Large Deviation Bound for a General Form of Ranking Problem

Proposed by: Shivani Agarwal

Problem Area: Machine learning

Tools: Probability, statistics

Motivation. Consider the following example. Alice is shown a finite number of movies $x_1, \ldots, x_n$, each drawn from some set X, and for each pair of movies (x_i, x_j) $(i < j)$, she is asked to provide a preference label $r_{ij} \in \mathbb{R}$ indicating her preference for movie x_i over x_j (with $r_{ij} > 0$ indicating a preference for movie x_i; $r_{ij} < 0$ indicating a preference for movie x_j; and $r_{ij} = 0$ indicating an equal liking for the two movies). Say we build a ranking or scoring function $f : X \to \mathbb{R}$ to predict Alice's preferences on new movies in X: the higher the score assigned by f to a movie x, the greater we predict that Alice will like that movie. Then given a new pair of movies $(x, x') \in X \times X$ for which Alice's true preference would be r, the error or loss we incur in predicting her preference for this pair of movies using f can be written as

$$\ell(f, (x, x', r)) \;=\; \mathbf{1}_{(r(f(x)-f(x'))<0)} + \frac{1}{2}\mathbf{1}_{(f(x)=f(x'))}\mathbf{1}_{(r\neq 0)}\,, \tag{1}$$

where $\mathbf{1}_{(\phi)}$ is 1 if ϕ is true and 0 otherwise. In order to design a good ranking function f, we need to estimate the error incurred by f on new movies. Under suitable probabilistic assumptions, we would like to find out whether we can use the observed or empirical error of f on the movies for which Alice has already provided preferences, to estimate the error of f in predicting her preferences on new movies.

Problem. More specifically, the problem we are interested in can be described as follows. There is an instance space X from which instances are drawn, and a set $R \subseteq \mathbb{R}$ from which pair-wise preference labels are drawn. Assume n instances $x_1, \ldots, x_n$ are drawn randomly and independently according to some (unknown) distribution $\mathcal{D}$ on X, and that for each pair (x_i, x_j) $(i < j)$, a preference label r_{ij} is drawn according to some (unknown) conditional distribution $\mathcal{Q}$ on R. The empirical error of a ranking function $f : X \to \mathbb{R}$ on these examples is given by

$$\widehat{L}_n(f) \;=\; \frac{1}{\binom{n}{2}} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \ell(f, (x_i, x_j, r_{ij}))\,. \tag{2}$$

The expected error of f on a new pair of examples drawn according to the same distributions is given by

$$L(f) \;=\; \mathbf{E}_{(x,x')\sim\mathcal{D}\times\mathcal{D}}\left[\mathbf{E}_{r\sim\mathcal{Q}(\cdot|(x,x'))}\left[\ell(f, (x, x', r))\right]\right]\,. \tag{3}$$

Can we estimate the expected error $L(f)$ based on the empirical error $\widehat{L}_n(f)$? In particular, can we derive a large deviation bound that provides an upper bound on the following probability (for $\epsilon > 0$):

$$\mathbf{P}\left(\left|\widehat{L}_n(f) - L(f)\right| > \epsilon\right)\,. \tag{4}$$

Impact. Such a bound can be useful in establishing generalization results for ranking algorithms in machine learning under more general probabilistic assumptions than those used to derive previous results.

Special Cases. It could be of interest to start with $R = \{-1, 0, 1\}$.

Previous Results. Large deviation bounds of the above type are known under the following probabilistic assumptions:

1. When the preference examples provided by Alice are of the same form as above, namely movies $x_1, \ldots, x_n$ together with pair-wise preferences r_{ij} for each $i < j$, but with r_{ij} given by
$$r_{ij} \; = \; y_i - y_j \, ,$$
where y_i denotes a 'relevance rating' (e.g. 1 to 5 stars) for movie x_i, and where the movie-rating pairs (x_i, y_i) are drawn randomly and independently according to a joint distribution $\mathcal{D}$ on $X \times Y$, where $Y \subseteq \mathbb{R}$ is the set of possible relevance ratings. See [1, 2].

2. When the preference examples provided by Alice are of the same form as above, namely movies $x_1, \ldots, x_n$ together with pair-wise preferences r_{ij} for each $i < j$, but with r_{ij} given by
$$r_{ij} \; = \; \pi(x_i, x_j) \, ,$$
where $\pi : X \times X \to \mathbb{R}$ is a deterministic 'truth' function, and where the movies x_i are drawn randomly and independently according to a distribution $\mathcal{D}$ on X. See [3].

References

[1] S. Agarwal, T. Graepel, R. Herbrich, S. Har-Peled, and D. Roth. Generalization bounds for the area under the ROC curve. *Journal of Machine Learning Research*, 6:393–425, 2005.

[2] S. Agarwal and P. Niyogi. Generalization bounds for ranking algorithms via algorithmic stability. *Journal of Machine Learning Research*, 10:441–474, 2009.

[3] C. Rudin and R. E. Schapire. Margin-based ranking and an equivalence between AdaBoost and RankBoost. *Journal of Machine Learning Research*, 10:2193–2232, 2009.