HOMEWORK 4:
WRITTEN EXERCISE PART

>>Arpit Jain<<
>>ajain74<<

1 Balls and Bins [25/3 pts]

Suppose we throw balls into n bins. Each ball is thrown independently and uniformly at random.
(1) [Birthday Paradox] Suppose we throw m balls. What is the probability that at least one bin has more than one balls? Write down the expression and then use the inequality $1 - x \leq e^{-x}$ to give a lower bound.

Consider two cases, first $m \leq n$

$\implies$ P(all balls in different bins) = $\frac{\binom{n}{m} m!}{n^m}$

$\implies$ P(at least one ball has more than one balls) = 1 - P(all balls in different bins)

$\implies$ P(at least one ball has more than one balls) = $1 - \frac{\binom{n}{m} m!}{n^m}$

lower bound for P(at least one ball has more than one balls) = $1 - \prod_i^{m-1} 1 - \frac{i}{n}$

P(at least one ball has more than one balls) $\geq 1 - \prod_i^{m-1} e^{-\frac{i}{n}}$

Lower Bound $= 1 - e^{-\frac{(m-1)m}{2n}}$

now, second case is $m > n$,
for this, we can use pigeon hole principle

$\implies$ P(at least one ball has more than one balls) = 1

(2) [Coupon Collecting] Let X denote the number of balls thrown until every bin has at least one ball. What is the expectation of X? Express it using the harmonic number $H_n = \sum_{i=1}^{n} 1/i$.

Consider n intervals, where k^{th} interval implies the frame where exactly $k-1$ different bins are not empty.

now, defining n_k as the total number of throws made in the k^{th} frame.

given, all throws are independent,
$\implies$ $n_1, n_2, n_3, ..., n_k, ...n_n$ are i.i.d. random variables.

Expected no of throws = $E_{total} = \sum_{k=1}^{n} E(n_k)$

now, P(throw each new ball in a new bin) = $\frac{n-k+1}{n}$, where k is the given time frame.
$\implies$ $p_k = \frac{n-k+1}{n}$ ... (1)

Using (1), we can establish that n_k is a random variable with geometric distribution, where probability is defined as p_k

$\implies E(n_k) = \frac{1}{p_k}$
$\implies E_{total} = \sum_{k=1}^{n} \frac{1}{p_k}$
$\implies E_{total} = \sum_{k=1}^{n} \frac{n}{n-k+1}$

$\implies E_{total} = \sum_{k=1}^{n} \frac{n}{k}$

$\implies \mathbf{E_{total} = nH_n}$

2 VC-dimension of Rectangles [25/3 pts]

What is the VC-dimension d of axis-parallel rectangles in R^3? Specifically, a legal target function is specified by three intervals $[x_{\min}, x_{\max}], [y_{\min}, y_{\max}], [z_{\min}, z_{\max}]$, and classifies an example (x, y, z) as positive if and only if $x \in [x_{\min}, x_{\max}], y \in [y_{\min}, y_{\max}]$, and $z \in [z_{\min}, z_{\max}]$. Justify your answer.

VC-dimension of axis-parallel rectangles in $\mathbf{R^3} = 6$

consider the points, $(1,0,0), (0,1,0), (-1,0,0), (0,-1,0), (0,0,1), (0,0,-1)$ .... (1)
this set of points can can be shattered

For understanding this, we can take the max and min points across the 1-dimensional spaces, say, across x-axis i.e. consider (1,0,0), (-1,0,0)

Now, these points are sufficient to label combinations of the type
{(class1, class1); (class2, class2); (class1, class2); (class2, class1)}, where class1 is (+) and class2 is (-).

Similarly, we can establish the same logic across y-axis and z-axis.

Thus, using (1), we can chose every labelling of the points in (1).

$\implies$ VC-dimension of axis-parallel rectangles in $R^3 \geq 6$

Consider 7 points as V, and we select the min set S which contains max-x, min-x, max-y, min-y, max-z, min-z, then we can have maximum 6 point and minimum 2 points.

Labeling points in S as class1 (+), and
Labeling points in V $\cap$ not in S as class2 (-).

If we take a axis parallel rectangle, which has all points in S, then it would also label the points not in S as class1 (+).

In other words, we can not have a labelling with 7 points where points inside the geometrical connection as class1 (+) and the remaining points as class2 (-).

3 Mistake Bound Model [25/3 pts]

CNF is the class of Conjunctive Normal Form formulas in the form $C_1 \wedge C_2 \wedge \ldots$, where each clause C_i is in the form $L_1 \vee L_2 \ldots$, and each Boolean literal L_i is either a boolean feature x or its negation $\neg x$. k-CNF is the class of CNF in which each clause has size at most k. For example, $x_4 \wedge (x_1 \vee x_2) \wedge (x_2 \vee \neg x_3 \vee x_5)$ is a 3-CNF. Give an algorithm to learn 3-CNF formulas over n boolean features in the mistake-bound model. Your algorithm should run in polynomial-time per example (so the "halving algorithm" is not allowed). How many mistakes does it make at most? (Hint: modify the FIND-S algorithm.)

