Properties of a Binary Similarity Measure

Ben Veal
Department of Mathematics
London School of Economics
London WC2A 2AE, U.K.
b.w.veal@lse.ac.uk

CDAM Research Report LSE-CDAM-2005-06

March 2005

Abstract

Say we have a set of data which can be represented by a set of distinct binary vectors $A \subseteq \{0,1\}^n$ (e.g. medical data: each vector could correspond to a patient and each entry to the presence or absence of a particular symptom), and each vector has a corresponding label of either 0 or 1 (for example this could represent whether a patient has a particular disease or not). We form some classification rule $h : \{0,1\}^n \to \{0,1\}$ that correctly labels all of the vectors in A. Now we are presented with a new vector $\mathbf{x} \in \{0,1\}^n \backslash A$ (with no assumptions of what distribution it has come from), and we would ideally want it to be correctly classified by h. We might expect that if $\mathbf{x}$ is 'similar' to the vectors in A then it is more likely to be correctly classified by h (since h is correct or 'consistent' on A), but how do we measure the similarity of $\mathbf{x}$ to A? Here we explore one particular similarity measure, investigating its usefulness in classification and its combinatorial and extremal properties.

1 Introduction

In this report we investigate a 'measure of similarity' for binary vectors first introduced in [AH04].

The original idea was motivated by problems in classification of medical data. A patient may be represented by a binary vector $\mathbf{x} \in \{0,1\}^n$ where each entry indicates the presence of absence of a particular symptom, the symptoms corresponding to the indices of the vector. A collection of entries of the vector corresponds to a particular 'syndrome' displayed by the patient.

We may want to diagnose patients based on the symptoms (or lack of symptoms) they display, i.e. based on these patient vectors. If we know the true condition of all the patients corresponding to vectors in some set A, we can use this information to hypothesize some rule for diagnosing other patients based on their vectors. But given a new patient, how sure can we be that our hypothesized rule will correctly diagnose that patient?

Many theories and results exist in the machine learning literature to help answer this question, but in those cases assumptions are made on the patients that are presented to us, usually that they are independent, and identically distributed. In real world applications we can rarely make such assumptions. In section 2 we show that the similarity measure investigated here may help to answer this question when no such probabilistic assumptions are made. In section 3 we present some combinatorial results linking the size of a vector set A with extremal values of our similarity measure. As sugested in [AH04] these extremal values give us a measure of the 'representativeness' of the vector set A. Finally in section 4 we give some asymptotic results on the values of previously mentioned quantities.

First let's explain the similarity measure we have talked about in terms of the patient data example given above. Say we have a set of patient vectors $A \subseteq \{0,1\}^n$ (where n is the total number of symptoms under consideration), and consider two new patients represented by the vectors $\mathbf{x}, \mathbf{y} \notin A$. If for every collection of k symptoms (k some number between 1 and n) there is a patient in A which is the same as $\mathbf{y}$ on those symptoms, (i.e. there is a patient in A displaying the same syndromes of length k as $\mathbf{y}$), but the

same cannot be said for $\mathbf{x}$, then $\mathbf{y}$ is considered more similar to A than $\mathbf{x}$ is. Equivalently if the shortest length syndrome of $\mathbf{y}$ that doesn't appear amongst any of the patients in A is longer than the corresponding syndrome for $\mathbf{x}$, then we consider $\mathbf{y}$ as being more similar to A than $\mathbf{x}$ is. Thus $\mathbf{x} \notin A$ is regarded as dissimilar to A if there is some 'short' syndrome observed in $\mathbf{x}$ that is absent from any vector in A, the extreme case being when there is a single symptom on which $\mathbf{x}$ differs from all vector of A.

Note that this definition of similarity measure is not a metric, since we are comparing a vector with a *set* of vectors. There are many other similarity measures we could use. For example the minimum Hamming distance of a vector $\mathbf{x}$ to a vector set A is defined as $\min\{d(\mathbf{x}, \mathbf{y}) : \mathbf{y} \in A\}$ where $d(\mathbf{x}, \mathbf{y})$ is the number of entries where $\mathbf{x}$ and $\mathbf{y}$ differ. However we believe the similarity measure investigated here is better suited to the medical data example explained above. The relationship between minimum Hamming distance and the similarity measure investigated here is explored in [AH04].

From now on I shall use mathematical terminology and talk of coordinates or indices of the vectors rather than symptoms. Also for an index set $I \subseteq [n]$ and $\mathbf{z} \in \{0,1\}^n$ we define $\mathbf{z}|_I \in \{0,1\}^{|I|}$ to be the vector formed by restricting $\mathbf{z}$ to the indices in I. (We may sometimes refer to $\mathbf{z}|_I$ as a *subvector* of $\mathbf{z}$.) Similarly for vector set $V \subseteq \{0,1\}^n$ we define $V|_I = \{\mathbf{z}|_I : \mathbf{z} \in V\}$.

We now give the definition of similarity studied in this paper.

Definition 1.1 *For $A \subseteq \{0,1\}^n$ and $\mathbf{x} \in \{0,1\}^n$, the* similarity *of $\mathbf{x}$ to A, $s(\mathbf{x}, A)$, is defined to be the largest k such that every subvector $\mathbf{x}|_I$ of dimension k appears also as a subvector $\mathbf{y}|_I$ of some observation $\mathbf{y} \in A$. That is,*

$$s(\mathbf{x}, A) = \max\{s : \forall I \subseteq [n], |I| \leq s, \exists \mathbf{y} \in A, \mathbf{y}|_I = \mathbf{x}|_I\}$$

where n is the dimension of the vectors (the number of symptoms considered). Also define $s(\mathbf{x}, \emptyset) = 0$ for all vectors $\mathbf{x}$.

Note that $0 \leq s(\mathbf{x}, A) \leq n - 1 \ \forall \mathbf{x} \in A^c$ (where $A^c = \{0,1\}^n \backslash A$), and $s(\mathbf{x}, A) = n \ \forall \mathbf{x} \in A$.

Note also that $s(\mathbf{x}, A)$ is one less than the size of the smallest syndrome of $\mathbf{x}$ that does *not* appear in any of the vectors in A. This is similar to the definition of 'specification number' defined in [ABST95], and 'witness set' explored in [KLRS96]. Similar ideas have appeared in the context of machine learning in [GK95, Heg94, MS91]. However in those cases we are considering vectors *within* our vector set, and the measure is used differently to the way in which it is used here. Specifically the specification number of $\mathbf{x} \in A$ is equivalent to $s(\mathbf{x}, A \backslash \{\mathbf{x}\}) + 1$. This tells us how many entries of vector $\mathbf{x}$ need to be revealed to a learner in order for them to be able to say with certainty which of the vectors of A it is (the set A being known to them).

2 Classification accuracy and similarity

In this section we present a result linking similarity measure with the accuracy of classification.

Before going any further, let's introduce some more terminology and assumptions that shall be used later. We assume that every vector in $\{0,1\}^n$ has a corresponding label of either 0 or 1, and call this the *classification* of the vector. This could represent for example whether a patient corresponding to that vector has a particular disease or not. (We assume that all patients with the same vector will have the same diagnosis - theoretically it's always possible to include enough symptoms so that this is true.) In other words we are assuming some function or *concept*, $c : \{0,1\}^n \to \{0,1\}$ that classifies the vectors in $\{0,1\}^n$ (the exact specification of c is unknown to us). We have some set of vectors A, called our *training set* for which we know the classifications, and we use this information to try and approximate c with another concept $h : \{0,1\}^n \to \{0,1\}$ (the hypothesized diagnosis rule mentioned in the example in the introduction). We call c our *underlying* or *true* concept and h our *hypothesis*. Most of the classifications given by h are called predictions since we don't know if they are correct or not. We shall assume that h is *consistent* with c on our training set, in other words $h(\mathbf{x}) = c(\mathbf{x})$ for all vectors $\mathbf{x}$ in A, then $\mathbf{x}$ is said to be 'misclassified' by h if $h(\mathbf{x}) \neq c(\mathbf{x})$.

There are many elegant and useful results in the machine learning literature on the classification accuracy of consistent hypotheses (see for example [BEHW89, Val84, Ant01, AB99]). A typical result from 'PAC' learning, for example, will tell us for a given confidence level and size of training set, the probability of misclassification by our hypothesis when the underlying concept is restricted to be in some known class of functions. However, as mentioned in the introduction, all these approaches depend on probabilistic assumptions about the sample data that is presented to us.

