Assessment of Case Influence in Support Vector Machine

Yoonkyung Lee*
Department of Statistics
The Ohio State University
*joint work with Shanshan Tu and Yunzhang Zhu

November 25, 2020
CIMAT
Guanajuato, Mexico

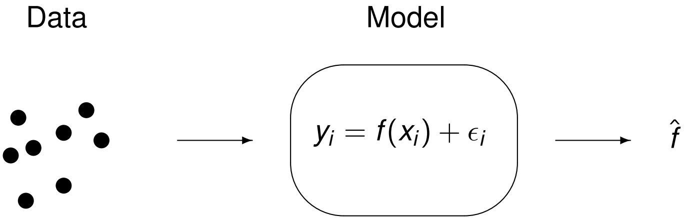
A Trip Down Memory Lane

March 2015

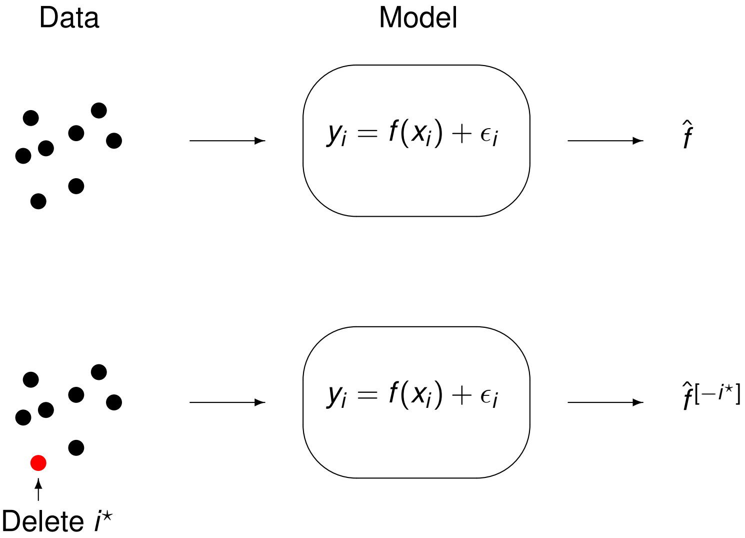
Introduction

- Stability and robustness is desired for modeling (or prediction) procedures
- Given a modeling procedure, how sensitive is the fitted model to some change in the data?
- How much does the model change if a case is deleted?
- Connected to privacy-preserving data analysis and adversarial machine learning

Modeling Process



Case Influence in Case Deletion Scheme



Overview

- Case deletion is considered for
  - model assessment (e.g. regression diagnostics)
  - model selection (e.g. leave-one-out CV)
  - measuring model complexity (model df)
- Extensively studied for mean regression with squared error loss (e.g. Cook's distance)
- Generalize the ideas of case influence to classification

Case Influence in Linear Regression

- Cook's distance for case $i^\star$ (Cook, 1977):

$$D_{i^\star} = \frac{1}{p\hat{\sigma}^2} \sum_{i=1}^{n} \left(\hat{f}(x_i) - \hat{f}^{[-i^\star]}(x_i) \right)^2$$

TECHNOMETRICS©, VOL. 19, NO. 1, FEBRUARY 1977

Detection of Influential Observation in Linear Regression

R. Dennis Cook

Department of Applied Statistics
University of Minnesota
St. Paul, Minnesota 55108

A new measure based on confidence ellipsoids is developed for judging the contribution of each data point to the determination of the least squares estimate of the parameter vector in full rank linear regression models. It is shown that the measure combines information from the studentized residuals and the variances of the residuals and predicted values. Two examples are presented.

