even in models with millions of parameters. Influence functions require differentiability and convexity constraints that are not satisfied in state of the art models (e.g. deep learning). The authors present how influence functions can be accurately approximated using 2nd order optimization techniques. Finally, the direct applications of this idea will be experimented through use cases of debugging a model, detecting dataset errors and engineering adversarial training points.

[*]Equal contribution [1]University of Paris Dauphine, Paris, France. Correspondence to: Eole Cervenka <eole.cervenka@dauphine.eu>, Geovani Rizk <geovani.rizk@dauphine.eu>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

2. Influence Function approach

Suppose that we want to build a model for a prediction problem from the input space $\mathcal{X}$ to the output space $\mathcal{Y}$. We use the same notation that the original authors and give the training points $z_1, ..., z_n$ where $z_i = (x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$. Let $L(z, \theta)$ be the loss value for a certain training point z and the parameters θ of the model, and let $\frac{1}{n}\sum_{i=1}^{n} L(z_i, \theta)$ be the empirical risk. By definition, the empirical risk minimizer is given by $\hat{\theta} \stackrel{\text{def}}{=} \arg\min_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(z_i, \theta)$. In the first parts of this section we suppose that the empirical risk is twice differentiable and strictly convex. In 2.X, we explain how influence function can be approximated accurately and provide useful informations even in non-convex and/or non-differentiable empirical risk context.

2.1. Training set modification

In this section, we will demonstrate how influence functions can measure the impact of a modification in the training set on the parameters and the model prediction. Two kinds of dataset modification can be tracked this way:

- removing a training point which can be used to debug the model and to detect dataset errors.
- perturbing a training point which can be used to create adversarial training examples.

In both types of dataset modification, measuring the impact with the brute-force method requires retraining the model which is not tractable.

2.1.1. Removing a training point

Suppose we want to see the effect of removing a point from the training set on parameters; it is equivalent to upweigth-

ing said point by $\epsilon = -\frac{1}{n}$. Influence functions allows us to express as a closed form expression the variation in θ with respect to a variation in ϵ. Thus, it can be used to linearly approximate the effect of upweighting a point in the training set on θ.

Influence of upweighting a point z on parameters is given by :

$$\begin{aligned}\mathcal{I}_{\text{up,params}}(z) &\stackrel{\text{def}}{=} \frac{d\hat{\theta_{\epsilon,z}}}{d\epsilon}|_{\epsilon=0} \\ &= -\mathcal{H}_{\hat{\theta}}^{-1}\nabla_\theta \mathcal{L}(z,\hat{\theta}) \end{aligned} \quad (1)$$

The variation of the loss on z_{test} with respect to the parameters is given by $\nabla_\theta \mathcal{L}(z_{\text{test}},\hat{\theta})$. By chaining, we can express the variation of the loss on z_{test} with respect to ϵ :

$$\begin{aligned}\mathcal{I}_{\text{up,loss}}(z) &= \nabla_\theta \mathcal{L}(z_{\text{test}},\hat{\theta})\ \frac{d\hat{\theta_{\epsilon,z}}}{d\epsilon}|_{\epsilon=0} \\ &= -\nabla_\theta \mathcal{L}(z_{\text{test}},\hat{\theta})\mathcal{H}_{\hat{\theta}}^{-1}\nabla_\theta \mathcal{L}(z,\hat{\theta}) \end{aligned} \quad (2)$$

2.1.2. Perturbing a training point

Perturbing a training point z is equivalent to remove it and add its pertubed version $z_\delta = (x+\delta, y)$. Using what we have previously introduce about adding weight to the point, the empirical risk minimizer can be expressed by $\hat{\theta}_{\epsilon,z_\delta,-z} \stackrel{\text{def}}{=} \arg\min_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(z_i,\theta) + \epsilon\mathcal{L}(z_\delta,\theta) - \epsilon\mathcal{L}(z,\theta)$.

Analogically, influence of perturbing a point z on parameters can be expressed as the influence of transfering the weight from point z to point z_δ :

$$\begin{aligned}\mathcal{I}_{\text{pert,params}}(z) &= \frac{d\hat{\theta}_{\epsilon,z_\delta,-z}}{d\epsilon}|_{\epsilon=0} \\ &= \mathcal{I}_{\text{up,params}}(z_\delta) - \mathcal{I}_{\text{up,params}}(z) \\ &= -\mathcal{H}_{\hat{\theta}}^{-1}\left(\nabla_\theta \mathcal{L}(z_\delta,\hat{\theta}) - \nabla_\theta \mathcal{L}(z,\hat{\theta})\right) \end{aligned} \quad (3)$$

Suppose that x is continuous and δ small enough, we can linearly approximate $\nabla_\theta \mathcal{L}(z_\delta,\hat{\theta}) - \nabla_\theta \mathcal{L}(z,\hat{\theta})$ by $\nabla_x \nabla_\theta \mathcal{L}(z,\hat{\theta})\delta$.

