Lecture 7, December 13, 2016

Machine Learning Theory

Nearest Neighbors

Ruth Urner

1 Nearest Neighbors and alternative notions of learnability

In this lecture we will introduce and analyze another basic learning algorithm, namely prediction based on *Nearest Neighbors*. It is one of the earliest learning methods developed and studied [1]. A 1-Nearest Neighbor predictor returns, for a point x, the label of the sample point in S nearest to x. This simple paradigm is employable as soon as the domain space $\mathcal{X}$ is equipped with some metric, that is, as soon as we have a notion of distance in the domain space. In this lecture however, we will assume that the domain is Euclidean, and, in particular, we assume $\mathcal{X} = [0,1]^d$ is a d-dimension unit cube.

Then, formally, given a sample $S = ((\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_m, y_m))$, the output of 1-Nearest Neighbor learning is the predictor:

$$h_{\mathrm{NN}}(\boldsymbol{x}) = y_i,$$

where i is the index of the sample point closest to $\boldsymbol{x}$, that is

$$i = \mathrm{argmin}_{j=1\ldots,m} \|\boldsymbol{x} - \boldsymbol{x}_j\|.$$

It is not hard to see that the class of all functions, which can be the output of 1-Nearest Neighbor learning on some sample S has infinite VC-dimension. In particular, we have $h_{\mathrm{NN}}(\boldsymbol{x}_i) = y_i$ for all sample points $(\boldsymbol{x}_i, y_i)$, thus we can shatter arbitrarily large sets. This means, that Nearest Neighbor learning does not enjoy a strong learnability guarantee as in Definition 1, Lecture 2. In particular, we can not hope to prove convergence to the loss of the best nearest neighbor predictor (as in that Definition) with sample sizes that are independent of the distribution.

A weaker notion of learnability is *consistency*. A consistent learner is one, whose output converges to having smallest possible risk for any distribution (but does not necessarily converge at the same rate for all distributions). This best possible risk is also called the *Bayes risk* of the distribution. For a distribution P over $\mathcal{X} \times \{0,1\}$ the Bayes risk is defined as

$$L_P^* = \inf_{h \in \{0,1\}^{\mathcal{X}}} L_P(h).$$

Consistency can then be phrased as follows:

Definition 1. *A learner $\mathcal{A}$ is* consistent *if, for all $\epsilon > 0, \delta > 0$ and all distributions P, there exists a sample size $m(\epsilon, \delta, P)$ such that, for all $m \geq m(\epsilon, \delta, P)$ we have*

$$\mathbb{P}_{S \sim P}[L(\mathcal{A}(S)) \leq \mathcal{L}_P^* + \epsilon] \;\geq\; 1 - \delta$$

Note that, while the above notion of consistency is weaker than our original notion of learnability in the sense that it does not require uniform rates over all distributions, it is a stronger requirement in a different aspect. A consistent learner is required to converge to the best possible risk for any distribution, and not just to the approximation error of some fixed class $\mathcal{H}$ (which could be significantly larger).

1-Nearest Neighbor learning is not consistent. However, a variant, namely k-Nearest Neighbor prediction, has been shown to be consistent. A k-Nearest Neighbor predictor averages, for some point $\boldsymbol{x}$ the labels of the k nearest points in the sample rounds that average to 0 or 1. If k grows at a certain rate with the sample size n, then this simple learning algorithm is consistent [3].

We will focus on analysis of 1-Nearest Neighbor prediction and we will see that, at least for certain distribution, the risk of a 1-Nearest Neighbor classifier converges to twice the Bayes risk.

Note that the no-free-lunch results implies that, for classes of unbounded VC-dimension, we can also not get finite sample guarantees for convergence to twice the Bayes risk uniformly over all distributions. In order to still get some guarantees for finite sample sizes, we will therefore restrict the class of distributions. We will show that Nearest Neighbor prediction enjoys finite sample guarantees in terms of the Lipschitzness of the regression function of the distribution P. The regression function $\eta : \mathcal{X} \to [0, 1]$ is the function that, for a point x assigns it the probability of having label 1 under P. Formally

$$\eta(\boldsymbol{x}) = \mathbb{P}_{(x,y) \sim P}[y = 1 \mid x].$$

Recall that a function $f : \mathcal{X} \to \mathbb{R}$ is *L-Lipschitz* if, for all $x, x' \in \mathcal{X}$ we have

$$|f(x) - f(x')| \leq L|x - x'|.$$

That is the function can not "change value too fast". Intuitively, if a distribution P has an L-Lipschitz regression function, then we can expect that close by points are likely to have the same label (or label probability). Clearly this is also something that we would expect to be useful for Nearest Neighbor prediction to be successful. We will analyze 1-Nearest Neighbor prediction for classes of distributions whose regression functions are L-Lipschitz for some fixed value L.

2 Analysis of Nearest Neighbor prediction

We will now show how to get finite sample performance guarantees for Nearest Neighbor prediction. In this section, for some point $\boldsymbol{x}$, we will denote the sample point closest to $\boldsymbol{x}$ by $\mathrm{NN}(\boldsymbol{x})$. The proofs of the results in this section were shown in class. However, here, we will focus on the discussion of the results and refer to the corresponding proofs in the textbook, rather than reproducing the proofs.

