Minimal complexity and asymptotic optimality of random projections

Mireille Boutin, Evzenie Coupkova (ecoupkov@purdue.edu)

Purdue University - Department of Mathematics

Abstract

The generalization error of a classifier is related to the complexity of the set of functions among which the classifier is chosen. We study a family of low-complexity classifiers consisting of thresholding a random one-dimensional feature. The feature is obtained by projecting the data on a random line after embedding it into a higher dimensional space parametrized by monomials of order up to k. More specifically, the extended data is projected n-times and the best classifier among those n, based on its performance on training data, is chosen. We obtain a bound on the generalization error of these low-complexity classifiers. The bound is less than that of any classifier with a non-trivial VC dimension, including linear classifiers. We also show that, given full knowledge of the class conditional densities, the error of the classifiers would converge to the optimal (Bayes) error as k and n go to infinity. On the other hand, if only a training dataset is given, we show that the classifiers will perfectly classify all the training points as k and n go to infinity.

Introduction of the method

Given a labeled training dataset $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$, where $\boldsymbol{x}_i \in \mathbb{R}^d$ and $y \in \{-1, 1\}$ we construct a classifier by

- Generate n vectors $\boldsymbol{r}_j \sim \text{Unif}(S^{d-1})$ independently.
- Project the data in the direction of each $\boldsymbol{r}_j$:

$$\widetilde{{}_j x_i} = \boldsymbol{x}_i \cdot \boldsymbol{r}_j$$

- Classify the projected data $\widetilde{{}_j x_i}$ in one dimension, compute the error of classification e_j.
- Choose the projection direction $\hat{\boldsymbol{r}}$ that results in the smallest classification error $\hat{e} = \min_j e_j$.
- For non-linear separation between classes first apply polynomial transformation of the original data of degree k.

Quantities of interest on a set of classifiers $\mathcal{F}$

- Hypothesis set

$$\mathcal{F} = \{f_{\boldsymbol{\theta}}(\boldsymbol{x}),\ \boldsymbol{\theta} \in \Theta\}$$

- Training error

$$R_{\text{train}}(f_\theta) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\{f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) \neq y_i\}$$

- Population error

$$R_{\text{popul}}(f_\theta) = \\ = \int_{\mathcal{E}\times\{-1,1\}} \rho_{\boldsymbol{X},Y}(\boldsymbol{x}, y)\mathbb{1}\{f_{\boldsymbol{\theta}}(\boldsymbol{x}) \neq y\} d(\boldsymbol{x}, y)$$

Bounds for the generalization gap

- With probability at least $1-\delta$:

$$\sup_{f\in\mathcal{F}} |R_{\text{popul}}(f) - R_{\text{train}}(f)| \leq R_{\mathcal{F}}(\delta, N)$$

- For an algorithm with a given VC dimension d_{VC}:

$$R_{\mathcal{F}} \leq \sqrt{\frac{8}{N}\ln\left(\frac{4(2N+1)^{d_{VC}}}{\delta}\right)}$$

- For thresholding after random projection (our result):

$$R_{\mathcal{F}} \leq \sqrt{\frac{8}{N}\ln\left(\frac{16nN}{\delta}\right)}$$

Comparison between the bounds for the generalization gap

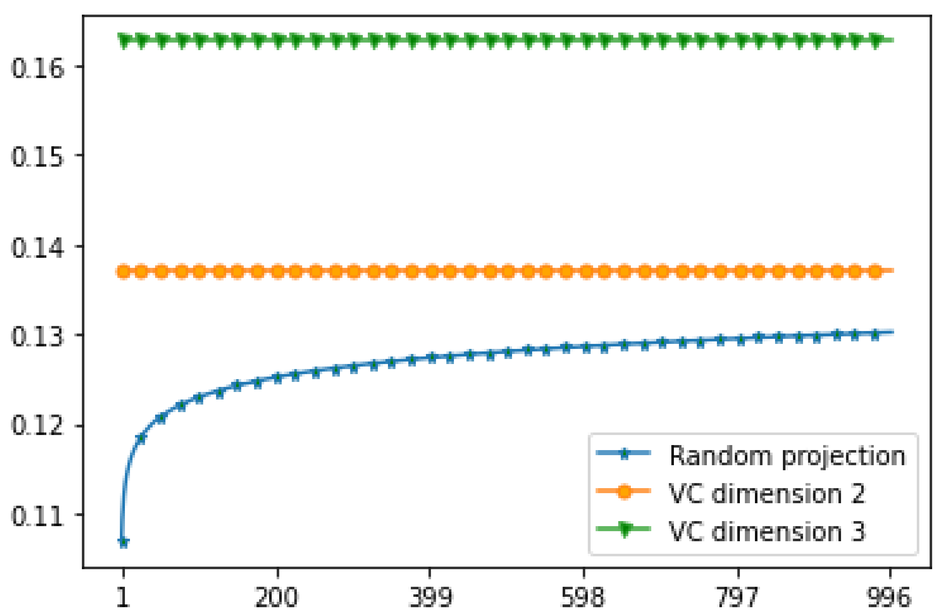

Figure 1:Upper bound for the generalization gap when $N = 10000$, $\delta = 0.1$ and n is between 1 and 1000 compared to the estimate given by Sauer-Shelah lemma [2] for VC dimensions 2 and 3.

Asymptotic optimality - consistence of the method

- Practical standpoint:
  - For k large enough, as n converges to infinity, the training error of the method converges to 0 in probability.
- Theoretical standpoint:
  - Given full knowledge of probability distributions, if Bayes decision rule results splits the support of the data distribution into two measurable sets, as k and n converge to infinity, the reducible error of the method converges to 0 in probability.

Future directions

This work might be expanded in several ways. First, it is possible to explore if applying a similar strategy would tighten even the better bound given by chaining technique [1]. Second, it might be interesting to figure out if the bound can be useful for more complex algorithms that use random projections as their building blocks.

References

[1] R. Dudley.
The Sizes of Compact Subsets of Hilbert Space and Continuity of Gaussian Processes, volume 1, pages 125–165.
07 2010.

[2] V. N. Vapnik and A. J. Červonenkis.
The uniform convergence of frequencies of the appearance of events to their probabilities.
Teor. Verojatnost. i Primenen., 16:264–279, 1971.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. 1826099. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation."