Problem Set 3

CS 6375

Due: 4/6/2022 by 11:59pm

Note: all answers should be accompanied by explanations for full credit. Late homeworks will not be accepted.

Problem 1: VC Dimension (25 pts)

1. Consider a binary classification problem for data points in $\mathbb{R}^3$ with a hypothesis space consisting of axis aligned 3-d boxes such that any point in the box is labeled with a $+$ and any point outside the box is labeled with a $-$. What is the VC dimension of this hypothesis space? Prove it. Can you generalize your argument to axis aligned boxes in $\mathbb{R}^d$?

Problem 2: Spherical Hypotheses (25 pts)

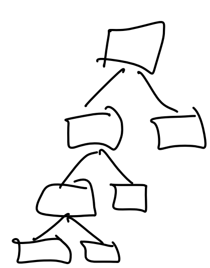
Given training data of the form $(x^{(1)}, y^{(1)}), \ldots, (x^{(M)}, y^{(M)})$, where $x^{(m)} \in \mathbb{R}^n$ and $y^{(m)} \in \{-1, 1\}$, consider the hypothesis space of n-dimensional spheres: each element of the hypothesis space is parameterized by a center $c \in \mathbb{R}^n$ and a radius $r > 0$ such that all points within distance r of the center c are classified as $+1$ and the remaining points are classified with a -1.

1. Assuming that the training data can be correctly classified under the sphereical hypothesis space, describe an optimization problem whose solution is a spherical hypothesis that is a max-margin perfect classifier.
2. Using the method of Lagrange multipliers, construct the dual of your optimization problem.

Problem 3: Medical Diagnostics (50 pts)

For this problem, you will use the data set provided with this problem set. The data has been divided into two pieces heart_train.data and heart_test.data. These data sets were generated using the UCI SPECT heart data set (follow the link for information about the format of the data). Note that the class label is the first column in the data set.

1. Suppose that the hypothesis space consists of all decision trees with exactly three attribute splits (repetition along the same path is allowed) for this data set.
   (a) Run the adaBoost algorithm for five rounds to train a classifier for this data set. Draw the 5 selected trees in the order that they occur and report the ϵ and α, generated by adaBoost, for each.



(b) Run the adaBoost algorithm for 10 rounds of boosting. Plot the accuracy on the training and test sets versus iteration number.

2. Now, suppose that the hypothesis space consists of only height 1 decision trees for this data set.

   (a) Use coordinate descent to minimize the exponential loss function for this hypothesis space over the training set. You can use any initialization and iteration order that you would like other than the one selected by adaBoost. What is the optimal value of α that you arrived at? What is the corresponding value of the exponential loss on the training set?

   (b) What is the accuracy of the resulting classifier on the test data?

   (c) What is the accuracy of adaBoost after 20 rounds for this hypothesis space on the test data? How does the α learned by adaBoost compare to the one learned by gradient descent?

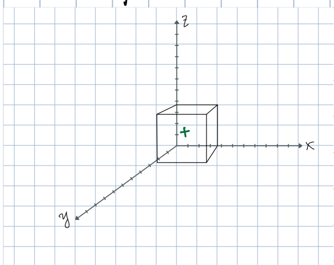
   (d) Use bagging, with 20 bootstrap samples, to produce an average classifier for this data set. How does it compare to the previous classifiers in terms of accuracy on the test set?

   (e) Which of these 3 methods should be preferred for this data set and why?

Problem 1: VC Dimension (25 pts)

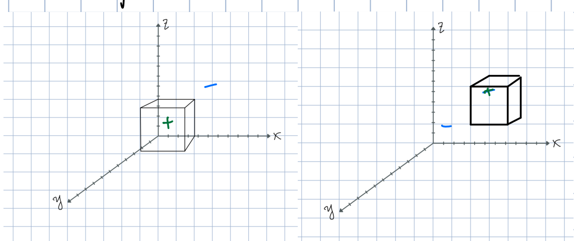
1. Consider a binary classification problem for data points in $\mathbb{R}^3$ with a hypothesis space consisting of axis aligned 3-d boxes such that any point in the box is labeled with a + and any point outside the box is labeled with a −. What is the VC dimension of this hypothesis space? Prove it. Can you generalize your argument to axis aligned boxes in $\mathbb{R}^d$?