Case Influence in Linear Regression

- Cook's distance for case $i^\star$ (Cook, 1977):

$$D_{i^\star} = \frac{1}{p\hat{\sigma}^2} \sum_{i=1}^{n} \left(\hat{f}(x_i) - \hat{f}^{[-i^\star]}(x_i) \right)^2$$

- It can be expressed using the residual and leverage:

$$D_{i^\star} = \frac{1}{p\hat{\sigma}^2} \left[\frac{h_{i^\star}}{(1 - h_{i^\star})^2} \right] r_{i^\star}^2,$$

where $r_{i^\star} = y_{i^\star} - \hat{f}(x_{i^\star})$ and $h_{i^\star}$ is the leverage of case $i^\star$ ($i^\star$th diagonal entry of the hat matrix $X(X^\top X)^{-1} X^\top$)

Support Vector Machine

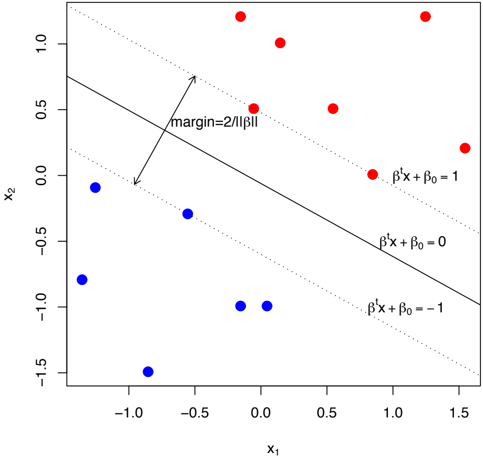
Vapnik (1996), *The Nature of Statistical Learning Theory*

$$y_i = \begin{cases} 1 & \text{for class 1} \\ -1 & \text{for class 2} \end{cases}$$

and $\phi(x) = \text{sign}(f(x))$.

Find $f(x) = \beta_0 + \beta^\top x$ with a large margin by minimizing

$$\sum_{i=1}^{n} (1 - y_i f(x_i))_+ + \frac{\lambda}{2}\|\beta\|^2.$$



Review of Support Vector Machine

- The solution as a discriminant function is shown to be of the form:

$$\hat{f}(x) = a + \frac{1}{\lambda}\sum_{i=1}^{n} \alpha_i y_i (x_i^\top x)$$

- The coefficients α_i are determined by solving a quadratic programming problem

- According to the optimality conditions, if $y_i \hat{f}(x_i) > 1$, $\hat{\alpha}_i = 0$

- If $\hat{\alpha}_i = 0$, then $\hat{f}^{[-i]} = \hat{f}$

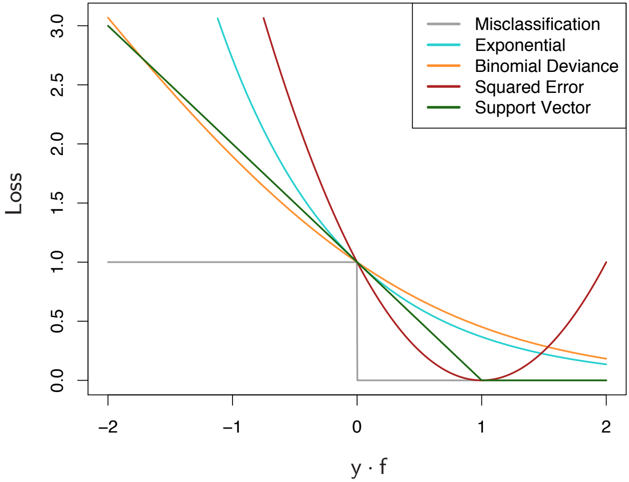
- Data points with $\hat{\alpha}_i > 0$ are called support vectors

Margin-Based Loss Function

For a real-valued discriminant function $f(x)$ which induces the rule $\phi(x) = \text{sign}(f(x))$,

- Misclassification (0-1): $I(yf(x) \leq 0)$
- SVM (hinge) : $(1 - yf(x))_+$
- Logistic regression (binomial deviance): $\log(1 + \exp(-yf(x)))$
- Boosting (exponential): $\exp(-yf(x))$

Loss Function



source: Hastie, Tibshirani & Friedman (2009)

Challenges in Extension to Classification

- Extension of Cook's distance appropriate for margin-based classification?
- How to calculate the leave-one-out (LOO) solution $\hat{f}^{[-i]}$ for $i = 1, \cdots, n$?

Case Influence Measure

- *Classification discrepancy rate*:

$$\begin{aligned} CD_{i^\star} &= \frac{1}{n}\sum_{i=1}^{n}\left| I(y_i\hat{f}(x_i) < 0) - I(y_i\hat{f}^{[-i^\star]}(x_i) < 0)\right| \\ &= \frac{1}{n}\sum_{i=1}^{n} I(\hat{f}(x_i)\hat{f}^{[-i^\star]}(x_i) < 0) \end{aligned}$$

- *Functional margin difference*:

$$MD_{i^\star} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i\hat{f}(x_i) - y_i\hat{f}^{[-i^\star]}(x_i)\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{f}(x_i) - \hat{f}^{[-i^\star]}(x_i)\right)^2$$

- *Loss difference*:

$$LD_{i^\star} = \frac{1}{n}\sum_{i=1}^{n}\left(L(y_i\hat{f}(x_i)) - L(y_i\hat{f}^{[-i^\star]}(x_i))\right)^2$$

Computation for Case Deletion

- Can we calculate the leave-one-out (LOO) solution $\hat{f}^{[-i]}$ efficiently from the full data solution $\hat{f}$?
- Take $\hat{f}$ as a warm start?
- Using a homotopy technique, examine the link between the two solutions by a case-weight adjusted solution path

Case-Weight Adjusted SVM

- For each case $i^\star$, consider minimizing

$$\sum_{i \neq i^\star} (1 - y_i(\beta_0 + x_i^\top \beta))_+ + \underbrace{\omega\,(1 - y_{i^\star}(\beta_0 + x_{i^\star}^\top \beta))_+}_{\text{weight-adjusted}} + \frac{\lambda}{2}\|\beta\|^2,$$

  with a case weight $\omega \in [0, 1]$

- Treat the case weight ω as a homotopy parameter linking the full data solution to the leave-one-out (LOO) solution:

Full data solution $\hat{f}$ $\Longrightarrow$ LOO solution $\hat{f}^{[-i^\star]}$

$\omega = 1$ ——— $\omega = 0$

Constrained Optimization

- Express the hinge loss with slack variable ξ:

$$(1 - yf)_+ = \begin{cases} \min_\xi & \xi \\ \text{s.t.} & 1 - yf \le \xi \\ & \xi \ge 0 \end{cases}$$

- The SVM problem can be formulated as a constrained optimization with linear inequalities
- The KKT optimality conditions can be derived for the solution $(\beta_{0,\omega}, \beta_\omega)$ given case weight ω for each $i^\star$
- Representation of the discriminant function:

$$f_\omega(x) = a_\omega + \frac{1}{\lambda} \sum_{i=1}^{n} \theta_{i,\omega} y_i (x_i^\top x)$$

Optimality Conditions

- The KKT conditions with dual variables $\theta_{i,\omega}$, $i = 1, \ldots, n$:

$$\sum_{i=1}^{n} \theta_{i,\omega} y_i = 0$$

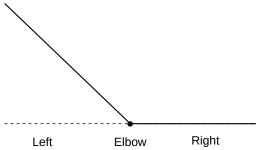
$$\theta_{i,\omega} = 0 \qquad \text{if} \quad y_i(\beta_{0,\omega} + x_i^\top \beta_\omega) > 1$$

$$\theta_{i,\omega} \in \begin{cases} [0,1], & \text{for } i \neq i^\star \\ [0,\omega], & \text{for } i = i^\star \end{cases} \qquad \text{if} \quad y_i(\beta_{0,\omega} + x_i^\top \beta_\omega) = 1$$

$$\theta_{i,\omega} = \begin{cases} 1, & \text{for } i \neq i^\star \\ \omega, & \text{for } i = i^\star \end{cases} \qquad \text{if} \quad y_i(\beta_{0,\omega} + x_i^\top \beta_\omega) < 1$$

- According to the margin $y_i f_\omega(x_i) = y_i(\beta_{0,\omega} + x_i^\top \beta_\omega)$, cases are categorized into

$\mathcal{R}_\omega = \{i : y_i f_\omega(x_i) > 1\}$ (right)
$\mathcal{E}_\omega = \{i : y_i f_\omega(x_i) = 1\}$ (elbow)
$\mathcal{L}_\omega = \{i : y_i f_\omega(x_i) < 1\}$ (left)



Piecewise Linearity of Solution Path

Proposition

The solution path $(a_\omega, \theta_\omega)$ satisfying the KKT conditions is piecewise linear in case weight ω.