???

2.2. Additionnal insights

The authors compare the insights from using euclidian distance versus using influence function in the context of searching for influential training point for a given prediction in a logistic regression model.

The euclidian distance (or cosine distance if the points are normalized) is given by $x \cdot x_{test}$.

The influence function that approximates the impact of upweighting a training point on the loss, in a logit regression is given by:

$$y_{\text{test}} y \cdot \sigma(-y_{\text{test}}\theta^\top x_{\text{test}}) \cdot \sigma(-y\theta^\top x) x_{test}^\top \mathcal{H}_{\hat{\theta}}^{-1} x \quad (4)$$

where $\sigma(t) = \frac{1}{1+\exp(-t)}$

- $\mathcal{H}_{\hat{\theta}}^{-1}$: The inverse Hessian (weighted covariance matrix) is responsible for upweighting the influence of points which add information that is not redundant with information added by other points of the training set
- $\sigma(-y\theta^\top x)$: The loss term is responsible for downweighting the influence of points which do not impact the predictions of the model

We will reproduce this experience... ?

2.3. Making influence function calculation tractable

Computing influence function pose a computational challenge for state-of-the-art models with typically millions of parameters. Specifically, forming the Hessian costs $O(np^2)$ and inverting it costs: $O(p^3)$. Thus, getting the hessian inverse in $\mathcal{I}_{\text{up,loss}}(z,z_{\text{test}}) = -\nabla_\theta \mathcal{L}(z_{\text{test}},\hat{\theta})^\top \mathcal{H}_{\hat{\theta}}^{-1}\mathcal{L}(z,\hat{\theta})$ costs $O(np^2+p^3)$.

The authors present two techniques for approximating $s_{\text{test}} = \mathcal{H}_{\hat{\theta}}^{-1}\nabla_\theta \mathcal{L}(z_{\text{test}},\hat{\theta})$; relying on a Hessian Vector Product technique introduced by Pearlmutter.

2.3.1. Conjugate gradients (CG)

The idea is to get $\mathcal{H}_{\hat{\theta}}^{-1}v$ by solving a minimisation problem. The function to minimise is define by $\mathcal{H}_{\hat{\theta}}^{-1}v = \arg\min_t \left(\frac{1}{2}t^\top \mathcal{H}_{\hat{\theta}} t - v^\top t\right)$. The CG approaches can iteratively gives the value of the solution by evaluate $\mathcal{H}_{\hat{\theta}}t$ in $O(np)$ p times with $\theta \in \mathbb{R}^p$. However CG approaches can give us a good estimation in less iterations than p.

2.3.2. Stochastic estimation

While CG requires to go through n training points per iteration, stochastic estimation only samples a single point per iteration, reducing the cost dramatically on large datasets. We use the Taylor expansion to approximate H^{-1}. In order to do that, we use the Taylor expansion of ln :

==========

2.4. Validation and extension

Influence functions are asymptotic approximation of the leave one out training with the underlying assumptions that the model parameters minimize the empirical risk and that the empirical risk is twice differentiable and convex.

The authors show by experiment that in logistic regression (GLM), the influence function does track change in loss for leave one out training really well .

Performance in non convex objectives is then studied and tested. The idea is to use quadratic approximation of the empiririca risk which will be smooth and convex, and apply influence function on the risk approximation function. The authors argue that the influence function track influence function even when they use a quadratic approximation of the empirical risk.

Theorems: a function is convex if and only if its hessian matrix is PSD contrapositive : non(convex) =¿ non(PSD) eigen value positive =¿ PSD contrapositive : non(PSD) =¿ non(eigen values positive) Hence : non(convex) =¿ non(eigen values positive)

We want to guarantee that the quadratic approximation of the loss is convex. It is convex iff the hessian matrix is PSD. We have no guarantee than the hessian of the loss is PSD, but the damping term allows us to guarantee that for Lambda large enough, (H + lambda * I) is PSD.

By experiment (fig2), influence function tracks change in loss for leave one out training in CNN (non GLM). We note that it tends to overestimate the influence (+ or -) of influential points and underestimate the influence of less influential points.

Performance non differentiable loss trick: smooth approximation

HINGE bad : cant get any information from the derivatives (1st and 2nd) SMOOTHHINGE -¿ better : approximate Hinge but get the information around non-derivative point of Hinge.