Recent work by Vovk ([Vov02]) outlines an algorithm that not only produces predictions but also an indication of their credibility. The credibility of the prediction depends on the example which we are trying to classify, and is ascertained by looking at how 'strange' it would be to give this example a given classification when compared to previously classified data. This is similar to the approach suggested in [AH04] for use with similarity measure. There it is suggested that perhaps we can be more confident of our predictions on data that are 'similar' to our training set A, where 'similarity' is measured according to definition 1.1. We could then form a hierarchy of classification confidence based on similarity measure, and maybe decide not to classify data that is low in this hierarchy. Empirical work carried out at Rutgers University [HSS04] seems to confirm this approach. Experimental results on real life data sets show a higher error rate for vectors with a low similarity measure.

Unlike current PAC learning theory, and the work of Vovk, it is suggested in [AH04] that the similarity measure approach might be useful even when no probabilistic assumptions about the sample data can be made. Here we present some theoretical results to give some weight to this idea.

First we need to introduce some background material. There is a standard way of representing a Boolean function $f : \{0,1\}^n \to \{0,1\}$ by a formula which shall now be explained.

For each index i of the input vectors we associate a *literal*, u_i, and its

negation, $\overline{u}_i$, which represent the following functions:

$$\begin{aligned} u_i(\mathbf{x}) &= \begin{cases} 1 & \text{if } \mathbf{x}_i = 1, \\ 0 & \text{if } \mathbf{x}_i = 0, \end{cases} \\ \overline{u}_i(\mathbf{x}) &= \begin{cases} 1 & \text{if } \mathbf{x}_i = 0, \\ 0 & \text{if } \mathbf{x}_i = 1, \end{cases} \end{aligned} \tag{1}$$

We can use parenthesis "(", ")" and the logical connectives $\wedge$ (AND), $\vee$ (OR) with the literals and their negations to form other Boolean formulae.
E.g. $\psi = (u_1 \wedge \overline{u}_2) \vee u_3$ represents the following function:

$$\psi(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbf{x}_1 = 1 \textit{ and } \mathbf{x}_2 = 0, \text{ or if } \mathbf{x}_3 = 1 \\ 0 & \text{otherwise} \end{cases}$$

We can then form more complex formulae by recursive use of parenthesis and logical connectives e.g. $(\psi \vee \phi) \wedge u_3$. A *disjunction* is that formed by use of the $\vee$ connective, e.g. $\psi \vee \phi$, and a *conjuction* by use of the $\wedge$ connective, e.g. $\psi \wedge \phi$. A *monomial* is a conjunction of literals and/or negations of literals. Often the $\wedge$ symbol is dropped from the notation in this case, e.g. $u_1\overline{u}_2u_3$ instead of $u_1 \wedge \overline{u}_2 \wedge u_3$. A *disjunctive normal form* (written DNF) is a disjunction of monomials, e.g. $(u_1\overline{u}_2u_3) \vee (\overline{u}_3u_4u_5) \vee (u_3\overline{u}_4u_5)$. Each monomial in the DNF is called a *term* of the DNF. It turns out that any Boolean function can be represented by a DNF formula (see for example [Ant01]).

DNF formulae are the natural formulae to think about when considering binary medical data as in the example in the introduction. Doctors usually determine whether a patient has a particular condition by looking for the presence of a collection of symptoms, e.g. a stuffy nose, sore throat and a cough may indicate a cold. Some symptoms may indicate a different condition than that which is being checked for, and so absence of these symptoms will be required for a positive classification, e.g. if the patient also has a fever, then they are unlikely to have a cold and more likely to have the flu. If we associate each literal with a symptom as we did for the corresponding vector indices in the introduction, then a monomial corresponds to a collection of symptoms and/or lack of symptoms, i.e. a syndrome. This monomial takes the value 1 if and only if the symptoms corresponding to the literals in the monomial are present, and those of the negated literals are absent,

i.e. if and only if the patient displays the syndrome. A monomial thus gives us a simple binary classification rule, based on one particular syndrome. It may be however that there are several syndromes that lead to a positive classification. For example sneezing and a sore throat with no fever may also indicate a cold. In this case a disjunction of monomials, i.e. a DNF, would be appropriate.

In the case where the initial vector set to be considered is not binary, for example one of the indices may correspond to a temperature reading, we can transform the vectors into binary vectors by a process called *binarization*. This consists of setting a number of cut-points for each non-binary index in the original vector, and then associating a binary entry in the transformed vector for each cut-point. The binary entry then takes the value 1 if and only if the correponding entry in the original vector is greater than the cut-point corresponding to this binary entry. See [BHIK97] for more details.

For many classification problems, especially the medical example given, we can expect that the number of terms (number of syndromes), and the maximum length (number of literals/symptoms) of each term (syndrome) are bounded. This leads to the following definition.

Definition 2.1 *A* k-term-l-DNF *is a DNF of at most* k *terms, and in which each term contains at most* l *literals.*

Sometimes we may be interested in negating the ouput of a Boolean function,

Definition 2.2 *The* negation *of a Boolean function* $f : \{0,1\}^n \to \{0,1\}$, *denoted* $\bar{f}$ *is obtained by negating the output of* f*:*

$$\bar{f}(\mathbf{x}) = \begin{cases} 1 & if\ f(\mathbf{x}) = 0 \\ 0 & if\ f(\mathbf{x}) = 1 \end{cases}$$

If we know a DNF for a Boolean function f we can find a formula for $\bar{f}$ by the process of *negation*. This consists of negating every literal that appears

in f and then replacing every $\wedge$ with a $\vee$, and every $\vee$ with a $\wedge$ (see [Ant01]). The resulting formula is in *conjunctive normal form*, or CNF. It consists of a conjunction of disjunctions of monomials. We can transform this into a DNF by using the distributive law for Boolean formulae:

$$u_3 \wedge (u_1 \vee u_2) = (u_1 \wedge u_3) \vee (u_2 \vee u_3)$$

The following simple example makes this process clearer.
Let $\phi = u_1u_2 \vee u_3u_4$, then negating gives: $\overline{\phi} = (\overline{u}_1 \vee \overline{u}_2) \wedge (\overline{u}_3 \vee \overline{u}_4)$, and then using the distributive law we get: $\overline{\phi} = \overline{u}_1\overline{u}_3 \vee \overline{u}_1\overline{u}_4 \vee \overline{u}_2\overline{u}_3 \vee \overline{u}_2\overline{u}_4$. Consequently we have the following lemma:

Lemma 2.3 *If $f : \{0,1\}^n \to \{0,1\}$ has a k-term l-DNF then $\bar{f}$ has an l^k-term k-DNF.*

Proof of 2.3: If we negate a k-term l-DNF then we end up with a CNF consisting of a conjunction of at most k disjunctions, with each disjunction containing at most l literals. Using the distributive law on these disjunctions gives us a disjunction of at most l^k monomials each of length at most k (since must take one literal from each of the k disjunctions in order to form a conjunction/monomial). □

For the following results we shall deal with k-term-l-DNF functions, and assume we know beforehand maximum values of k and l for our underlying concept. By previous remarks these are not unrealistic assumptions.

We will need the following lemma.

Lemma 2.4 *If we have a hypothesis $h : \{0,1\}^n \to \{0,1\}$ that is consistent with some underlying concept $c : \{0,1\}^n \to \{0,1\}$ on $A \subseteq \{0,1\}^n$, and such that there is a k_h-term l_h-DNF representing h and a k_c-term l_c-DNF representing c, then there is a DNF ϕ_m of degree $\leq \max\{k_h + l_c, k_c + l_h\}$ such that*

$$\phi_m(\mathbf{x}) = 1 \Leftrightarrow c(\mathbf{x}) \neq h(\mathbf{x})$$

Proof of 2.4: Let $\bar{h} : \{0,1\}^n \to \{0,1\}$ be the negation of h, i.e. $\bar{h}(\mathbf{x}) = 1 \Leftrightarrow h(\mathbf{x}) = 0$, similarly let $\bar{c}$ be the negation of c. Now let ϕ_h be a k_h-term l_h-DNF representing h, and ϕ_c a k_c-term l_c-DNF representing c. By lemma 2.3 there is a k_h-DNF for $\bar{h}$, say $\phi_{\bar{h}}$, and a k_c-DNF for $\bar{c}$, say $\phi_{\bar{c}}$. Now let $m : \{0,1\}^n \to \{0,1\}$ be defined by: $m(\mathbf{x}) = 1 \Leftrightarrow h(\mathbf{x}) \neq c(\mathbf{x})$, i.e. $m(\mathbf{x})$ indicates whether $\mathbf{x}$ is misclassified by h or not. Then:

$$\begin{aligned} m(\mathbf{x}) = 1 \quad &\Leftrightarrow \quad \text{either } h(\mathbf{x}) = 1 \text{ and } c(\mathbf{x}) = 0, \text{ or } h(\mathbf{x}) = 0 \text{ and } c(\mathbf{x}) = 1 \\ &\Leftrightarrow \quad \text{either } h(\mathbf{x}) = 1 \text{ and } \bar{c}(\mathbf{x}) = 1, \text{ or } \bar{h}(\mathbf{x}) = 1 \text{ and } c(\mathbf{x}) = 1 \\ &\Leftrightarrow \quad \text{either } \mathbf{x} \text{ satisfies a term of } \phi_h \text{ and a term of } \phi_{\bar{c}} \\ &\qquad\qquad \text{or } \mathbf{x} \text{ satisfies a term of } \phi_{\bar{h}} \text{ and a term of } \phi_c \end{aligned}$$

