Consistency of Nearest Neighbor Classification under Selective Sampling

Oussama ENNAFII

January 4, 2015

Abstract

Consistency of the nearest neighbor, in the supervised learning framework, is well established with interesting convergence rates. Basically, It turns out that supervised learning needs, in average, for a precision rate of ϵ a sample of labeled points of the order of $\frac{1}{\epsilon}$. The problem arises when labeling points becomes very expensive. One should look for another method to yield the same order of precision with fewer labels. Active learning is a framework when this may be possible. Indeed, in this framework the learner chooses to label interesting points that yields the most information he can get. However, there is no guaranty that classifiers remain consistent in this kind of learning. This report is mainly influenced by the work of S.Dasgupta in: "Consistency of Nearest Neighbor Classification under Selective Sampling". In this article, we explain why some a priori sound strategies do not work. It also proposes a way to make nearest neighbor classifiers consistent with theoretical guaranties.

1 Introduction:

Active learning is a framework adapted to classification problems where labeling points turns out to be very expensive. In fact, we look to get a low error classifier with as few samples as possible. For instance, in the pharmaceutical industry, one wants to know whether a molecule binds with the target or not. Every labeling demand conducting an, expensive and time consuming, experiment on the molecule. Another example is classification

for landmines: every experiment is dangerous. The issue of interest here is that most heuristics yield unrepresentative points, and thus biased sampling.

To make our point, let us start by recalling some basics. In a binary classification problem, the instance space is noted as $\mathscr{X}$ and the label space as $\mathscr{Y}$. (X, Y) are generated from $\mathscr{X}$ x $\mathscr{Y}$ with distribution $\mathbf{P}$. Let $\eta(x) = \mathbb{E}(Y|X = x)$. A classifier $h : \mathscr{X} \leftarrow \mathscr{Y}$ has a risk:$R = \mathbb{P}(h(X) \neq Y)$. The minimizer of the the risk is the Bayes classifer $h^* = \mathbb{I}(\eta(x > 1/2)$. We note:$R^* = \mathbb{P}(h^*(X) \neq Y)$.

In supervised learning, the nearest neighbor is well studied. In [Fix and Hodges] and [Cover and Hart], One can find that the k_n -NN classifier's risk goes to at:

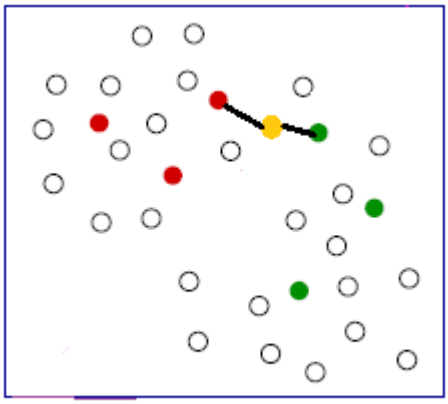
- $2R^*$, for $k_n = 1$.
- $R^* + O(\frac{1}{\sqrt{k}})$, for $k_n = k$ constant.
- R^*, for $k_n = o(n)$.

Our goal is to see how supervised learning affects nearest neighbor consistency. Let X_t be the observed sample at time t. The learner has to ask in the spot if he has to ask for its label or not. An aggressive strategy is to ask for labels only in the area, where the sample is, is unsure (see figure). A sounding heuristic is thus to ask for a label if its two nearest already quired neighbors do not agree.

The problem of such strategy is that it does not preserve consistency, as we sample only the subset of $\mathscr{X}$ where $\eta(x)$ is around $1/2$. An example where this strategy fails is when X is uniformly distributed in $[0, 1]$ and $Y = \mathbb{I}(X \in [1/2 - \alpha, 1/2 + \alpha]$, where $\alpha > 0$. We choose α so small that it is very likely to have the firt two samples labeled with O. It will never ask for other labels. The asymptotic risk is 2α whereas $R^* = 0$.

This is due to the fact that $\eta(x)$ may vary very much from a region to another unless there are some strong assymptions of regularity. It is then essential to probe every ball of the space $\mathscr{X}$ with some probability, say at least $1/n$.