USE CASES

Understanding model behavior

While the inverse of pixel wise euclidean distance is meaningful in SVM to track influence of training points in the prediction of a given test point, it is less so in Inception (CNN). This is due to the fact that SVM will use points that are matching superficial patterns of the class while Inception matches distinctive characteristics of the class. For SVMs, helpful training point belong to the class of the test point and harmful training points belong to some other class. For Inception, helpful training point may be of a different class than that of the test point.

We will produce experiment to track the most influential training point in a similar prediction problems, across two distinct models.

Adversarial training examples

Debugging domain mismatch

Fixing mislabeled examples

==========

Papers must not exceed eight (8) pages, including all figures, tables, and appendices, but excluding references and acknowledgements. When references and acknowledgements are included, the paper must not exceed ten (10) pages. Acknowledgements should be limited to grants and people who contributed to the paper. Any submission that exceeds this page limit, or that diverges significantly from the specified format, will be rejected without review.

The text of the paper should be formatted in two columns, with an overall width of 6.75 inches, height of 9.0 inches, and 0.25 inches between the columns. The left margin should be 0.75 inches and the top margin 1.0 inch (2.54 cm). The right and bottom margins will depend on whether you print on US letter or A4 paper, but all final versions must be produced for US letter size.

The paper body should be set in 10 point type with a vertical spacing of 11 points. Please use Times typeface throughout the text.

2.5. Title

The paper title should be set in 14 point bold type and centered between two horizontal rules that are 1 point thick, with 1.0 inch between the top rule and the top edge of the page. Capitalize the first letter of content words and put the rest of the title in lower case.

2.6. Author Information for Submission

ICML uses double-blind review, so author information must not appear. If you are using LaTeX and the `icml2018.sty` file, use `\icmlauthor{...}` to specify authors and `\icmlaffiliation{...}` to specify affiliations. (Read the TeX code used to produce this document for an example usage.) The author information will not be printed unless `accepted` is passed as an argument to the style file. Submissions that include the author information will not be reviewed.

2.6.1. Self-Citations

If you are citing published papers for which you are an author, refer to yourself in the third person. In particular, do not use phrases that reveal your identity (e.g., "in previous work (Langley, 2000), we have shown ...").

Do not anonymize citations in the reference section. The only exception are manuscripts that are not yet published (e.g., under submission). If you choose to refer to such unpublished manuscripts (Author, 2018), anonymized copies have to be submitted as Supplementary Material via CMT. However, keep in mind that an ICML paper should be self

contained and should contain sufficient detail for the reviewers to evaluate the work. In particular, reviewers are not required to look at the Supplementary Material when writing their review.

2.6.2. Camera-Ready Author Information

If a paper is accepted, a final camera-ready copy must be prepared. For camera-ready papers, author information should start 0.3 inches below the bottom rule surrounding the title. The authors' names should appear in 10 point bold type, in a row, separated by white space, and centered. Author names should not be broken across lines. Unbolded superscripted numbers, starting 1, should be used to refer to affiliations.

Affiliations should be numbered in the order of appearance. A single footnote block of text should be used to list all the affiliations. (Academic affiliations should list Department, University, City, State/Region, Country. Similarly for industrial affiliations.)

Each distinct affiliations should be listed once. If an author has multiple affiliations, multiple superscripts should be placed after the name, separated by thin spaces. If the authors would like to highlight equal contribution by multiple first authors, those authors should have an asterisk placed after their name in superscript, and the term "*Equal contribution" should be placed in the footnote block ahead of the list of affiliations. A list of corresponding authors and their emails (in the format Full Name <email@domain.com>) can follow the list of affiliations. Ideally only one or two names should be listed.

A sample file with author names is included in the ICML2018 style file package. Turn on the `[accepted]` option to the stylefile to see the names rendered. All of the guidelines above are implemented by the LaTeX style file.

2.7. Abstract

The paper abstract should begin in the left column, 0.4 inches below the final address. The heading 'Abstract' should be centered, bold, and in 11 point type. The abstract body should use 10 point type, with a vertical spacing of 11 points, and should be indented 0.25 inches more than normal on left-hand and right-hand margins. Insert 0.4 inches of blank space after the body. Keep your abstract brief and self-contained, limiting it to one paragraph and roughly 4–6 sentences. Gross violations will require correction at the camera-ready phase.

2.8. Partitioning the Text

You should organize your paper into sections and paragraphs to help readers place a structure on the material and understand its contributions.