Hence we can make a DNF for m by taking all conjunctions of either a term from ϕ_h with a term from $\phi_{\bar{c}}$, or a term from $\phi_{\bar{h}}$ with a term from ϕ_c, and then forming the disjunction of all these resulting terms. The resulting DNF has terms of length at most $\max\{k_h + l_c, k_c + l_h\}$. □

Theorem 2.5 *If we have a hypothesis $h\colon \{0,1\}^n \to \{0,1\}$ that is consistent with some underlying concept $c : \{0,1\}^n \to \{0,1\}$ on $A \subseteq \{0,1\}^n$, and such that there is a k_h-term l_h-DNF representing h, and a k_c-term l_c-DNF representing c, then for any $\mathbf{x} \in \{0,1\}^n$:*

$$s(\mathbf{x}, A) \geq \max\{k_h + l_c, k_c + l_h\} \Rightarrow h(\mathbf{x}) = c(\mathbf{x})$$

Proof of 2.5: For a given DNF term T, let $V(\mathrm{T}) \subseteq \{0,1\}^n$ denote the set of all vectors satisfying T, and let $I_{\mathrm{T}} \subseteq [n]$ be the set of indices of the literals in T. Then for any $\mathbf{x} \in V(\mathrm{T})$ we have $\mathbf{x}|_{I_{\mathrm{T}}} = \mathbf{y}|_{I_{\mathrm{T}}} \Leftrightarrow \mathbf{y} \in V(\mathrm{T})$, i.e. $V(\mathrm{T})$ consists of all vectors in $\{0,1\}^n$ that are the same as $\mathbf{x}$ on the indices of I_{T}. Now let ϕ_m be defined as in Lemma 2.4, and $\mathbf{x}$ be any misclassified vector (i.e. $h(\mathbf{x}) \neq c(\mathbf{x})$). Then $\mathbf{x}$ must satisfy at least one term of ϕ_m. Let T_m be such a term, and I_{T_m} the set of indices of its literals. Then $V(\mathrm{T}_m) \cap A = \emptyset$ since h is consistent with c on A, but any vector that satisfies T_m is misclassified by c (since ϕ_m is a DNF for the misclassified vectors). Hence, putting this together with Lemma 2.4 we get:

$$h(\mathbf{x}) \neq c(\mathbf{x}) \Rightarrow \mathbf{x}|_{I_{\mathrm{T}_m}} \notin A|_{I_{\mathrm{T}_m}} \Rightarrow s(\mathbf{x}, A) < |I_{\mathrm{T}_m}| \leq \max\{k_h + l_c, k_c + l_h\}$$

for some term T_m of ϕ_m.
□
We can also say something about the number of misclassified vectors.

Corollary 2.6 *If h misclassifies one vector, then it misclassifies at least $2^{n-\max\{k_h+l_c,k_c+l_h\}}$ vectors, where k_h, l_h, k_c and l_c are as in (2.4).*

Proof of 2.6: Let ϕ_m be as in (2.4), then since there's at least one misclassified vector, ϕ_m must have at least one term, say T_m. There are $2^{n-|\mathrm{T}_m|}$ vectors that satisfy T_m, all of them misclassified by h, and by (2.4) we have

$$2^{n-\max\{k_h+l_c,k_c+l_h\}} \leq 2^{n-|\mathrm{T}_m|}$$

Hence there are at least $2^{n-\max\{k_h+l_c,k_c+l_h\}}$ misclassified vectors. □

The following simple example shows that the bound of Theorem 2.5 is tight.
Let $n = 3$, and

$$A = \left\{ \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \right\}$$

Suppose a DNF for the underlying concept, c, is $\phi_c = u_2 \vee u_3$, and a DNF for our hypothesis, h, is $\phi_h = u_1 \vee u_2$ (so h is consistent on A). Then we have $l_c = l_h = 1$, and $k_c = k_h = 2$.
The vector $v = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$ is misclassified by h, ($h(v) = 1$, but $c(v) = 0$), and we have $s(v, A) = 2 = \max\{k_h + l_c, k_c + l_h\} - 1$. (The only vectors with similarity greater than or equal to $\max\{k_h + l_c, k_c + l_h\}$ are the vectors A themselves, and h is consistent on these.)

If we are prepared to look at our hypothesis and training set in a bit more detail we can improve on the bound of Theorem 2.5.
We need a bit more notation before we can proceed.

Let A^+ be the set of all vectors in A that are labelled as 1 by our underlying concept c, and A^- the ones that are labelled 0. For a DNF term/monomial T denote by I_T the set of indices corresponding to the literals appearing in T, I_T^c the set of indices corresponding to literals not appearing in T, and V_T for the set of all vectors in $\{0,1\}^n$ that satisfy the term T.
We also need the following definition.

Definition 2.7 *For vector set $A \subseteq \{0,1\}^n$, vector $\mathbf{x} \in \{0,1\}^n$, and index set $I \in [n]$ let $s_I(\mathbf{x}, A)$ be the similarity of $\mathbf{x}$ to A if we restrict the vectors to the indices in I, i.e.*

$$s_I(\mathbf{x}, A) = s(\mathbf{x}|_I, A|_I)$$

With these definitions we can now present an improvement on Theorem 2.5.

Theorem 2.8 *If we have a hypothesis $h\colon \{0,1\}^n \to \{0,1\}$ that is consistent with some underlying concept $c : \{0,1\}^n \to \{0,1\}$ on $A \subseteq \{0,1\}^n$, and c can be represented by a k_c-term l_c-DNF (with $k_c > 0$), then for any $\mathbf{x} \in \{0,1\}^n$ we have:*

$$h(\mathbf{x}) = 1 \text{ and } s_{I^c_{T_+}}(\mathbf{x}, V_{T_+} \cap A) \geq k_c \;\Rightarrow\; c(\mathbf{x}) = 1 \tag{2}$$

$$h(\mathbf{x}) = 0 \text{ and } s_{I^c_{T_-}}(\mathbf{x}, V_{T_-} \cap A) \geq l_c \;\Rightarrow\; c(\mathbf{x}) = 0 \tag{3}$$

where T_+ is any term of h satisfied by $\mathbf{x}$, and T_- is any term of $\bar{h}$ satisfied by $\mathbf{x}$.

Proof of 2.8: Assume that $h(\mathbf{x}) = 1$ but $c(\mathbf{x}) = 0$. Then $\mathbf{x}$ satisfies a term of a DNF for h, say T, and a term of a DNF for $\bar{c}$ (the negation of c), say T'. By lemma 2.3 $\bar{c}$ has a $l_c^{k_c}$-term k_c-DNF, and so we may assume $|I_{T'}| \leq k_c$. Since h is consistent with c on A, A cannot contain any vectors that satisfy both T and T'. So none of the vectors in $V_T \cap A$ satisfy T'. Since $\mathbf{x}$ *does* satisfy T' we cannot have $\mathbf{x}|_{I_{T'} \setminus I_T} \in (V_T \cap A)|_{I_{T'} \setminus I_T}$ (since otherwise we have a vector in $V_T \cap A$ that satisfies T'). If $s_{I_T^c}(\mathbf{x}, V_T \cap A) \geq k_c$ then we must have $\mathbf{x}|_{I_{T'} \setminus I_T} \in (V_T \cap A)|_{I_{T'} \setminus I_T}$ since $(I_{T'} \setminus I_T) \subseteq I_T^c$, and $|I_{T'} \setminus I_T| \leq |I_{T'}| \leq k_c$.

This is a contradiction. Hence we must have $s_{I_T^c}(\mathbf{x}, V_T \cap A) < k_c$ in this case. The proof for (3) is similar.
□

The following corrolary shows why Theorem 2.8 is a tighter bound than Theorem 2.5.

Corollary 2.9 *For h, c, and A as is Theorem 2.5, and any $\mathbf{x} \in A^c$*

$$s(\mathbf{x}, A) \geq \max\{k_h+l_c, k_c+l_h\} \Rightarrow s_{I^c_{T_+}}(\mathbf{x}, V_{T_+}\cap A) \geq k_c \text{ or } s_{I^c_{T_-}}(\mathbf{x}, V_{T_-}\cap A) \geq l_c$$

For some term T_+ of a DNF for h satisfied by $\mathbf{x}$, or some term T_- of a DNF for $\bar{h}$ satisfied by $\mathbf{x}$.