Figure 1: Aggressive strategy: the yellow point is unsure.



We will develop, first, the theory for the separable case : where $R^* = 0$. This case requires fewer conditions since we already have valuable information: the space is separable. It means that we can separate the space into two subsets in which the label is constant. We just have to determine the boundaries of such subspaces.

We will see, via a counter-example, how the requirements for the separable case do not work for the non-separable one. Afterwards, we derive a general strategy that guaranties consistency.

2 Preliminaries:

Let $(\mathscr{X}, \rho)$ be a metric space. For $x \in \mathscr{X}$ and $r \geq 0$, $\mathscr{B}(x, r) := \{y \in \mathscr{X} : \rho(y, x) \leq r\}$ is the closed ball of radius r and x as center.

Let $Z_n := ((X_1, Y_1), (X_2, Y_2), ..., (X_t, Y_t), ..., (X_n, Y_n))$ be independent realizations of $\mathbf{P}$. We denote by Q_n the set of intances whose labels have been queried, and by $\tilde{Y}_t(x)$ the label of $x \in Q_n$. We denote also, by $\Gamma(x, S)$ the nearest neighbor of x in the set S and by $\Gamma_k(x, S)$ the set of k nearest neighbors of x in S.

We can write formaly:

$$T_n^1(x) = \tilde{Y}_n(\Gamma(x, Q_n)), \forall x$$

and

$$T_n^k(x) = \mathbb{I}(\sum_{z \in \Gamma_k(x, Q_n)} \tilde{Y}_n \geq \lfloor k/2 \rfloor), \forall x$$

$$R_n = \mathbb{P}(T_n(X) \neq Y)$$

where T_n^k is the k-NN classifier. Consistency is when $\mathbb{E}(R_n) \to R^*$. The expectations is over all possible realizations Z_n. Strong consistency, on the other hand, means that with probability 1, $R_n \to R^*$.

We suppose also that we have a sequence $(u_t)_{t \in \mathbb{N}^*}$. We define $s_n = \sum_{t \leq n} u_t$.

(R_1) Every Y_t is queried with probability at least u_t, such that: $\frac{u_n}{k_n} \to \infty$.

3 The separable case:

We start off with the separable case. In this setting, The space is divided up to two regions:

$$\mathscr{X}_l := \{x \in \mathscr{X} : \exists r > 0 \text{ s.t. } \forall z \in \mathscr{B}(x, r), \eta(z) = l \text{ and } \mu(\mathscr{B}(x, r)) > 0\}, \forall l = 0, 1$$

We assume herein that:

(A_1) $\mu(\mathscr{X}_0 \cup \mathscr{X}_1) = 1$.

μ is the marginal probability measure of X.

We want to see if (R_1) is sufficient to get the desired consistency.
Let:

$$Y_t^{(q)} = \begin{cases} Y_t, & \text{if } \textit{queried} \\ ?, & \text{if } \textit{not} \end{cases}$$

and

$$\mathscr{F}_n = \sigma((X_t, Y_t^{(q)})_{t \leq n})$$

(R_1) means that:

$$\mathbb{P}(Y_n \text{ is queried } | \mathscr{F}_{n-1}, X_n) \geq u_n \tag{1}$$

The first thing to verify is wheither this condition lets us span all the space.

Lemma 3.1 *Pick a ball B with $\mu(B) > 0$, if k_n is a non-decreasing sequence of positive integers and $s_n/k_n \to \infty$, then:*

$$\exists n_0 > 0 \ s.t. \ \forall n \geq n_0, |Q_n \cap B| \geq k_n$$

Proof We define the event:

$$E_t = \{X_t \in Q_n \cap B\}$$

We know that:

$$\mathbb{P}(E_t|\mathscr{F}_{t-1}) \geq \mu(B)u_n$$

It means that:

$$\sum_{t \leq n} \mathbb{P}(E_t|\mathscr{F}_{t-1}) \geq \mu(B)s_n \to \infty$$

We conclude with Levy's extension of Borel-Cantelli's lemma [Williams]:

$$\frac{\sum_{t \leq n} \mathbb{I}(E_t)}{\sum_{t \leq n} \mathbb{P}(E_t|\mathscr{F}_{t-1})} \to 1, \text{ a.s.}$$