In particular, for $\omega_{m+1} < \omega < \omega_m$, if $y_{i^\star} f_{\omega_m}(x_{i^\star}) \geq 1$, $(a_\omega, \theta_\omega)$ is constant; otherwise, $(a_\omega, \theta_\omega)$ changes linearly.

Corollary

The slope of $y_i f_\omega(x_i)$ on $[\omega_{m+1}, \omega_m)$ for $i = 1, \cdots, n$ is 0 if $y_{i^\star} f_{\omega_m}(x_{i^\star}) \geq 1$, and nonzero constant otherwise.

Monotonicity of Functional Margin Path

Proposition

The functional margin of the weighted case, $y_{i^\star} f_\omega(x_{i^\star})$, is piecewise linear and nondecreasing in ω.

Remark

This result is analogous to the monotonicity of a residual in case weight in regression.

Path-Following Algorithm

- Similar to the results for SVM and kernel QR solution paths (Rosset and Zhu 2007, Li et al. 2007)
- Devise a path-following algorithm
- By tracking changes in the three sets with ω, we can identify breakpoints $0 \leq \omega_M < \cdots < \omega_1 < \omega_0 = 1$ and corresponding solutions.
- Can generate the entire case-weight adjusted solution path as ω decreases from 1 to 0

Example:

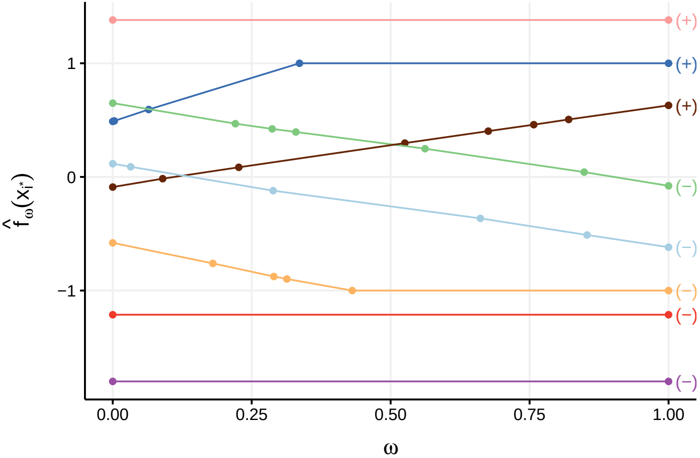
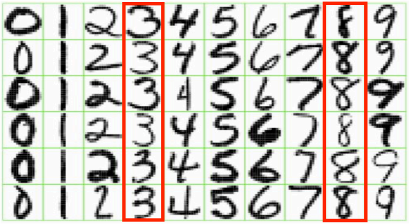


Figure: The discriminant score path $\hat{f}_\omega(x_{i^\star})$ for SVM with radial kernel starting from the original full-data fit at $\omega = 1$ to the fit at $\omega = 0$ when case $i^\star$ is removed.

Example: Detection of Mislabeled Cases



- Subset 100 cases of digits 3 and 8 from handwritten digit data (Le Cun et al. 1990)
- Randomly flip the class labels for 10% of the cases for each digit
- Rank cases according to influence measures for SVM in case deletion scheme to detect mislabeled cases

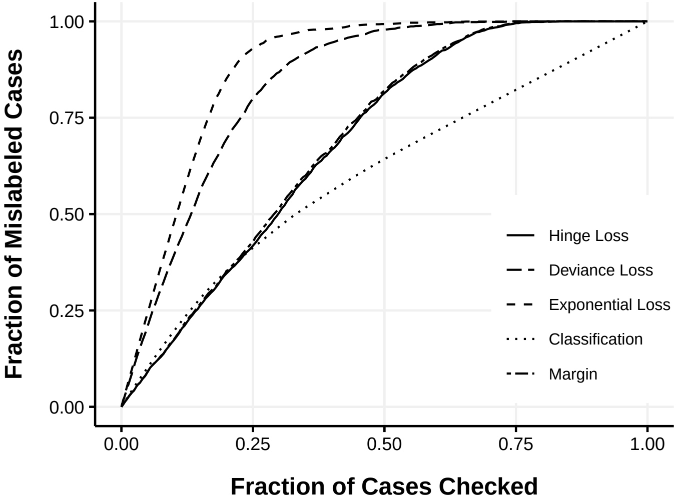


Figure: Operating characteristic curves of five influence measures in detecting mislabeled cases.