Proof of 2.9: If $h(\mathbf{x}) = 1$ then there is a term T_+ of a DNF for h of length at most l_h satisfied by $\mathbf{x}$. By definition there is a substring of $\mathbf{x}|_{I^c_{T_+}}$ of length $s_{I^c_{T_+}}(\mathbf{x}, V_{T_+} \cap A) + 1$ not appearing in $A|_{I^c_{T_+}}$. Since $V_{T_+} \cap A$ consists of *all* vectors in A that are the same as $\mathbf{x}$ on I_{T_+}, there must be a substring of $\mathbf{x}$ of length $|I_{T_+}| + s_{I^c_{T_+}}(\mathbf{x}, V_{T_+} \cap A) + 1$ not appearing in A.
So we have:

$$s(\mathbf{x}, A) \leq |I_{T_+}| + s_{I^c_{T_+}}(\mathbf{x}, V_{T_+} \cap A) \leq l_h + s_{I^c_{T_+}}(\mathbf{x}, V_{T_+} \cap A)$$

Since $s(\mathbf{x}, A) \geq k_c + l_h$ we must have:

$$s_{I^c_{T_+}}(\mathbf{x}, V_{T_+} \cap A) \geq k_c$$

If $h(\mathbf{x}) = 0$ then there is a term T_- of a DNF for $\bar{h}$ of length at most k_h satisfied by $\mathbf{x}$. Using the same agument again but with l_c and k_h instead of k_c and l_h yields:

$$s_{I^c_{T_-}}(\mathbf{x}, V_{T_-} \cap A) \geq l_c$$

□

So any vector that satisfies the bound of Theorem 2.5 will also satisfy the bound of Theorem 2.8.
The following example shows that the converse is not true.
Let $n = 4$, and

$$A = \left\{ \begin{pmatrix} 1 \\ 1 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} \right\}$$

Suppose a DNF for the underlying concept, c, is $\phi_c = \overline{u}_2\overline{u}_3 \vee \overline{u}_1\overline{u}_4$, and a DNF for our hypothesis, h, is $\phi_h = \overline{u}_1$ (so h is consistent on A).
Then we have $l_c = k_c = 2$.
The vector $v = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$ satisfies $\overline{u}_1$, and we have

$$s_{I^c_{\overline{u}_1}}(v, V_{\overline{u}_1} \cap A) = s_{\{2,3,4\}} \left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \left\{ \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} \right\} \right) = 2 = k_c$$

However, $s(v, A) = 2 < 4 = k_c + l_c$. So Theorem 2.5 in not useful in this case, whereas Theorem 2.8 tells us that h is definitely consistent with c on v (given that we know $k_c = 2$).

Theorems 2.5 and 2.8 give some weight to the idea of using similarity measure for a hierarchy of classification confidence as proposed by Anthony and Hammer. The following example however, shows that this approach will not always work.
Let $n = 4$ and

$$A = \left\{ \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \right\}$$

Let our underlying concept c have a DNF $\phi_c = u_1u_2 \vee u_1u_3 \vee \overline{u}_1u_3\overline{u}_4$, and suppose our hypothesis h has DNF $\phi_h = u_1$ (this is the most obvious choice

for h, and it is the function that would be found using a standard ID3 method outlined in [Qui86]). Then we get $\phi_{\bar{c}} = \overline{u}_1\overline{u}_3 \vee \overline{u}_1 u_4 \vee \overline{u}_2\overline{u}_3$, and $\phi_{\bar{h}} = \overline{u}_1$. By considering terms of ϕ_c with terms of $\phi_{\bar{h}}$, and terms of $\phi_{\bar{c}}$ with terms of ϕ_h we get a DNF for the vectors misclassified by h:

$$\phi_m = \bar{u}_1 u_3 \bar{u}_4 \vee u_1 \bar{u}_2 \bar{u}_3$$

This gives the following misclassified vectors:

$$\mathbf{v_1} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \mathbf{v_2} = \begin{pmatrix} 0 \\ 1 \\ 1 \\ 0 \end{pmatrix}, \mathbf{v_3} = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \mathbf{v_4} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \end{pmatrix}$$

We have $s(\mathbf{v_1}, A) = 2$, $s(\mathbf{v_2}, A) = 1$, $s(\mathbf{v_3}, A) = 1$, and $s(\mathbf{v_4}, A) = 0$. All other vectors in A^c have similarity 0 to A since they all have a '1' in the bottom entry (whereas the vectors of A have a '0').

So we see in this example that all of the misclassified vectors have similarity measure greater than or equal to the similarity measure of the correctly classified ones. We can find examples like this of arbitrary dimension, by just adding dimensions to the vectors in this example, and including in our training set all vectors that are the same as A in the example above on the first four indices.

3 Combinatorial results

First a simple but useful result:

Theorem 3.1 *For any vector sets A and B, and any vector $\mathbf{x}$ we have:*

$$\begin{aligned} \max\{s(\mathbf{x}, A), s(\mathbf{x}, B)\} \leq s(\mathbf{x}, A \cup B) &\leq s(\mathbf{x}, A) + s(\mathbf{x}, B \backslash A) + 1 \\ &\leq s(\mathbf{x}, A) + s(\mathbf{x}, B) + 1 \end{aligned}$$

Proof of 3.1: The first inequality is obvious since $A, B \subseteq A \cup B$, and so any set of indices I for which $\mathbf{x}|_I \in A|_I$ or $\mathbf{x}|_I \in B|_I$ will also satisfy

$\mathbf{x}|_I \in (A \cup B)|_I$. The second inequality comes from the fact that $s(\mathbf{x}, A)$ is one less than the size of the smallest set of indices on which $\mathbf{x}$ differs from all vectors in A, and that $\mathbf{x}|_I \notin A|_I$ and $\mathbf{x}|_J \notin (B \backslash A)|_J \Rightarrow \mathbf{x}|_{I \cup J} \notin (A \cup B)|_{I \cup J}$. So we get $s(\mathbf{x}, A \cup B) + 1 \leq (s(\mathbf{x}, A) + 1) + (s(\mathbf{x}, B \backslash A) + 1)$.
The third inequality comes from $(B \backslash A) \subseteq B$. □

It would be useful to have some measure of how representative a dataset is of the whole of $\{0,1\}^n$. The following two definitions (introduced in [AH04]) attempt to do just that. Assume from now on that A is a proper non-empty subset of $\{0,1\}^n$ (i.e. $\emptyset \neq A \neq \{0,1\}^n$).

Definition 3.2 *The* pervasiveness*, $P(A)$, of $A \subseteq \{0,1\}^n$ is defined to be the minimum similarity of $\mathbf{x} \in \{0,1\}^n$ to A; that is,*

$$P(A) = \min_{\mathbf{x} \in \{0,1\}^n} s(\mathbf{x}, A) = \min_{\mathbf{x} \in A^c} s(\mathbf{x}, A)$$

Definition 3.3 *The* extent *of $A \subseteq \{0,1\}^n$, $e(A)$, is defined to be the maximum similarity of $\mathbf{x} \notin A$ to A; that is,*

$$e(A) = \max_{\mathbf{x} \in A^c} s(\mathbf{x}, A)$$

Note that, for $A \neq \{0,1\}^n$, $0 \leq P(A) \leq e(A) \leq n - 1$.
Some alternative characterisations of pervasiveness and extent are given in [AH04].
We now present some more relationships linking $P(A)$, $e(A)$, and $|A|$ (note that $|A^c| = 2^n - |A|$). The following definitions will be needed for the proof of Theorem 3.6.

Definition 3.4 *A set of indices $I \subseteq [n]$ is* shattered *by a vector set A if $A|_I$ contains all possible binary vectors of dimension $|I|$.*

Definition 3.5 *The* VC-dimension *of a vector set A (written* VCdim(A)*) is the size of the largest set of indices that is shattered by A.*

Note that we must have $2^{\mathrm{VCdim}(A)} \leq |A|$, and so $\mathrm{VCdim}(A) \leq \lfloor \log_2 |A| \rfloor$.

Theorem 3.6 *With the above notation,*

$$n - 1 - \lfloor \log_2 |A^c| \rfloor \leq P(A) \leq \lfloor \log_2 |A| \rfloor$$

Proof of 3.6: First the lower bound.
We must have some vector $\mathbf{x} \in A^c$ such that $s(\mathbf{x}, A) = P(A)$ (by definition of $P(A)$). So there exists some $I \subseteq [n]$ such that $|I| = P(A)+1$, and $\mathbf{x}|_I \notin A|_I$. Therefore for all $\mathbf{y}$ such that $\mathbf{y}|_I = \mathbf{x}|_I$, we have $\mathbf{y} \in A^c$, and so

$$2^{n-(P(A)+1)} \leq |A^c| \tag{4}$$

Taking logarithms and rearranging gives $n - 1 - \lfloor \log_2 |A^c| \rfloor \leq P(A)$.
For the upper bound, let $I \subseteq [n]$ such that $|I| = P(A)$. Then we have:

$$\begin{array}{lll} s(\mathbf{x}, A) & \geq P(A) & \forall \mathbf{x} \in \{0,1\}^n \\ \Rightarrow & \mathbf{x}|_I \in A|_I & \forall \mathbf{x} \in \{0,1\}^n \\ \Rightarrow & I \text{ is shattered by } A & \\ \Rightarrow & |I| \leq \mathrm{VCdim}(A) & \\ \Rightarrow & P(A) = |I| \leq \mathrm{VCdim}(A) \leq \lfloor \log_2 |A| \rfloor & \end{array}$$

□

Proposition 3.7 *For any non-empty $A \subset \{0,1\}^n$, the pervasiveness and extent satisfy,*

$$P(A) \leq e(A) < |A|$$

Proof of 3.7: The first inequality comes from the definitions of $P(A)$ and $e(A)$.
Now, let $\mathbf{x}$ be a vector with $s(\mathbf{x}, A) = e(A)$. For each vector $\mathbf{y} \in A$ we can find an index that differentiates it from $\mathbf{x}$, this gives a total of at most $|A|$ indices to distinguish $\mathbf{x}$ from all vectors in A. Hence $s(\mathbf{x}, A) < |A|$, i.e. $e(A) < |A|$. □

4 Asymptotic results

In this section we present some asymptotic results on the expected value of the similarity measure, and pervasiveness of a random vector set.

Theorem 4.1 *For any* $\mathbf{x} \in \{0,1\}^n$*, a random vector set* A*, and any constant* $0 < p < 1$*, then almost surely we have,*

$$\lim_{n\to\infty} \left[s(\mathbf{x}, A) - \left(n - \left\lfloor \log_2 \left[\log_2(\log_{\frac{1}{1-p}}(n)) \log_{\frac{1}{1-p}}(n) \right] \right\rfloor \right) \right] = 0$$

Where A *is chosen according to the distribution* $U_{\mathbf{x}}$ *in which each vector in* $\{0,1\}^n \backslash \{\mathbf{x}\}$ *is chosen independently with probability* p *to be in* A*.*

Proof of 4.1: Note that by symmetry the distribution of $s(\mathbf{x}, A)$ under $U_{\mathbf{x}}$ is the same for all vectors $\mathbf{x}$, hence we can limit our investigations to $s(\mathbf{0}, A)$ under $U_{\mathbf{0}}$.
For vector $\mathbf{x} \neq \mathbf{0}$ let

$$\mathcal{D}_{\mathbf{x}} = \{\mathbf{y} \in \{0,1\}^n : \forall i \in [n], \mathbf{x}_i = 0 \Rightarrow \mathbf{y}_i = 0\}$$

So $\mathcal{D}_{\mathbf{x}}$ consists of all vectors that have at least the same 0's as $\mathbf{x}$ and possibly more. Note that there is a one-to-one correspondence between the vectors of weight d and the index sets $I \subseteq [n]$ such that $|I| = n - d$.

We have:

$$\begin{aligned} s(\mathbf{0}, A) - (n-d) \geq 0 \quad &\Leftrightarrow \quad s(\mathbf{0}, A) \geq n - d \\ &\Leftrightarrow \quad \forall I \subseteq [n] \text{ such that } |I| = n-d,\ \mathbf{0}|_I \in A|_I \\ &\Leftrightarrow \quad \forall \mathbf{x} \in \{0,1\}^n \text{ such that } w(\mathbf{x}) = d,\ A \cap \mathcal{D}_{\mathbf{x}} \neq \emptyset \\ &\Leftrightarrow \quad \nexists \mathbf{x} \in \{0,1\}^n \text{ such that } w(\mathbf{x}) = d \text{ and } \mathcal{D}_{\mathbf{x}} \subseteq A^c \end{aligned} \tag{5}$$

Now let us assume that, apart from $\mathbf{0}$, each vector in $\{0,1\}^n$ is chosen independently with probability p to be in A, ($\mathbf{0}$ is not chosen).

For any vector $\mathbf{x}$ of weight d we have $|\mathcal{D}_{\mathbf{x}}\backslash\{\mathbf{0}\}| = 2^d - 1$, and so:

$$\mathbb{P}_{U_{\mathbf{0}}}\left(\mathcal{D}_{\mathbf{x}} \subseteq A^c\right) = (1-p)^{2^d-1}$$

(Remember $\mathbf{0}$ is in A^c with probability 1.)
There are $\binom{n}{d}$ vectors of weight d. Hence by the union bound:

$$\mathbb{P}_{U_{\mathbf{0}}}\left(\exists \mathbf{x} \in \{0,1\}^n : \ w(\mathbf{x}) = d, \text{ and } \mathcal{D}_{\mathbf{x}} \subseteq A^c\right) \leq \binom{n}{d}(1-p)^{2^d-1} < n^d(1-p)^{2^d-1}$$

If we let $d = \left\lfloor \log_2 \left[b \log_2(\log_{\frac{1}{1-p}}(n)) \log_{\frac{1}{1-p}}(n) \right] \right\rfloor + 1$ for any fixed $b > \frac{1}{2}$ then this gives:

$$\begin{aligned} n^d(1-p)^{2^d-1} &= n^d(1-p)^{2[b\log_2(\log_{\frac{1}{1-p}}(n))\log_{\frac{1}{1-p}}(n)]-1} \\ &= \frac{n^{d-2b\log_2(\log_{\frac{1}{1-p}}(n))}}{1-p} \\ &= \frac{n^{(1-2b)\log_2(\log_{\frac{1}{1-p}}(n))+\log_2[b\log_2(\log_{\frac{1}{1-p}}(n))]}}{1-p} \\ &\to 0 \quad \text{as } n \to \infty \end{aligned}$$

Since $1 - 2b < 0$, and the first term of the exponent of n grows much faster than the second we get

$$\frac{n^{(1-2b)\log_2(\log_{\frac{1}{1-p}}(n))+\log_2[b\log_2(\log_{\frac{1}{1-p}}(n))]}}{1-p} \longrightarrow 0 \quad \text{as} \quad n \longrightarrow \infty$$

and so by (5), with probability 1, for any $b > \frac{1}{2}$:

$$\begin{aligned} & \lim_{n\to\infty} \left[s(\mathbf{0}, A) - \left(n - \left\lfloor \log_2 \left[b \log_2(\log_{\frac{1}{1-p}}(n)) \log_{\frac{1}{1-p}}(n) \right] \right\rfloor - 1 \right) \right] \geq 0 \\ \Rightarrow \ & \lim_{n\to\infty} \left[s(\mathbf{0}, A) - \left(n - \left\lfloor \log_2 \left[\log_2(\log_{\frac{1}{1-p}}(n)) \log_{\frac{1}{1-p}}(n) \right] \right\rfloor - \lfloor \log_2(b) \rfloor - 1 \right) \right] \geq 0 \\ \Rightarrow \ & \lim_{n\to\infty} \left[s(\mathbf{0}, A) - \left(n - \left\lfloor \log_2 \left[\log_2(\log_{\frac{1}{1-p}}(n)) \log_{\frac{1}{1-p}}(n) \right] \right\rfloor \right) \right] \geq 0 \end{aligned} \tag{6}$$

(Since $\lfloor \log_2(b) \rfloor = -1$ for b close to $\frac{1}{2}$.)

For an upper bound, let

$$\#\mathcal{D}_d = |\{\mathbf{x} : w(\mathbf{x}) = d \text{ and } \mathcal{D}_{\mathbf{x}} \subseteq A^c\}|$$

I shall proceed by showing that for certain values of d, $\lim_{n\to\infty} \mathbb{P}_{U_{\mathbf{0}}}(\#\mathcal{D}_d = 0) = 0$, and so by (5) this will imply $\lim_{n\to\infty}[s(\mathbf{0}, A) - (n-d)] < 0$ almost surely. We have already seen that if $w(\mathbf{x}) = d$ then we have

$$\mathbb{P}_{U_{\mathbf{0}}}(\mathcal{D}_{\mathbf{x}} \subseteq A^c) = (1-p)^{2^d-1}$$

hence

$$\mathbb{E}(\#\mathcal{D}_d) = \binom{n}{d}(1-p)^{2^d-1}$$

(since there are $\binom{n}{d}$ vectors of weight d).
For $\alpha \subseteq [n]$ let 1_α be an indicator variable which is equal to $1 \Leftrightarrow \mathcal{D}_{\mathbf{0}^\alpha} \subseteq A^c$, and 0 otherwise, where $\mathbf{0}^\alpha$ is the weight $|\alpha|$ vector with 0's at every index apart from those in α.
Then we have:

$$\begin{aligned}
\mathbb{E}_{U_{\mathbf{0}}}\left(\#\mathcal{D}_d^2\right) &= \mathbb{E}_{U_{\mathbf{0}}}\Big[\Big(\sum_{\alpha:|\alpha|=d} 1_\alpha\Big)^2\Big] \\
&= \mathbb{E}_{U_{\mathbf{0}}}\left(\sum_{\alpha:|\alpha|=d}\sum_{\beta:|\beta|=d}[1_\alpha \times 1_\beta]\right) \\
&= \sum_{\alpha:|\alpha|=d}\sum_{\beta:|\beta|=d}\mathbb{P}_{U_{\mathbf{0}}}(\mathcal{D}_{\mathbf{0}^\alpha} \subseteq A^c \text{ and } \mathcal{D}_{\mathbf{0}^\beta} \subseteq A^c) \\
&= \sum_{\alpha:|\alpha|=d}\Big\{\mathbb{P}_{U_{\mathbf{0}}}(\mathcal{D}_{\mathbf{0}^\alpha} \subseteq A^c)\sum_{\beta:|\beta|=d}\mathbb{P}_{U_{\mathbf{0}}}(\mathcal{D}_{\mathbf{0}^\beta}\backslash\mathcal{D}_{\mathbf{0}^\alpha} \subseteq A^c)\Big\} \\
&= \sum_{\alpha:|\alpha|=d}\Big\{(1-p)^{2^d-1}\sum_{l=0}^{d}\sum_{\beta:|\alpha\cap\beta|=l}\mathbb{P}_{U_{\mathbf{0}}}(\mathcal{D}_{\mathbf{0}^\beta}\backslash\mathcal{D}_{\mathbf{0}^\beta}\cap\mathcal{D}_{\mathbf{0}^\alpha} \subseteq A^c)\Big\} \\
&= \binom{n}{d}(1-p)^{2^d-1}\sum_{l=0}^{d}\binom{n-d}{d-l}\binom{d}{l}(1-p)^{2^d-1-(2^l-1)}
\end{aligned}$$

Chebyshev's inequality tells us that for a non-negative random variable x with mean μ and variance σ, and any $\epsilon > 0$ we have:

$$\mathbb{P}(|x-\mu| \geq \epsilon) \leq \frac{\sigma^2}{\epsilon^2}$$

Taking $\epsilon = \mu$ we get:

$$\mathbb{P}\left(|x-\mu| \geq \mu\right) \leq \frac{\sigma^2}{\mu^2}$$

$$\text{that is} \qquad \mathbb{P}\left(x = 0 \text{ or } x \geq 2\mu\right) \leq \frac{\sigma^2}{\mu^2}$$

$$\text{so} \qquad \mathbb{P}\left(x = 0\right) \leq \frac{\sigma^2}{\mu^2} \tag{7}$$

Applying (7) to the random variable $\#\mathcal{D}_d$ we get:

$$\begin{aligned}
&\mathbb{P}_{U_{\mathbf{0}}}\left(\#\mathcal{D}_d = 0\right) \leq \frac{\mathbb{E}_{U_{\mathbf{0}}}\left(\#\mathcal{D}_d^2\right) - \mathbb{E}_{U_{\mathbf{0}}}\left(\#\mathcal{D}_d\right)^2}{\mathbb{E}_{U_{\mathbf{0}}}\left(\#\mathcal{D}_d\right)^2} \\
&= \frac{\left[\binom{n}{d}(1-p)^{2^d-1}\sum_{l=0}^{d}\binom{n-d}{d-l}\binom{d}{l}(1-p)^{2^d-1-(2^l-1)}\right] - \binom{n}{d}^2(1-p)^{2^{(d+1)}-2}}{\binom{n}{d}^2(1-p)^{2^{(d+1)}-2}} \\
&= \frac{\sum_{l=0}^{d}\binom{n-d}{d-l}\binom{d}{l}(1-p)^{1-2^l} - \binom{n}{d}}{\binom{n}{d}} \\
&= \binom{n}{d}^{-1}\sum_{l=0}^{d}\binom{n-d}{d-l}\binom{d}{l}\left((1-p)^{1-2^l}-1\right) \quad \left(\text{since } \tbinom{n}{d} = \textstyle\sum_{l=0}^{d}\tbinom{n-d}{d-l}\tbinom{d}{l}\right) \\
&< \binom{n}{d}^{-1}\sum_{l=1}^{d}\binom{n-d}{d-l}d^l\Big(\frac{1}{1-p}\Big)^{2^l} \\
&< \sum_{l=1}^{d}\frac{(n-d)!^2}{n!(n-2d+l)!}\,d^{2l}\Big(\frac{1}{1-p}\Big)^{2^l} \\
&= \sum_{l=1}^{d}\frac{(n-d)\cdots(n-2d+l+1)}{n(n-1)\cdots(n-d+1)}\,d^{2l}\Big(\frac{1}{1-p}\Big)^{2^l}
\end{aligned}$$

There are d terms on the top of the first fraction, and $d-l$ terms on the bottom, hence dividing top and bottom by n^d gives,

$$\mathbb{P}_{U_{\mathbf{0}}}\left(\#\mathcal{D}_d = 0\right) < \sum_{l=1}^{d}\frac{\left(1-\frac{d}{n}\right)\cdots\left(1-\frac{(2d-l-1)}{n}\right)}{1\left(1-\frac{1}{n}\right)\cdots\left(1-\frac{d-1}{n}\right)}\,\frac{d^{2l}}{n^l}\Big(\frac{1}{1-p}\Big)^{2^l}$$

For every term of the sum, the top of the first fraction is a product of factors smaller than 1, with no factor smaller than $1-\frac{2d}{n}$. So, for each term, taking

just d of these factors and dividing by $\left(1-\frac{2d}{n}\right)^d$ we get,

$$\mathbb{P}_{U_{\mathbf{0}}}\left(\#\mathcal{D}_d=0\right) < \frac{1}{\left(1-\frac{2d}{n}\right)^d}\frac{\left(1-\frac{d}{n}\right)\cdots\left(1-\frac{(2d-1)}{n}\right)}{1\left(1-\frac{1}{n}\right)\cdots\left(1-\frac{d-1}{n}\right)}\sum_{l=1}^{d}\left(\frac{d^2}{n}\right)^l\left(\frac{1}{1-p}\right)^{2^l}$$

If we pair up the factors on the top and bottom of the second fraction we see that the top is smaller, hence this fraction is less than 1. Also if we take $n > 4d$, we get $\frac{1}{(1-\frac{2d}{n})} > 2$, hence;

$$\mathbb{P}_{U_{\mathbf{0}}}\left(\#\mathcal{D}_d=0\right) < 2^d\sum_{l=1}^{d}\left(\frac{d^2}{n}\right)^l\left(\frac{1}{1-p}\right)^{2^l} \qquad (\text{for } n>4d). \tag{8}$$

Let $d = \left\lfloor \log_2\left[b\log_2(\log_{\frac{1}{1-p}}(n))\log_{\frac{1}{1-p}}(n)\right]\right\rfloor$ for any constant $b < \frac{1}{2}$. Then for large n we have $\frac{d^2}{n} < 1$ and $\frac{1}{1-p} > 1$. Hence for large n the terms in (8) either increase with l, or decrease at first and then increase.
Hence:

$$\mathbb{P}_{U_{\mathbf{0}}}\left(\#\mathcal{D}_d=0\right) < 2^d\, d\left[\left(\frac{d^2}{n}\right)\left(\frac{1}{1-p}\right)^2+\left(\frac{d^2}{n}\right)^d\left(\frac{1}{1-p}\right)^{2^d}\right]$$