It means that:

$$\exists n_1 > 0 \text{ s.t. } \forall n \geq n_1, |Q_n \cap B| \geq 1/2 \sum_{t \leq n} \mathbb{P}(E_t|\mathscr{F}_{t-1}) \geq \mu(B)s_n/2$$

$s_n/k_n \to \infty$ means that:

$$\exists n_2 > 0 \text{ s.t. } \forall n \geq n_2, s_n/k_n \geq 2\mu(B)$$

It follows:

$$\forall n \geq n_0 := max(n_1, n_2) \ , |Q_n \cap B| \geq k_n$$

It follows from this lemma that:

Theorem 3.2 *Let k_n a non-decreasing sequence of positive integers. For a selective sampling scheme that meets the requirement (R_1). If (A_1) holds then the resulting k_n-NN classifier is strongly consistent.*

The proof of this theorem is in[dasgupta].

This means that, in the separable case, nearest neighbor classifiers are almost surely consistent as soon as we can span the whole space. Requirement (R_1) gives a way to span this space, while the additionnal assymption means that we can always label instances. In this case, with for instance, $u_n = 1/n, \forall n$ we can ensure consitency with paying $O(log(n))$ more labels.

4 Counter-example: biased sampling:

In the general case, we cannot always separate the space as before. The requirement (R_1) is no longer sufficient as we may sub-query a kind of instances in an region of the space. It results that the classifier is erroneous over that region and thus biased over all.

We consider in this example that:

$$\mu \sim U([0,1])$$

and Y is a - π - Bernoulli variable independent from X.

For $\pi < 1/2$:

$$h^* \equiv 0$$

and

$$R^* = \pi$$

We devise the strategy:

(S_1) Given X_t:

(i). $\tilde{Y}_{t-1}(\Gamma(X_t, Q_{t-1})) = 0$, querry Y_t. (Type-0 querry)

(ii). $\tilde{Y}_{t-1}(\Gamma(X_t, Q_{t-1})) = 1$, querry Y_t with probability $1/n$. (Type-1 querry)

Set $x \in [0,1]$. We define:

$$H(n,m) = \sum_{n \le t \le m} 1/t$$

The main idea is that after time D_n, there are two neighbors of x in the interval $I_n := [x - r_n, x + r_n]$ that have been queried: we denote their labels (with $t \geq D_n$) as $\tilde{Y}_t^L$ and $\tilde{Y}_t^R$ for the left and right neighbor respectively.

The probability that some X_t is in the left side of I_n and to be quieried is at least r_n/t. It follows that the probability so as not to get before D_n a quiered neighbor at the left is:

$$\prod_{t \le D_n} (1 - r_n/t) \le exp(-r_n H(1, D_n))$$

Lemma 4.1 *The probabilty that before D_n, we did not have a quiered neighbor of x at each side in I_n is below $2exp(-r_nH(1,D_n))$*

We call,a_n, the time at which the nearest neighbor changes arrival time (the c_n^{th} arrival after the first phase). The fact is between D_n and a_n, $\tilde{Y}_t^L = \tilde{Y}_t^R = 1$ with big probability. That is because the probability to get, between two arrivals, $\tilde{Y}_t^L = \tilde{Y}_t^R = 1$ is constant Cst[dasgupta]. It follows that:

Lemma 4.2 *The chance to not get $\tilde{Y}_t^L = \tilde{Y}_t^R = 1$ between D_n and a_n is below: $(1 - Cst)^{c_n - 1}$.*

Moreover, we can get $a_n \leq n$ also with big probability.