We start by showing that the expected loss of 1-Nearest Neighbor prediction is bounded by twice the Bayes risk plus a term that depends on the Lipschitzness and the expected distance of points to their Nearest Neighbors in the sample.

Lemma 1. *Let $\mathcal{X} = [0,1]^d$ and let P be a distribution over $\mathcal{X} \times \{0,1\}$ with L-Lipschitz regression function η. Then, for all sample sizes m, we have*

$$\mathbb{E}_{S\sim P^m}[L_P(h_S)] \leq 2\mathcal{L}_P(h^*) + L \cdot \mathbb{E}_{S\sim P^m, \boldsymbol{x}\sim P}[\|\boldsymbol{x} - \mathrm{NN}(x)\|]$$

Proof. Proof of Lemma 19.1 in [2]. □

That is, we can control the loss of Nearest Neighbor prediction if we manage to control the distance of testpoints to their Nearest Neighbors. The following lemma will allow us to do that. For some collection of subsets over the domain space, the lemma bounds the joint mass (according to some distribution) of sets in the collection that are not hit by a sample form the distribution.

Lemma 2. *Let $C_1, \ldots, C_r$ be a collection of subsets of the domain $\mathcal{X}$ and let $P_\mathcal{X}$ be a distribution over $\mathcal{X}$. Then*

$$\mathbb{E}_{S\sim P_\mathcal{X}^m}[\sum_{i:C_i\cap S=\emptyset} P_\mathcal{X}(C_i)] \leq \frac{r}{m\mathrm{e}}$$

Proof. Proof of Lemma 19.2 in [2]. □

We will make use of this lemma in the following way. We will partition our domain space, the unit cube $\mathcal{X} = [0,1]^d$, into small sub-cubes of a fixed small side-length ϵ. There are $(1/\epsilon)^d$ many subcubes in such a partition. Then, we will choose m large enough so that the joint mass of cubes that are not hit by a sample from the distribution is small (using the bound in the above lemma). Now, for a testpoint drawn from the distribution, we can bound the probability of error by the probability that it falls into an empty cube (which is small by the above consideration) plus the probability that it falls into a cube which was hit and there is an error. For that case, the distance to its nearest neighbor is at most $\epsilon\sqrt{d}$, since there is a sample point in the same cube. Given that, we can invoke the bound of Lemma 1 to obtain the following overall bound on the error.

Theorem 1. *Let $\mathcal{X} = [0,1]^d$, and let the regression function η be L-Lipschitz. Then we have*

$$\mathbb{E}_{S\sim P^m}[L_P(h_{\mathrm{NN}})] \;\leq\; 2L_P(h^*) + 2L\sqrt{d}m^{\frac{1}{d+1}}$$

Proof. Proof of Theorem 19.3 in [2]. □

Note that, in order for the term $2L\sqrt{d}m^{\frac{1}{d+1}}$ to be smaller than some ϵ, we need $m \geq \left(\frac{2L\sqrt{d}}{\epsilon}\right)^{d+1}$. That is, the sample sizes needed to guarantee small error grow exponentially with the dimension of the space. In the next section, we will show that this exponential dependence is not avoidable.

3 Lower bounds

We now show that any algorithm, that has similar performance guarantees as the ones we saw for the 1-nearest neighbor prediction in the above section needs sample sizes that grow exponentially with the dimension of the space. Specifically, the next theorem says that, if the sample size is restricted to be smaller than $(L+1)^{d/2}$, then there is a distribution with

L-Lipschitz regression function on which the learner will have an expected error lower bounded by a constant $(1/4)$. Moreover, the Bayes risk of that distribution is 0. Thus, if the samples is smaller than $(L+1)^{d/2}$, then the learner's risk will not approach twice the Bayes risk.

Theorem 2. *For any learner $\mathcal{A}$ and any Lipschitz constant L, there is a distribution P over $[0,1]^d \times \{0,1\}$ with a L-Lipschitz regression function η and such that $L_P^* = 0$, but also such that for sample sizes $m \leq (L+1)^{d/2}$, we have*

$$\mathbb{E}_{S \sim P^m}[L_P(\mathcal{A}(S))] \geq 1/4$$

Proof. Proof of Theorem 19.4 in [2]. The idea is to construct distributions with support on points on a grid G of sidelength $1/L$. There are L^d points in such a grid. Note that all functions f from G to $\{0,1\}$ are L-Lipschitz since for any two different grid points x and x', we have

$$|f(x) - f(x')| \leq 1 = L\frac{1}{L} \leq L|x - x'|$$

We can now repeat the construction of the no-free-lunch theorem (Lecture 4, Theorem) using all functions on the grid G. □

References

[1] Thomas M. Cover and Peter E. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13:21–27, 1967.

[2] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning*. Cambridge University Press, 2014.

[3] Charles J. Stone. Consistent nonparametric regression. *The Annals of Statistics*, 5(4):595–620, 07 1977.