For d as above we have:

$$2^d\, d\Big(\frac{d^2}{n}\Big)\Big(\frac{1}{1-p}\Big)^2 \le \left(\frac{b\log_2(\log_{\frac{1}{1-p}}(n))\log_{\frac{1}{1-p}}(n)d^3}{n}\right)\Big(\frac{1}{1-p}\Big)^2$$

$$= \left(\frac{b\log_2(\log_{\frac{1}{1-p}}(n))\log_{\frac{1}{1-p}}(n)\left\lfloor\log_2\left[b\log_2(\log_{\frac{1}{1-p}}(n))\log_{\frac{1}{1-p}}(n)\right]\right\rfloor^3}{n}\right)\Big(\frac{1}{1-p}\Big)^2$$

$$\longrightarrow 0 \quad \text{as} \quad n\to\infty$$

Also:

$$2^d\, d\Big(\frac{d^2}{n}\Big)^d\Big(\frac{1}{1-p}\Big)^{2^d} \le 2^d\,\frac{d^{2d+1}}{n^d}n^{b\log_2(\log_{\frac{1}{1-p}}(n))} = 2^d\,\frac{d^{2d+1}}{n^{d-(b\log_2(\log_{\frac{1}{1-p}}(n)))}}$$

$$
\begin{aligned}
&< \frac{2^d d^{3d}}{n^{\log_2\ b\log_2(\log_{\frac{1}{1-p}}(n))\log_{\frac{1}{1-p}}(n)\ -\ b\log_2(\log_{\frac{1}{1-p}}(n))\ -1}} \\
&= \frac{(2d^3)^d}{n^{(1-b)\log_2(\log_{\frac{1}{1-p}}(n))+\log_2(b\log_2(\log_{\frac{1}{1-p}}(n)))-1}} \\
&< \frac{(2d^3)^d}{n^{\frac{1}{2}\log_2(\log_{\frac{1}{1-p}}(n))+\frac{1}{2}\log_2\ b\log_2(\log_{\frac{1}{p}}(n))}} \qquad \text{since } b < \frac{1}{2} \\
&\leq \frac{(2d^3)^d}{n^{\frac{1}{2}d}} = \left(\frac{2d^3}{\sqrt{n}}\right)^d \longrightarrow 0 \quad \text{as} \quad n \longrightarrow \infty
\end{aligned}
$$

(Since $\sqrt{n}$ grows much faster than d^3.)

Hence for this value of d, as $n \to \infty$, $\mathbb{P}_{U_{\mathbf{0}}}(\#\mathcal{D}_d = 0) \to 0$,
i.e. $\mathbb{P}_{U_{\mathbf{0}}}(\nexists \mathbf{x} \in \{0,1\}^n$ such that $w(\mathbf{x}) = d \; and \; \mathcal{D}_{\mathbf{x}} \subseteq A^c) \to 0$
By (5) this implies that with probability 1, for any $b < \frac{1}{2}$:

$$
\begin{aligned}
& \lim_{n\to\infty} \left[s(\mathbf{0},A) - \left(n - \left\lfloor \log_2\left[b\log_2(\log_{\frac{1}{1-p}}(n))\log_{\frac{1}{1-p}}(n)\right]\right\rfloor\right)\right] < 0 \\
\Rightarrow\ & \lim_{n\to\infty} \left[s(\mathbf{0},A) - \left(n - \left\lfloor \log_2\left[\log_2(\log_{\frac{1}{1-p}}(n))\log_{\frac{1}{1-p}}(n)\right]\right\rfloor - \lceil\log_2(b)\rceil\right)\right] < 0 \\
\Rightarrow\ & \lim_{n\to\infty} \left[s(\mathbf{0},A) - \left(n - \left\lfloor \log_2\left[\log_2(\log_{\frac{1}{1-p}}(n))\log_{\frac{1}{1-p}}(n)\right]\right\rfloor\right)\right] < 1 \qquad (9)
\end{aligned}
$$

(Since $\lceil\log_2(b)\rceil = -1$ for b close to $\frac{1}{2}$.)
Finally, $\left[s(\mathbf{0},A) - \left(n - \left\lfloor \log_2\left[\log_2(\log_{\frac{1}{1-p}}(n))\log_{\frac{1}{1-p}}(n)\right]\right\rfloor\right)\right]$ must be an integer, hence putting (6) and (9) together, and by the comments at the beginning of the proof we have proved the theorem. □

Note that $\log_2(\log_{\frac{1}{1-p}}(n)) = \log_2(\log_2(n)) - \log_2(\log_2(\frac{1}{1-p}))$, and so the contribution due to p in Theorem 4.1 is independent of n.

In Theorem 4.1 the size of our set A is a random variable dependent on fixed p and n, but what about for fixed $|A|$?
In this case the probability of finding an index in which all $|A|$ vectors have entry 1 tends to one as the number of indices tends to infinity, and so we get $\lim_{n\to\infty} s(\mathbf{0},A) = 0$ (almost surely) in this case.

Theorem 4.2 *For fixed $0 < p < 1$, and a randomly chosen vector set A, almost surely we have,*

$$\lim_{n\to\infty}\left[P(A) - \left(n - \lfloor \log_2(n \log_{\frac{1}{1-p}}(2))\rfloor - 1\right)\right] = 0$$

where A is chosen according to the distribution U in which each vector from $\{0,1\}^n$ is chosen independently with probability p.

Proof of 4.2:
The proof follows the same format as that for Theorem 4.1, i.e. the probabilistic method.
First a definition. A *k-cube* is a set, $C_k \subseteq \{0,1\}^n$, of 2^k vectors such that there is some index set $I \subseteq [n]$ with $|I| = n-k$ and $\mathbf{x}|_I = \mathbf{y}|_I \ \forall \mathbf{x}, \mathbf{y} \in C_k$ (the *dimensions* of the cube are the indices in $[n]\backslash I$).
This leads to the following:

$$\begin{aligned} P(A) \leq n-k-1 &\Leftrightarrow \exists \mathbf{x} \in A^c \text{ and } I \in [n] \text{ with } |I| = n-k \text{ such that } \mathbf{x}|_I \notin A|_I \\ &\Leftrightarrow \exists \text{ a } k\text{-cube in } A^c \text{ (since } \mathbf{x}|_I \notin A|_I \text{ and } \mathbf{y}|_I = \mathbf{x}|_I \Rightarrow \mathbf{y} \notin A) \end{aligned} \tag{10}$$

For each index set there are 2^{n-k} possible 0-1 patterns on those indices, and there are $\binom{n}{n-k} = \binom{n}{k}$ different index sets of size $n-k$. This gives a total of $\binom{n}{k}2^{n-k}$ k-cubes. Putting this together with (10) gives:

$$\begin{aligned} \mathbb{P}_U\left(P(A) - (n-k-1) \leq 0\right) &= \mathbb{P}_U\left(P(A) \leq n-k-1\right) \\ &= \mathbb{P}_U\left(\exists \text{ a } k\text{-cube in } A^c\right) \\ &< \binom{n}{k} 2^{n-k}(1-p)^{2^k} \text{ (by the union bound)} \\ &< \frac{n^k}{k!} 2^{n-k}(1-p)^{2^k} \\ &< 2^n n^k (1-p)^{2^k} \end{aligned}$$

Let $k = \lfloor \log_2(n \log_{\frac{1}{1-p}}(2+\epsilon))\rfloor = \lfloor \log_2(\log_{\frac{1}{1-p}}((2+\epsilon)^n))\rfloor$ for some $\epsilon > 0$.
Then:

$$2^n n^k (1-p)^{2^k} < n^k \left(\frac{2}{2+\epsilon}\right)^n = \mathrm{e}^{k\log_2(n) + n\log_2(\frac{2}{2+\epsilon})} \to 0 \text{ as } n \to \infty \tag{11}$$

(Since $\log_2(\frac{2}{2+\epsilon}) < 0$ and n grows much faster than $k\log_2(n)$.)
So for this value of k we have $\lim_{n\to\infty}[P(A) - (n-k-1)] \geq 0$ almost surely.
i.e.

$$\lim_{n\to\infty}\left[P(A) - (n - \lfloor \log_2(n\log_{\frac{1}{1-p}}(2+\epsilon))\rfloor - 1)\right] \geq 0 \tag{12}$$

Now for the upper bound.
Let $\#C_k$ denote the number of k-cubes in A^c. We shall use Chebyshev's inequality again to find a value of k such that as $n \to \infty$, $\#C_k > 0$ almost surely. By (10) this will imply that $\lim_{n\to\infty}[P(A) - (n-k-1) \leq 0]$ almost surely.
We have already seen that the number of different k-cubes in $\{0,1\}^n$ is $\binom{n}{k}2^{n-k}$, and the probability of a given k-cube appearing in A^c is $(1-p)^{2^k}$. Hence:

$$\mathbb{E}_U\left(\#C_k\right) = \binom{n}{k}2^{n-k}(1-p)^{2^k} \tag{13}$$

Let 1_{C_k} be an indicator variable which is equal to $1 \Leftrightarrow C_k \in A^c$, and 0 otherwise.
Then we get:

$$\mathbb{E}_U\left(\#C_k^2\right) = \mathbb{E}_U\left[\left(\sum_{C_k} 1_{C_k}\right)^2\right]$$

$$= \mathbb{E}_U\left(\sum_{C_k}\sum_{C_k'} 1_{C_k}1_{C_k'}\right)$$

$$= \sum_{C_k}\sum_{C_k'}\mathbb{P}_U\left(C_k \subseteq A^c \text{ and } C_k' \subseteq A^c\right)$$

