Bootstrap the Support Vector Machine

Thomas Goerttler, Christian Koopmann, Patricia Craja

Humboldt-University of Berlin

Introduction

Support Vector Machines are one of the most successful methods of Machine Learning. By reducing non-linear complex decisions problems to linear problems through application of the Kernel-Trick, they represent a computationally efficient way to tackle these problems. SVMs based on certain kernels (e.g. Gaussian RBF Kernel) are non parametric methods. Since the distribution of the underlying data is generally unknown so is the finite sample distribution of these methods. There has been considerable research on the asymptotic distribution of SVMs, which have been shown to be asymptotically normally distributed under certain conditions. An alternative idea to estimate these distributions is using Efrons empirical bootstrap. The idea behind this method is to repeatedly draw samples with replacement from the full data according to the empirical distribution function of the data. Through the repeated calculation of the statistic of interest one can get an estimate of its distribution. For the SVM this estimate has been shown to be consistent under relatively mild conditions.

Goal

The goal of this project is to apply the bootstrap method to the SVM algorithm to estimate the uncertainty of its predictions and the way in which this uncertainty is correlated with other aspects of the SVM (in particular the number of support vectors). This will be done using both simulated as well as real datasets. On simulated datasets the resulting confidence intervals can be compared to known actual values, according to the distribution used for simulation, whereas in the case of real data it will be compared to asymptotic results. The project will be implemented in Python using the Liblinear-Algorithm for training individual SVMs, and applying this algorithm to different Bootstrap samples in a parallelized manner. The results will be summarized in a short paper.

Theorie

Hierzu sollte auch etwas stehen.

Eine	Tabelle	kann
auch	hilfreich	sein

Theorie

Formeln macht man so:

$$\int_a^b f(x)\,dx \approx (b-a)\frac{f(a)+f(b)}{2}$$

Umsetzung

Hier wird das Vorgehen erklärt:

- ...
- ...
- ...

Umsetzung

Ein Algorithmus zur Lösung des Problems:

1. Wähle Startwerte für die Parameter.
2. Fülle die fehlenden Daten auf.
3. Berechne über die aufgefüllten Daten neue Parameterwerte.
4. Führe Schritte 2 und 3 bis zur Konvergenz aus.

(Simulations-) Ergebnisse

So schreibt man **fett**.

(Simulations-) Ergebnisse

Kursiv geht auch

Fazit

- Erkennisse
- Schlussfolgerungen
- Ausblick.
-

- Burgard, J.P.; Münnich, R. (2010): Modelling over and undercounts for design-based Monte Carlo studies in small area estimation: An application to the German register-assisted census. Computational Statistics and Data Analysis.
- Gabler, S,; Ganninger,M.; Münnich, R. (2010): Optimal allocation of the sample size to strata under box constraints. Metrika.
- Gelman, A.; (2007): Struggles with Survey Weighting and Regression Modeling. Statistical Science.
- Alfons, A.; Filzmoser, P.; Hulliger, B.,Kolb, J.P.; Kraft, S.; Münnich, R. und Templ, M. (2011): The AMELI simulation study. Research Project Report WP6 - D6.1, FP7-SSH-2007-217322 AMELI.
- Alfons, A.; Filzmoser, P.; Hulliger, B., Kolb, J.-P.; Kraft, S.; Münnich, R., und Templ, M. (2011): Synthetic data generation of SILC Data. Research Project Report WP6 - D6.2, FP7-SSH-2007-217322 AMELI.

For further informationen

Contact:

Name	**E-Mail**
Thomas Goerttler	thomas.goerttler@gmail.com
Christian Koopmann	c.k.e.koopmann@gmail.com
Patricia Craja	Patricia.craja@gmx.de