Rank correlation estimators and their limiting distributions

Wojciech Rejchel

Nicolaus Copernicus University, Toruń, Poland

Abstract

We consider the following rank correlation problem: using independent copies of vectors (X, Y) and (X', Y'), where $X, X' \in \mathbf{R}^d$ and $Y, Y' \in \mathbf{R}$, we want to predict, based on observations X and X', whether $Y < Y'$ or $Y > Y'$ with maximal accuracy, i.e. we look for the ranking rule ϕ that maximizes

$$E\,\mathbf{I}(Y < Y')\mathbf{I}(\phi(X) < \phi(X')). \tag{1}$$

We restrict to linear rules $\phi(z) = \theta^T z, \theta \in \mathbf{R}^d$. Han (1987) showed that a maximizer of the sample analogue of (1) is a consistent and asymptotically normal estimator of an unknown parameter. However, with respect to θ, the function $\mathbf{I}(y < y')\mathbf{I}(\theta^T x < \theta^T x')$ is not continuous, so calculating this estimator is computationally difficult. We propose (analogously to support vector machines in the classification theory; Vapnik, 1998) finding a minimizer of the following convex, with respect to θ, function

$$\frac{1}{n(n-1)} \sum_{i \neq j} \mathbf{I}(Y_i > Y_j)\, \psi(\theta^T X_i - \theta^T X_j), \tag{2}$$

where $\psi(z) = (1 - z)_+$. There exist effective algorithms that allowed us to compute the minimizer of (2) (Bobrowski and Niemiro, 1984). We show strong consistency and asymptotic normality of the new estimator.

Keywords

Rank correlation problem, Linear ranking rule, Discontinuous criterion function, Support vector machines, Convex minimization, U-statistics.

References

Bobrowski, L. and W. Niemiro (1984). A method of synthesis of linear discriminant function in the case of nonseparability. *Pattern Reco. 17*, 205–210.

Han, A.K. (1987). Non-parametric analysis of a generalized regression model. *J. of Econometrics 35*, 303–316.

Vapnik, V.N. (1998). *Statistical Learning Theory.* New York: J. Wiley.