# Analyzing Dataset Consistency

Charles Davi

March 8, 2021

**Abstract**

Many consumer devices can be used to perform parallel computations, and in a series of approximately five-hundred research notes,[1] I introduced a new and comprehensive model of artificial intelligence rooted in information theory and parallel computing that allows for classification and prediction in worst-case polynomial time, using what is effectively AutoML, since the user only needs to provide the datasets. This results in run-times that are simply incomparable to any other approach to A.I. of which I'm aware, with classifications at times taking seconds over datasets comprised of tens of millions of vectors, even when run on consumer devices. Below is a series of formal lemmas, corollaries, and proofs, that form the theoretical basis of my model of A.I.

## 1 Local Consistency

A dataset $A$ is ***locally consistent*** if there is a value $\delta_x > 0$, for each $x \in A$, such that the set of all points for which $||x - y_i|| \leq \delta_x$, $B_x = \{y_1, \ldots, y_k\}$, has a single classifier, equal to the classifier of $x$, and $B_x$ is non-empty, for all $x$. That is, all of the points within $B_x$ are within $\delta_x$ of $x$, and are of a single class. We say that a dataset is locally consistent around a single point $x \in A$, if $B_x$ is non-empty.

For reference, the nearest neighbor algorithm simply returns the vector within a dataset that has the smallest Euclidean distance from the input vector. Expressed symbolically, if $f(x, A)$ implements the nearest neighbor algorithm, as applied to $x$, over $A$, and $y = f(x, A)$, for some $y \neq x$, then $||x-y|| \leq ||x-z||$, for all $z \in A$. Further, assume that if there is at least one $y_i \in A$, such that $||x - y|| = ||x - y_i||$, then the nearest neighbor algorithm will return a set of outputs, $\{y, y_1, \ldots, y_k\}$.

[1] All of the research notes and applicable code are publicly available on my ResearchGate homepage.

**Lemma 1.1.** *If a dataset $A$ is locally consistent, then the nearest neighbor algorithm will never generate an error.*

*Proof.* Assume that the nearest neighbor algorithm is applied to some dataset $A$, and that $A$ is locally consistent. Further, assume that for some input vector $x$, the output of the nearest neighbor algorithm, over $A$, contains some vector $y$, and that $y$ has a classifier that is not equal to the classifier of $x$, thereby generating an error. Since $y$ is contained in the output of the nearest neighbor algorithm, as applied to $x$, over $A$, it follows that $||x - y||$ is the minimum over the entire dataset. Because $A$ is locally consistent, there is some nonempty set $B_x$, that contains every vector $y_i$, for which, $||x - y_i|| \leq \delta_x$. Since $||x - y||$ is minimum, it must be the case that $||x - y|| \leq ||x - y_i|| \leq \delta_x$, for all $y_i$. But this implies that $y \in B_x$, which contradicts our assumption that the classifiers of $x$ and $y$ are not equal. Therefore, if we nonetheless assume their classifiers are not equal, then $A$ is not locally consistent, which leads to a contradiction, thereby completing the proof. □

What Lemma 1.1 says is that the nearest neighbor algorithm will never generate an error when applied to a locally consistent dataset, since the nearest neighbor of an input vector $x$ is always an element of $B_x$, which guarantees accuracy. As a practical matter, if the nearest neighbor isn't within $\delta_x$ of an input vector $x$, then we can flag the prediction as possibly erroneous. This could of course happen when your training dataset is locally consistent, and your testing dataset adds new data that does not fit into any $B_x$.

**Corollary 1.2.** *A dataset is locally consistent if and only if the nearest neighbor algorithm does not generate any errors.*

*Proof.* Lemma 1.1 shows that if a dataset $A$ is locally consistent, then the nearest neighbor algorithm will not generate any errors. So now assume that the nearest neighbor algorithm generates no errors when applied to $A$, and for each $x \in A$, let $y_x$ denote a nearest neighbor of $x$, and let $\delta_x = ||x - y_x||$. Let $B_x = y_x$, and so each $B_x$ is non-empty. If more than one vector is returned by the nearest neighbor algorithm for a given $x$, then include all such vectors in $B_x$. By definition, $||x - y_x|| \leq \delta_x$. Therefore, there is a value $\delta_x > 0$, for each $x \in A$, such that the set of all points for which $||x - y_i|| \leq \delta_x$, $B_x$, has a single classifier, and $B_x$ is non-empty for all such $x$, which completes the proof. □