Case Influence Graph

- Case-weight adjusted Cook's distance (Cook, 1986):

$$D_{i^\star}(\omega) = \frac{\sum_{i=1}^{n}(\hat{f}(x_i) - \hat{f}_{\omega}^{i^\star}(x_i))^2}{p\hat{\sigma}^2}$$

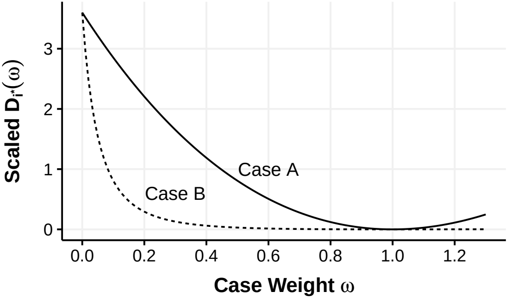


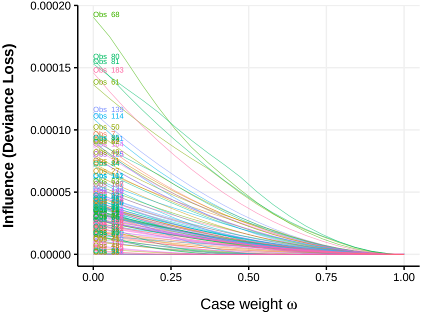
Figure: An illustrative example of case-influence graphs in least squares regression based on Figure 1 in Cook (1986)

Case Influence Graph for SVM

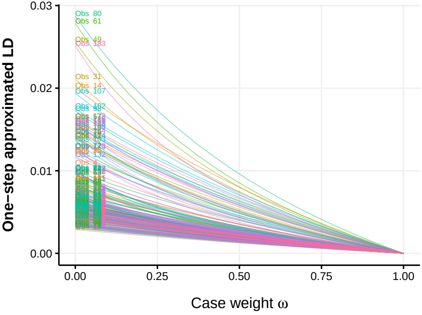
- Case-weight adjusted loss difference:

$$M_{i^\star}(\omega) = \frac{1}{n}\sum_{i=1}^{n}\left(L(y_i\hat{f}(x_i)) - L(y_i\hat{f}^{i^\star}_{\omega}(x_i))\right)^2$$

- Area under the influence graph as an alternative measure



(a) Linear SVM



(b) Logistic regression

Figure: Comparison of case influences in linear SVM and logistic regression defined as the (square rooted) area under the case influence graph. Red circles represent the mislabeled cases.

Global and Local Influence Measures

- Global influence:

$$G_{i^\star} = \int_0^1 M_{i^\star}(\omega)d\omega$$

- Local influence (Cook, 1986):
  the curvature of the case influence graph at $\omega = 1$

$$\ell_{i^\star} = \left.\frac{\partial^2 M_{i^\star,\omega}}{\partial \omega^2}\right|_{\omega=1}$$

- Since $M_{i^\star,\omega} = 0$ at $\omega = 1$, if $\left.\frac{\partial M_{i^\star,\omega}}{\partial \omega}\right|_{\omega=1} = 0$, then the local influence provides a quadratic approximation to $G_{i^\star}$