Lemma 4.3 *The probability that $a_n > n$ is below: $\frac{4}{H(D_n+1,n)}$ if : $c_n \leq H(D_n + 1, n)/2$.*

To finish, the probability of a Type 1 query in I_n at time t is at most $2r_n/t$. Thus:

Lemma 4.4 *The probability to have queries of Type-1 in I_n after the initial phase is below: $2r_nH(D_n + 1, n)$.*

If we choose then, for instance:

$$r_n = \frac{1}{\sqrt{log\ n}}\ ,\ D_n = \frac{n}{log\ n}\ ,\ c_n = \sqrt{log\ log\ n}$$

We conclude:

Theorem 4.5 *Under (S_1):*

$$\forall x \in [0,1]\ ,\ \mathbb{P}(T_n^1(x) = 1) \to 1$$

It means that:

$$R_n \to 1 - \pi$$

while $R^* = \pi$.

5 The general case:

In the case when η is continuous over $[0,1]$, the requirement (R_1) is not sufficient to ensure consistency. Indeed, in the previous example, we favored the label '1'. The resulting 1-NN classifier predicted 1 everywhere. A way to solve this issue is by querying labels dependeing on the homogeneity of the neighborhood of x. We pay close attention then to the granularity effect, as we can have for example pockets of '0' labels in a wider region of '1' region.

The probleme with such a strategy is that is more complex. A simpler way to remedy to the issue is by separating queried labels to two equal size sets. The first one, Q_n is for querying labels. The second, F_n is to be used afterwards to classify. It looks like we use active learning to choose the interesting points before learning over these points as in the supervised learning framework.

We summerize this in the following: We define:

$$Y_t^{(q,1)} = \begin{cases} Y_t, & \text{if queried and not in } F_n \\ !, & \text{if queried and in } F_n \\ ?, & \text{if not queried} \end{cases}$$

$$Y_t^{(q,2)} = \begin{cases} Y_t, & \text{if queried and in } F_n \\ !, & \text{otherwise} \end{cases}$$

and

$$\mathscr{G}_t = \sigma(X_1, Y_1^{(q,1)}, ..., X_t, Y_t^{(q,1)})$$

(R_2) The strategy satisfies:

- Decision about querying Y_t and placing X_n in F_t are based only on $\mathscr{G}_{t-1}$ and X_t.
- $\mathbb{P}(X_t \in F_n | \mathscr{G}_{t-1}, X_t) \geq u_n$

For example, we can use any strategy by making sure that we query X_t with probability $2u_t$ and then with probability $1/2$ put it in F_n.

We assume that:

(A_2) The metric space $(\mathscr{X}, \rho)$ is seperable. A consequence of this property is that the support of μ is of mass 1.

(A_3) For all x, almost surely, either $\mu(x) > 0$ or η is a continuous at x.

With the same argument as for the separable case, we can deduce:

Lemma 5.1 *Pick a ball B with $\mu(B) > 0$, if k_n is a non-decreasing sequence of positive integers and $s_n/k_n \to \infty$, then:*

$$\exists n_0 > 0 \ s.t. \ \forall n \geq n_0, |Q_n \cap B| \geq k_n$$

In [dasgupta] we can prove an important result:

Lemma 5.2 *Set: $\epsilon > 0$, $\eta < 1$. Let $B, B_1, ..., B_k$ Bernoulli variables with parameters $\eta, \eta_1, ...\eta_k$ respectively, such that:*

$$|\eta_l - \eta| \leq \epsilon \ , \ \forall l = 1, ..., k$$

Let V_k the majority vote over $B_1, ..., B_k$ breaking ties with a fair coin flip. Define $C(k, \epsilon, \eta)$ the supremum of $\mathbb{P}(V_k \neq B)$.

$$C(k, \epsilon, \eta) \leq \begin{cases} 2min(\eta, 1-\eta) + \epsilon, \text{if } k = 1 \\ min(\eta, 1-\eta) + 2/\sqrt{k}, \text{if } k > 1 \text{ and either } \eta = 1/2 \text{ or } \epsilon \leq |1 - 2\eta|/4 \end{cases}$$

Now we set:
$\epsilon_0 > 0$ and

$$\epsilon(x) := \begin{cases} \epsilon_0, \text{if } \eta(x) = 1/2 \\ min(\epsilon_0, |1 - 2\eta(x)|/4), \text{otherwise} \end{cases}$$

We use then the previous lemmas as in[dasgupta] to deduce:

Theorem 5.3 *Under the assymptions (A_2, A_3) and the requirement (R_2).*