For 3-CNF, we will have clauses with at most 3 boolean literals.
.... (1) now, we need an algorithm to learn 3-CNF formulas over n boolean features

using (1), number of clauses for n boolean features = $N = 2^1\binom{n}{1} + 2^2\binom{n}{2} + 2^3\binom{n}{3}$
$\implies$ N is a $poly(n)$

thus, total 3-CNF formulas $T = 2^N$
$\implies$ $T = 2^{poly(n)}$

using above, we can create new variable x_i for each of these clauses, and the corresponding reduction should take polynomial time.

now, applying FIND-S to obtain the learned hypothesis

Mistake 1:
Clauses with size $2^1\binom{n}{2}$ will reduce to $2^1\binom{n}{2} - \frac{1}{2}2^1\binom{n}{2}$,
Clauses with size $2^2\binom{n}{2}$ will reduce to $2^2\binom{n}{2} - \frac{1}{2^2}2^2\binom{n}{2}$, and
Clauses with size $2^3\binom{n}{3}$ will reduce to $2^3\binom{n}{3} - \frac{1}{2^3}2^3\binom{n}{3}$.
....
This pattern will repeat on each subsequent mistake, where at max 1 entry will be reduced.
Thus, the maximum number of mistakes = $max(\binom{n}{2}, 3\binom{n}{2}, 7\binom{n}{3}) + 1$
$= 7\binom{n}{3} + 1$

4 Extra Credit: VC-dimension of Linear Separators [20 pts]

In this problem, you will prove that the VC-dimension of the class H_n of halfspaces (another term for linear threshold functions $f_{w,b}(x) = \text{sign}(w^\top x + b)$) in n dimensions is $n + 1$. We will use the following definition: The convex hull of a set of points S is the set of all convex combinations of points in S; this is the set of all points that can be written as $\sum_{x_i \in S} \lambda_i x_i$, where each $\lambda_i \geq 0$, and $\sum_i \lambda_i = 1$. It is not hard to see that if a halfspace has all points from a set S on one side, then the entire convex hull of S must be on that side as well.
(a) [lower bound] Prove that VC-dim$(H_n) \geq n + 1$ by presenting a set of $n + 1$ points in n-dimension space such that one can partition that set with halfspaces in all possible ways, i.e., the set of points are shattered by H_n. (And, show how one can partition the set in any desired way.)
(b) [upper bound part 1] The following is Radon's Theorem, from 1920's.

Theorem 1. Let S be a set of $n + 2$ points in n dimensions. Then S can be partitioned into two (disjoint) subsets S_1 and S_2 whose convex hulls intersect.

Show that Radon's Theorem implies that the VC-dimension of halfspaces is at most $n + 1$. Conclude that VC-dim$(H_n) = n + 1$.
(c) [upper bound part 2] Now we prove Radon's Theorem. We will need the following standard fact from linear algebra. If $x_1, \ldots, x_{n+1}$ are $n + 1$ points in n-dimensional space, then they are linearly dependent. That is, there exist real values $\lambda_1, \ldots, \lambda_{n+1}$ not all zero such that $\lambda_1 x_1 + \ldots + \lambda_{n+1} x_{n+1} = 0$. You may now prove Radon's Theorem however you wish. However, as a suggested first step, prove the following. For any set of $n + 2$ points $x_1, \ldots, x_{n+2}$ in n-dimensional space, there exist $\lambda_1, \ldots, \lambda_{n+2}$ not all zero such that $\sum_i \lambda_i x_i = 0$ and $\sum_i \lambda_i = 0$. (This is called affine dependence.)

a) As we know, VC dimension is the size of largest set of points that can be shattered by the hypothesis.
Target function labels any example by the function: $sign(w > x + b)$, for halfspaces.

Let us consider a set of $n+1$ points,

$(0\ 0\ 0...0)^T, (1\ 0\ 0\ ...0)^T, (0\ 1\ 0\ ...0)^T, (0\ 0\ 1\ ...0)^T, (0\ 0\ 0\ ...1)^T$
Each of the points is in a n-dimensional space.
Point 0: is the origin
Point 1: is the point with coordinate in first dimension as 1, and remaining of the coordinates are 0.
Point 2: is the point with coordinate in second dimension as 1, and remaining of the coordinates are 0.
....
....
Point i: is the point with coordinate in i^{th} dimension as 1, and remaining of the coordinates are 0.
....
....
Point n: is the point with coordinate in n^{th} dimension as 1, and remaining of the coordinates are 0.

Now, in order to partition the above set of points into just 2 sets, say $S1$ (contains origin) and $S2$, we will need to use a Hyperplane.

So, for the lower bound, we just need to chose the Hyperplane.

Hence, the lower bound will be

$$\sum_{j:a_j \in S_2} x_j = \frac{1}{2}$$

So, VC dim (class of halfspaces in dim-n) $\geq$ n+1

b) Let us consider a Set S, which comprises of $n+2$ points.

Now, from Radons Theorem, we know that

Any set of $n+2$ points in R^n can always be partitioned in two subsets V_1 and V_2 such that the convex hulls of V_1 and V_2 intersect.

$\implies$ let us partition S in two subsets S_1 and S_2 s.t. the convex hulls of S_1 and S_2 intersect.

consider a point $a \in S1$ in the intersection
assume that a hyperplane exists

$$w^T x_j + c \leq w_0, \forall x_j \in S_1$$
$$w^T x_j + c > w_0, \forall x_j \in S_2$$

$\implies w^T a \leq w_o$ ..... (1)

We also have

$w^T a = \sum_{j:x_j \in S_2} \lambda_j w^T x_i$
$\implies \sum_{j:x_j \in S_2} \lambda_j w^T x_i > (\sum_{j:x_j \in S_2} \lambda_j)\ min_{j:x_j \in S_2}(w^T x_j)$

also,

$(\sum_{j:x_j \in S_2} \lambda_j)\ min_{j:x_j \in S_2}(w^T x_j) = min_{j:x_j \in S_2}\ (w^T x_j)$

$min_{j:x_j \in S_2} (w^T x_j) > w_0$ ..... (2)

Now, (1) and (2) contradicts each other

Set S of $n+2$ points can NOT be shattered
VC-dimension of halfspaces is at most $n+1$

$$\text{VC-dim}(H_n) = n + 1$$

c) Let $y_i =< x_i, 1 >$, where $i \in (1, 2..n+2)$
$\implies$ y_i is linearly dependent

Using the fact that $\{y_i\}$ is linearly dependent, we can say that

$\exists \lambda_1, \lambda_2, ...\lambda_{n+2}$

s.t. $\sum_{i=1}^{n} \lambda_i y_i = 0$

where λ_i are scalars and not all of them are zero.

now, let $V_1 = \{x_i | \lambda_i \geq 0\}$ and $V_2 = \{x_i | \lambda_i < 0\}$

So, we can define the following

$\implies$ $\lambda_{total} = \sum_{j:x_j \in V_1} \lambda_j$ ..... (1)

As we have defined, $V_2 = \{x_i | \lambda_i < 0\}$,
$\implies$ $\lambda_{total} = -\sum_{j:x_j \in V_2} |\lambda_j|$ ..... (2)

similarly,

$\implies$ $x_{total} = \sum_{j:x_j \in V_1} \lambda_j x_j$ ..... (3)
$\implies$ $x_{total} = -\sum_{j:x_j \in V_2} |\lambda_j| x_j$ ..... (4)

using (1) and (3)

$\implies$ $x_{total} = \sum_{j:x_j \in V_1} \frac{\lambda_j . \lambda_{total}}{\lambda_{total}} x_j$
$\implies$ $\frac{x_{total}}{\lambda_{total}} = \sum_{j:x_j \in V_1} \frac{\lambda_j}{\lambda_{total}} x_j$ ... (5)

similarly, using (2) and (4)

$\implies$ $x_{total} = -\sum_{j:x_j \in V_2} \frac{|\lambda_j| . \lambda_{total}}{\lambda_{total}} x_j$
$\implies$ $\frac{x_{total}}{\lambda_{total}} = -\sum_{j:x_j \in V_2} \frac{|\lambda_j|}{\lambda_{total}} x_j$ ... (6)

Using (5) and (6), the point $\frac{x_{total}}{\lambda_{total}}$ lies in the convex hull of both V_1 and V_2

5 Part 2 - Compare Naive Bayes and TAN

P value and statistical significance:
The two-tailed P value is less than 0.0001
By conventional criteria, this difference is considered to be extremely statistically significant.

Mean of Naive = 0.8817049
Standard Deviation of Naive = 0.023662628577804

Mean of TAN = 0.9354762
Standard Deviation of TAN = 0.011968345005797

Difference of means = 0.0537713

Confidence interval:
The mean of Naive minus TAN equals 0.05377130 95% confidence interval of this difference: From 0.07138853 to 0.03615407

Intermediate values used in calculations:
t-statistic = 6.4124
df = 18
standard error of difference = 0.008

Naive Bayes	TAN
0.8956	0.9225
0.8246	0.93225
0.888426	0.942
0.8821	0.9262
0.90213	0.95012
0.86178	0.95775
0.887212	0.942
0.9065	0.925302
0.87858	0.93169
0.890121	0.92495