Local Influence

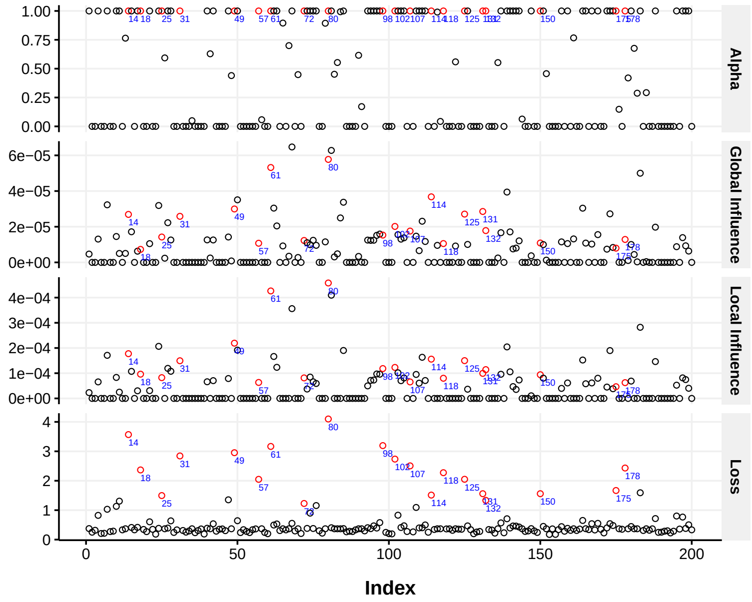
Lemma

For each $i = 1, \cdots, n$, the rate of change in the discriminant score at $\omega = 1$, $\frac{\partial \hat{f}_{\omega}^{i^\star}(x_i)}{\partial \omega}\Big|_{\omega=1}$, in SVM is 0, if the functional margin of the weighted case, $y_{i^\star}\hat{f}(x_{i^\star}) \geq 1$; otherwise, obtained explicitly.

Proposition

Let $M_{i^\star,\omega}$ be the case-weight adjusted loss difference with continuously differentiable loss $L(\cdot)$. Then the local influence $\ell_{i^\star}$ of each case $i^\star \in \{1, \cdots, n\}$ in SVM is 0 if $y_{i^\star}\hat{f}(x_{i^\star}) \geq 1$; otherwise,

$$\ell_{i^\star} = \frac{\partial^2 M_{i_\star,\omega}}{\partial \omega^2}\Bigg|_{\omega=1} = \frac{2}{n}\sum_{i=1}^{n}\left(L'(y_i\hat{f}(x_i))\right)^2 \cdot \left(\frac{\partial \hat{f}_{\omega}^{i^\star}(x_i)}{\partial \omega}\Bigg|_{\omega=1}\right)^2.$$



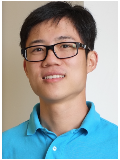
Alpha
Global Influence
Local Influence
Loss
1.00
0.75
0.50
0.25
0.00
6e−05
4e−05
2e−05
0e+00
4e−04
3e−04
2e−04
1e−04
0e+00
4
3
2
1
0
0
50
100
150
200
Index

Remarks

- Extended case influence statistics for SVM
- Presented a homotopy method for a case-weight adjusted solution path that connects the full data solution to LOO solutions for SVM
- How to extend the framework for case influence assessment to other classification methods (e.g. boosting)?
- How to define model complexity in classification using the notion of case sensitivity?

Acknowledgments

Shanshan Tu
@ Argo AI



Yunzhang Zhu
@ Statistics, OSU

DMS-15-13566
DMS-20-15490

References

Dennis Cook.
Detection of influential observations in linear regression.
Technometrics, 19(1):15 – 18, 1977.

Dennis Cook.
Assessment of local influence.
Journal of the Royal Statistical Society. Series B (Methodological), 48(2):133 – 169, 1986.

Youjuan Li, Yufeng Liu, and Ji Zhu.
Quantile regression in reproducing kernel Hilbert spaces.
Journal of American Statistical Association, 102(477):255 – 268, March 2007.

Daryl Pregibon.
Logistic regression diagnostics.
The Annals of Statistics, 9(4):705–724, 1981.

Saharon Rosset and Ji Zhu.
Piecewise linear regularized solution paths.
The Annals of Statistics, 35(3):1012 – 1030, 2007.

Shanshan Tu.
Case Influence and Model Complexity in Regression and Classification.
PhD thesis, The Ohio State University, 2019.

Grace Wahba.
Spline Models for Observational Data.
The Society for Industrial and Applied Mathematics, 1990.

Jianming Ye.
On measuring and correcting the effects of data mining and model selection.
Journal of American Statistical Association, 93(441):120 – 131, 1998.