① 1 point



✓

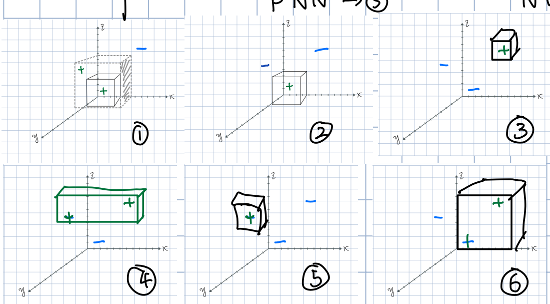
② 2 points



✓

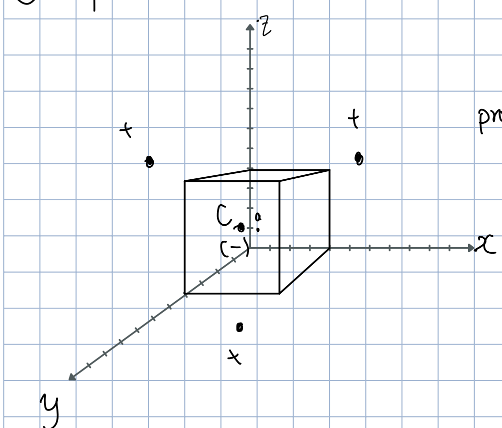
③ 3 points

PPP − All in big box
PPN → ①
PNP → ④
PNN → ⑤
NPP → ⑥
NPN → ②
NNP → ③
NNN → box NOT cover points



VC dimension = 3

④ 4 points ✗ (Does not work)



when points ≥ 4
There must be a point whose distances to other points are the shortest. Let it be the point C

prove: If point c is negatively labeled, while other points are positively labeled, It is impossible to fit all positive points in the same box.

Assume there exist box A centred at $a(x, y, z)$ covering all positive points, the distance from the farthest + point to a $= \sqrt{(x_a - x_{fur})^2 + (y_a - y_{fur})^2 + (z_a - z_{far})^2}$ from any positive points

Since c from any other positive points have the shortest distance

$\Rightarrow$ Therefore location of c must be inside the box.

c: Not possible

$\because$ Distance to other points not shortest

$\Rightarrow 0 \le \underset{\text{any point's}}{\text{Distance to c}} \le \underset{\text{furthest + point's}}{\text{Distance to a}}$

But c's label is negative

$\therefore$ Such a box A that contains only + label points doesn't exist in more than 3 points

$\therefore$ VC Dimension = 3

In n-dimensional box, Once there exists 4 or more points

$\Rightarrow$ a point "c" whose distance to other points is shortest will exist

$\Rightarrow$ If this point "c" is labeled negative while the other points are positive, then point c will be forced to be in the n-D box

$\Rightarrow$ The box that puts only + points inside cannot exist.

Problem 2: Spherical Hypotheses (25 pts)

Given training data of the form $(x^{(1)}, y^{(1)}), \ldots, (x^{(M)}, y^{(M)})$, where $x^{(m)} \in \mathbb{R}^n$ and $y^{(m)} \in \{-1, 1\}$, consider the hypothesis space of n-dimensional spheres: each element of the hypothesis space is parameterized by a center $c \in \mathbb{R}^n$ and a radius $r > 0$ such that all points within distance r of the center c are classified as $+1$ and the remaining points are classified with a -1.

1. Assuming that the training data can be correctly classified under the sphereical hypothesis space, describe an optimization problem whose solution is a spherical hypothesis that is a max-margin perfect classifier.
2. Using the method of Lagrange multipliers, construct the dual of your optimization problem.

⊛ $\lim_{\to\infty}$ not just $\frac{d}{dx} = 0$

∴ nm answer

$r^2 - 2 r r_0 \cos(\varphi - \gamma) + r_0^2 = a^2$ $\quad r(\varphi) =$

max r_1 => min $r_2 - r_1$

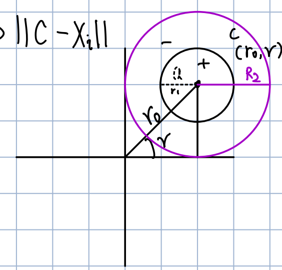
C => point whose distance to other points $\sqrt{(x_c - x_i)^2 + (y_c - y_i)^2} \Rightarrow \|C - X_i\|$
are smallest

∴ R_2 circle covers all points

R_1 circle covers only the + points

R_2 : Distance from c to furthest point

R_1 : Distance from c to furthest + point



$\arg\max_{r_1 r_2} \sqrt{(x_c - x_i)^2 + (y_c - y_i)^2}$