$$= \sum_{C_k}\mathbb{P}_U\left(C_k \subseteq A^c\right)\sum_{C_k'}\mathbb{P}_U\left(C_k'\backslash C_k \subseteq A^c\right)$$

$$= \sum_{C_k}\mathbb{P}_U\left(C_k \subseteq A^c\right)\sum_{i=0}^{k}\sum_{\substack{C_k': \\ C_k' \text{ shares } i \\ \text{dimensions} \\ \text{with } C_k}}\mathbb{P}_U\left(C_k'\backslash C_k \subseteq A^c\right)$$

$$= \sum_{C_k} \mathbb{P}_U\left(C_k \subseteq A^c\right) \sum_{i=0}^{k} \Big[\sum_{\substack{C'_k : C'_k \text{ shares} \\ i \text{ dimensions} \\ \text{with } C_k \text{ and} \\ C'_k \cap C_k = \emptyset}} \mathbb{P}_U\left(C'_k \subseteq A_c\right) + \sum_{\substack{C''_k : C''_k \text{ shares} \\ i \text{ dimensions} \\ \text{with } C_k \text{ and} \\ C''_k \cap C_k \neq \emptyset}} \mathbb{P}_U\left(C''_k \backslash C_k \subseteq A_c\right) \Big]$$

$$= \binom{n}{k} 2^{n-k} (1-p)^{2^k} \sum_{i=0}^{k} \binom{n-k}{k-i} \binom{k}{i} \left[\left(2^{n-k} - 2^{k-i}\right) (1-p)^{2^k} + 2^{k-i} (1-p)^{\left(2^k - 2^i\right)} \right]$$

$$= \binom{n}{k} 2^{n-k} (1-p)^{2^k} \sum_{i=0}^{k} \binom{n-k}{k-i} \binom{k}{i} (1-p)^{2^k} \left[2^{n-k} + 2^{k-i} \left((1-p)^{-2^i} - 1 \right) \right] \quad (14)$$

By Chebyshev's inequality (7) we get:

$$\mathbb{P}_U\left(\#C_k = 0\right) \leq \frac{\mathbb{E}_U\left(\#C_k^2\right) - \mathbb{E}_U\left(\#C_k\right)^2}{\mathbb{E}_U\left(\#C_k\right)^2}$$

Substituting in (13) and (14) gives:

$$\begin{aligned}
& \mathbb{P}_U\left(\#C_k = 0\right) \\
\leq\ & \frac{\sum_{i=0}^{k} \binom{n-k}{k-i} \binom{k}{i} (1-p)^{2^k} \left[2^{n-k} + 2^{k-i} \left((1-p)^{-2^i} - 1 \right) \right] - \binom{n}{k} 2^{n-k} (1-p)^{2^k}}{\binom{n}{k} 2^{n-k} (1-p)^{2^k}} \\
=\ & \frac{\sum_{i=0}^{k} \binom{n-k}{k-i} \binom{k}{i} \left[2^{n-k} + 2^{k-i} \left((1-p)^{-2^i} - 1 \right) \right] - \binom{n}{k} 2^{n-k}}{\binom{n}{k} 2^{n-k}} \\
=\ & \frac{\sum_{i=0}^{k} \binom{n-k}{k-i} \binom{k}{i} 2^{k-i} \left((1-p)^{-2^i} - 1 \right)}{\binom{n}{k} 2^{n-k}} \qquad \left(\text{since } \tbinom{n}{k} = \textstyle\sum_{i=0}^{k} \tbinom{n-k}{k-i} \tbinom{k}{i}\right) \\
=\ & \binom{n}{k}^{-1} \sum_{i=0}^{k} \binom{n-k}{k-i} \binom{k}{i} 2^{-n+2k-i} \left((1-p)^{-2^i} - 1 \right) \\
<\ & 2^k \sum_{i=0}^{k} \left(\frac{k^2}{n} \right)^i 2^{-n+2k-i} \left(\frac{1}{1-p} \right)^{2^i} \quad \text{(by the same sequence of steps as in (8))} \\
=\ & 2^{-n+3k} \sum_{i=0}^{k} \left(\frac{k^2}{2n} \right)^i \left(\frac{1}{1-p} \right)^{2^i} \qquad (15)
\end{aligned}$$

(for $n > 4k$).

Let $k = \lfloor \log_2(n \log_{\frac{1}{1-p}}(2)) \rfloor = \lfloor \log_2(\log_{\frac{1}{1-p}}(2^n)) \rfloor$. Then for large n we have $\frac{k^2}{2n} < 1$, also $\frac{1}{1-p} > 1$. Hence for large n the terms in (15) either increase with i, or decrease at first and then increase.
Hence:

$$\mathbb{P}_U\left(\#C_k = 0\right) < \frac{(k+1)}{2^{n-3k}} \left[\frac{1}{1-p} + \left(\frac{k^2}{2n}\right)^k \left(\frac{1}{1-p}\right)^{2^k} \right]$$

Obviously for k as above we get:

$$\frac{(k+1)}{2^{n-3k}} \frac{1}{1-p} \to 0 \quad \text{as} \quad n \to \infty$$

also:

$$\begin{aligned} \frac{(k+1)}{2^{n-3k}} \left(\frac{k^2}{2n}\right)^k \left(\frac{1}{1-p}\right)^{2^k} &\leq \frac{(k+1)}{2^{n-3k}} \left(\frac{k^2}{2n}\right)^k 2^n \\ &= (k+1) \left(\frac{4k^2}{n}\right)^k \to 0 \\ &\quad \text{as } n \to \infty \end{aligned} \tag{16}$$

So for this value of k we almost surely have a k-cube in A^c as $n \to \infty$.
Hence with probability 1 we have:

$$\lim_{n\to\infty} \left[P(A) - (n-k-1)\right] \leq 0$$

i.e.

$$\lim_{n\to\infty} \left[P(A) - (n - \lfloor \log_2(n \log_{\frac{1}{1-p}}(2)) \rfloor - 1)\right] \leq 0 \tag{17}$$

Since (12) is true for arbitrarily small ϵ, putting (12) and (17) together proves the theorem. □

There are a couple of things to note about the previous proof. Firstly, since $\log_2(n \log_{\frac{1}{1-p}}(2)) = \log_2(n) - \log_2(\log_2(\frac{1}{1-p}))$, we get that, as in Theorem 4.1, the contribution due to p is independent of n. Secondly the rate of convergence of the lower bound is highly dependent on ϵ in (11), and is much slower than that for the upper bound (16) which is independent of ϵ. So we can expect slow convergence from below to the value in the statement of the theorem.

References

[AB99] M. Anthony and P. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.

[ABST95] M. Anthony, G. Brightwell, and J. Shawe-Taylor. On specifying boolean functions by labelled examples. *Discrete Applied Mathematics*, pages 1–25, 1995.

[AH04] M. Anthony and P.L. Hammer. A boolean measure of similarity. research report 27-2004, Rutgers University, RUTCOR, Rutgers University, 640 Bartholomew Road, Piscataway, New Jersey 08854-8003, U.S.A., August 2004.

[Ant01] M. Anthony. *Discrete Mathematics of Neural Networks*, chapter 2, pages 9–18. SIAM Monographs on Discrete Mathematics and Applications. SIAM, Philadelphia, 2001.

[BEHW89] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Learnability and the vapnik-chervonenkis dimension. *Journal of the ACM*, 36(4):929–965, 1989.

[BHIK97] E. Boros, P.L. Hammer, T. Ibaraki, and A. Kogan. Logical analysis of numerical data. research report 04-97, Rutgers University, RUTCOR, Rutgers University, 640 Bartholomew Road, Piscataway, New Jersey 08854-8003, U.S.A., February 1997.

[GK95] S. A. Goldman and M. J. Kearns. On the complexity of teaching. *Journal of Computer and Systems Sciences*, 1(50):20–31, 1995.

[Heg94] T. Hegedüs. Combinatorial results on the complexity of teaching and learning. In *Proceeding of the* 19^{th} *International Symposium on Mathematical Foundations of Computer Science*. Springer-Verlag, August 1994.

[HSS04] P.L. Hammer, E. Subasi, and M. Subasi. Classification results from similarity measure experiments. Personal correspondance with Martin Anthony, 2004.

[KLRS96] E. Kushilevitz, N. Linial, Y. Rabinovich, and M. Saks. Witness sets for families of binary vectors. *Journal of Combinatorial Theory*, pages 376–380, 1996.

[MS91] S. Miyano and A. Shinohara. Teachability in computational learning. *New Generation Computing*, pages 337–347, 1991.

[Qui86] J.R. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986.

[Val84] L. G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

[Vov02] V. Vovk. On-line confidence machines are well-calibrated. www.vovk.net/kp, April 2002.