**Corollary 1.3.** *A dataset is locally consistent around $x \in A$, if and only if the nearest neighbor algorithm does not generate an error, when applied to $x$, over $A$.*

*Proof.* Assume that the nearest neighbor algorithm does not generate an error when applied to $x \in A$. It follows that there is some $B = \{y_1, \ldots, y_k\} \subseteq A$,

generated by the nearest neighbor algorithm, as applied to $x$, over $A$, for which $||x - y_i||$ is minimum, and the classifiers of $x$ and $y_i$ are all equal. Let $B_x = B$, and let $\delta_x = ||x - y_i||$. Since, the classifiers of $x$ and $y_i$ are all equal, this implies that $A$ is locally consistent around $x$.

Now assume that $A$ is locally consistent around $x$, and let $B = \{y_1, \ldots, y_k\} \subseteq A$ be the output of the nearest neighbor algorithm, as applied to $x$, over $A$. It follows that $||x - y_i||$ is minimum over $A$, and therefore, $B \subseteq B_x$. This implies that the classifier of $x$ equals the classifier of each $y_i$, which completes the proof. □

# 2 Clustering

We can analogize the idea of local consistency to geometry, which will create intuitive clusters that occupy some portion of Euclidean space. We say that a dataset $A$ is ***clustered consistently***, if for every $x \in A$, there exists some $\delta_x > 0$, such that all points contained within a sphere of radius $\delta_x$ (including its boundary), with an origin of $x$, denoted $S_x$, have the same classifier as $x$.

**Lemma 2.1.** *A dataset is locally consistent if and only if it is clustered consistently.*

*Proof.* Assume $A$ is locally consistent. It follows that for each $x \in A$, there exists a non-empty set of vectors $B_x$, each of which have the same classifier as $x$. Moreover, for each $y_i \in B_x$, it is the case that $||x - y_i|| \leq \delta_x$. This implies that all vectors within $B_x$ are contained within a sphere of radius $\delta_x$, with an origin of $x$. Therefore, $A$ is clustered consistently.

Now assume that $A$ is clustered consistently. It follows that all of the vectors $y_i$ within $S_x$ have a distance of at most $\delta_x$, from $x$. Let $B_x$ be the set of all vectors within $S_x$. This implies that for every $x \in A$, there exists a non-empty set of vectors $B_x$, all with the same classifier, such that for all $y_i \in B_x$, $||x - y_i|| \leq \delta_x > 0$. Therefore, $A$ is locally consistent, which completes the proof. □

What Lemma 2.1 says, is that the notion of local consistency is equivalent to an intuitive geometric definition of a cluster that has a single classifier.

**Lemma 2.2.** *If a dataset $A$ is locally consistent, and two vectors $x_1$ and $x_2$ have unequal classifiers, then there is no vector $y \in A$, such that $y \in C = (S_{x_1} \cap S_{x_2})$.*

*Proof.* If the regions bounded by the spheres $S_{x_1}$ and $S_{x_2}$ do not overlap, then $C$ is empty, and therefore cannot contain any vectors at all. So assume instead

that the regions bounded by the spheres $S_{x_1}$ and $S_{x_2}$ do overlap, and that therefore, $C$ is a non-empty set of points, and further, that there is some $y \in C$, such that $y \in A$.

There are three possibilities regarding the classifier of $y$:

(1) It is equal to the classifier of $x_1$;

(2) It is equal to the classifier of $x_2$;

(3) It is not equal to the classifiers of $x_1$ or $x_2$.

In all three cases, it follows that $y \in B_{x_1}$, and $y \in B_{x_2}$, which leads to a contradiction, since the Lemma assumes that the classifiers of $x_1$ and $x_2$ are not equal. Since these three cases cover all possible values for the classifier of $y$, it must be the case that $y$ does not exist, which completes the proof. □

What Lemma 2.2 says, is that if two clusters overlap as spheres, then their intersecting region contains no vectors from the dataset.

**Lemma 2.3.** *If $A$ is locally consistent, then any clustering algorithm that finds the correct value of $\delta_x$, for all $x \in A$, can generate clusters over $A$ with no errors.*

*Proof.* This follows immediately from Lemma 2.2, since any such algorithm could simply generate the sphere $S_x$, for all $x \in A$. □