2.8.1. Sections and Subsections

Section headings should be numbered, flush left, and set in 11 pt bold type with the content words capitalized. Leave 0.25 inches of space before the heading and 0.15 inches after the heading.

Similarly, subsection headings should be numbered, flush left, and set in 10 pt bold type with the content words capitalized. Leave 0.2 inches of space before the heading and 0.13 inches afterward.

Finally, subsubsection headings should be numbered, flush left, and set in 10 pt small caps with the content words capitalized. Leave 0.18 inches of space before the heading and 0.1 inches after the heading.

Please use no more than three levels of headings.

2.8.2. Paragraphs and Footnotes

Within each section or subsection, you should further partition the paper into paragraphs. Do not indent the first line of a given paragraph, but insert a blank line between succeeding ones.

You can use footnotes[1] to provide readers with additional information about a topic without interrupting the flow of the paper. Indicate footnotes with a number in the text where the point is most relevant. Place the footnote in 9 point type at the bottom of the column in which it appears. Precede the first footnote in a column with a horizontal rule of 0.8 inches.[2]

2.9. Figures

You may want to include figures in the paper to illustrate your approach and results. Such artwork should be centered, legible, and separated from the text. Lines should be dark and at least 0.5 points thick for purposes of reproduction, and text should not appear on a gray background.

Label all distinct components of each figure. If the figure takes the form of a graph, then give a name for each axis and include a legend that briefly describes each curve. Do not include a title inside the figure; instead, the caption should serve this function.

Number figures sequentially, placing the figure number and caption *after* the graphics, with at least 0.1 inches of space before the caption and 0.1 inches after it, as in Figure 1. The figure caption should be set in 9 point type and centered unless it runs two or more lines, in which case it should be flush left. You may float figures to the top or bottom of a

[1] Footnotes should be complete sentences.

[2] Multiple footnotes can appear in each column, in the same order as they appear in the text, but spread them across columns and pages if possible.

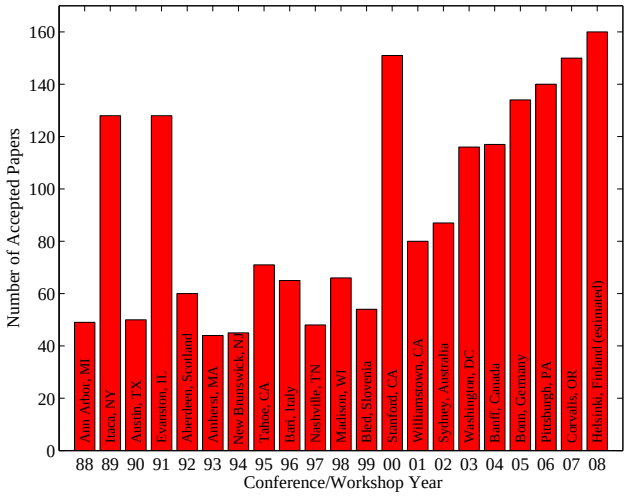

Figure 1. Historical locations and number of accepted papers for International Machine Learning Conferences (ICML 1993 – ICML 2008) and International Workshops on Machine Learning (ML 1988 – ML 1992). At the time this figure was produced, the number of accepted papers for ICML 2008 was unknown and instead estimated.

Algorithm 1 Bubble Sort

Input: data x_i, size m
repeat
  Initialize $noChange = true$.
  for $i = 1$ **to** $m - 1$ **do**
    if $x_i > x_{i+1}$ **then**
      Swap x_i and x_{i+1}
      $noChange = false$
    end if
  end for
until $noChange$ is $true$

column, and you may set wide figures across both columns (use the environment `figure*` in LaTeX). Always place two-column figures at the top or bottom of the page.

2.10. Algorithms

If you are using LaTeX, please use the "algorithm" and "algorithmic" environments to format pseudocode. These require the corresponding stylefiles, algorithm.sty and algorithmic.sty, which are supplied with this package. Algorithm 1 shows an example.

2.11. Tables

You may also want to include tables that summarize material. Like figures, these should be centered, legible, and numbered consecutively. However, place the title *above* the table with at least 0.1 inches of space before the title and the same after it, as in Table 1. The table title should be set in 9 point type and centered unless it runs two or more lines, in which case it should be flush left.

Table 1. Classification accuracies for naive Bayes and flexible Bayes on various data sets.

DATA SET	NAIVE	FLEXIBLE	BETTER?
BREAST	95.9± 0.2	96.7± 0.2	√
CLEVELAND	83.3± 0.6	80.0± 0.6	×
GLASS2	61.9± 1.4	83.8± 0.7	√
CREDIT	74.8± 0.5	78.3± 0.6	
HORSE	73.3± 0.9	69.7± 1.0	×
META	67.1± 0.6	76.5± 0.5	√
PIMA	75.1± 0.6	73.9± 0.5	
VEHICLE	44.9± 0.6	61.5± 0.4	√

Tables contain textual material, whereas figures contain graphical material. Specify the contents of each row and column in the table's topmost row. Again, you may float tables to a column's top or bottom, and set wide tables across both columns. Place two-column tables at the top or bottom of the page.

2.12. Citations and References

Please use APA reference format regardless of your formatter or word processor. If you rely on the LaTeX bibliographic facility, use `natbib.sty` and `icml2018.bst` included in the style-file package to obtain this format.

Citations within the text should include the authors' last names and year. If the authors' names are included in the sentence, place only the year in parentheses, for example when referencing Arthur Samuel's pioneering work (1959). Otherwise place the entire reference in parentheses with the authors and year separated by a comma (Samuel, 1959). List multiple references separated by semicolons (Kearns, 1989; Samuel, 1959; Mitchell, 1980). Use the 'et al.' construct only for citations with three or more authors or after listing all authors to a publication in an earlier reference (Michalski et al., 1983).

Authors should cite their own work in the third person in the initial version of their paper submitted for blind review. Please refer to Section 2.6 for detailed instructions on how to cite your own papers.

Use an unnumbered first-level section heading for the references, and use a hanging indent style, with the first line of the reference flush against the left margin and subsequent lines indented by 10 points. The references at the end of this document give examples for journal articles (Samuel, 1959), conference publications (Langley, 2000), book chapters (Newell & Rosenbloom, 1981), books (Duda et al.,

2000), edited volumes (Michalski et al., 1983), technical reports (Mitchell, 1980), and dissertations (Kearns, 1989).

Alphabetize references by the surnames of the first authors, with single author entries preceding multiple author entries. Order references for the same authors by year of publication, with the earliest first. Make sure that each reference includes all relevant information (e.g., page numbers).

Please put some effort into making references complete, presentable, and consistent. If using bibtex, please protect capital letters of names and abbreviations in titles, for example, use {B}ayesian or {L}ipschitz in your .bib file.

2.13. Software and Data

We strongly encourage the publication of software and data with the camera-ready version of the paper whenever appropriate. This can be done by including a URL in the camera-ready copy. However, do not include URLs that reveal your institution or identity in your submission for review. Instead, provide an anonymous URL or upload the material as "Supplementary Material" into the CMT reviewing system. Note that reviewers are not required to look a this material when writing their review.

Acknowledgements

Do not include acknowledgements in the initial version of the paper submitted for blind review.

If a paper is accepted, the final camera-ready version can (and probably should) include acknowledgements. In this case, please place such acknowledgements in an unnumbered section at the end of the paper. Typically, this will include thanks to reviewers who gave useful comments, to colleagues who contributed to the ideas, and to funding agencies and corporate sponsors that provided financial support.

References

Author, N. N. Suppressed for anonymity, 2018.

Duda, R. O., Hart, P. E., and Stork, D. G. *Pattern Classification*. John Wiley and Sons, 2nd edition, 2000.

Kearns, M. J. *Computational Complexity of Machine Learning*. PhD thesis, Department of Computer Science, Harvard University, 1989.

Langley, P. Crafting papers on machine learning. In Langley, Pat (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Michalski, R. S., Carbonell, J. G., and Mitchell, T. M. (eds.). *Machine Learning: An Artificial Intelligence Approach, Vol. I*. Tioga, Palo Alto, CA, 1983.

Mitchell, T. M. The need for biases in learning generalizations. Technical report, Computer Science Department, Rutgers University, New Brunswick, MA, 1980.

Newell, A. and Rosenbloom, P. S. Mechanisms of skill acquisition and the law of practice. In Anderson, J. R. (ed.), *Cognitive Skills and Their Acquisition*, chapter 1, pp. 1–51. Lawrence Erlbaum Associates, Inc., Hillsdale, NJ, 1981.

Samuel, A. L. Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, 3(3):211–229, 1959.

A. Do *not* have an appendix here

Do not put content after the references. Put anything that you might normally include after the references in a separate supplementary file.

We recommend that you build supplementary material in a separate document. If you must create one PDF and cut it up, please be careful to use a tool that doesn't alter the margins, and that doesn't aggressively rewrite the PDF file. pdftk usually works fine.

Please do not use Apple's preview to cut off supplementary material. In previous years it has altered margins, and created headaches at the camera-ready stage.