

New results on the robustness of convex risk minimization methods

Andreas Christmann

University of Dortmund, Department of Statistics, 44221 Dortmund,
GERMANY.

christmann@statistik.uni-dortmund.de

Keywords: Influence function, Robustness, Kernel logistic regression, Support Vector Machine, AdaBoost.

Abstract

Kernel logistic regression, support vector machine and AdaBoost belong to modern statistical machine learning methods, c.f. Vapnik (1998), Friedman, Hastie, and Tibshirani (2000), and Schölkopf and Smola (2002). Such methods are useful in analyzing high dimensional complex data sets and can fit even complicated dependency structures between variables in an automatic way by using non-linear kernels.

Consider a training data set $\{(\mathbf{x}_i, y_i) \in \mathbb{R}^p \times \{-1, +1\}, i = 1, \dots, n\}$. The goal is to learn from the training data set such that good predictions $\hat{y} = \text{sign}(\hat{f}(x) + \hat{b})$ can be made for the response y given a vector \mathbf{x} , where the function \hat{f} and \hat{b} are estimated from the training data set. Logistic regression, which is a parametric model, is often used for this problem. However, the classical maximum likelihood estimator in this model has two serious drawbacks: it is non-robust and it does not exist for all data sets, see Rousseeuw and Christmann (2003) for an alternative model and a robust estimation method.

The assumption in statistical machine learning is much weaker, because one assumes that the observations (x_i, y_i) from the training data set are independent and identically generated from an underlying unknown distribution \mathbb{P} for a pair of random variables (X_i, Y_i) , where \mathbb{P} is fully unspecified. The quality of the predictor $f(x_i) + b$ is measured by some convex loss function $L(y_i, f(x_i) + b)$. Consider the following empirical regularized risk:

$$(\hat{f}_{n,\lambda}, \hat{b}_{n,\lambda}) = \arg \min_{f \in H, b \in \mathbb{R}} \lambda \|f\|_H^2 + \frac{1}{n} \sum_{i=1}^n L(y_i, f(x_i) + b), \quad (1)$$

where $\lambda > 0$ is a small regularization parameter, H is a reproducing kernel Hilbert space (RKHS) of a kernel k , and b is an unknown real-valued offset. The decision function is $\text{sign}(\hat{f}_{n,\lambda} + \hat{b}_{n,\lambda})$. The choice of the kernel k enables the above methods to efficiently estimate not only linear, but also non-linear functions. Of special importance is the Gaussian radial basis function (RBF) kernel given by $k(x, x') = \exp(-\gamma \|x - x'\|^2)$, $\gamma > 0$, which is a universal kernel on every compact subset of \mathbb{R}^d . Popular loss functions

depend on y and $f + b$ via $v = y(f(x) + b)$. Some popular specifications of L are: kernel logistic regression based on $L(v) = \ln(1 + \exp(-v))$, AdaBoost based on $L(v) = \exp(-v)$, and the support vector machine based on $L(v) = \max(1 - v, 0)$.

Problem (1) can be interpreted as a stochastic approximation of the minimization of the theoretical regularized risk defined by

$$(f_{\mathbb{P},\lambda}, b_{\mathbb{P},\lambda}) = \arg \min_{f \in H, b \in \mathbb{R}} \lambda \|f\|_H^2 + \mathbb{E}_{\mathbb{P}} L(Y, f(X) + b). \quad (2)$$

Robustness properties of such convex risk minimization methods have not yet received much attention. Christmann and Steinwart (2003) showed that F.R. Hampel's influence function exists for some of these convex risk minimization methods provided that all explanatory variables \mathbf{x}_i belong to a compact set $\mathbf{X} \subset \mathbb{R}^p$.

In the present talk, it will be shown that the influence functions of a broad class of convex risk minimization methods exist under much weaker assumptions. Further, bounds on the influence function, J.W. Tukey's sensitivity curve and the maxbias will be given. Therefore, such convex risk minimization methods have reasonable robustness properties.

References

- A. Christmann, I. Steinwart** (2003). *On robust properties of convex risk minimization methods for pattern recognition*. University of Dortmund, SFB-475, Technical Report 15/2003.
- J. Friedman, T. Hastie, R. Tibshirani** (2000). Additive logistic regression: a statistical view of boosting (with discussion). *Ann. Statist.*, 28, 337-407.
- F.R. Hampel, E.M. Ronchetti, P.J. Rousseeuw, W.A. Stahel** (1986). *Robust Statistics: the Approach based on Influence Functions*. Wiley & Sons, New York.
- P.J. Rousseeuw, A. Christmann** (2003). Robustness against separation and outliers in logistic regression. *Computational Statistics & Data Analysis*, 43, 315-332.
- B. Schölkopf, A.J. Smola** (2002). *Learning with Kernels. Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge.
- V. Vapnik** (1998). *Statistical Learning Theory*. Wiley & Sons, New York.