10-601: Homework 5

Due: 25 October 2014 11:59pm (Autolab)
TAs: Abhinav Maurya, Ying Yang

Name: Dawei Wang
Andrew ID: daweiwan

Please answer to the point, and do not spend time/space giving irrelevant details. You should not require more space than is provided for each question. If you do, please think whether you can make your argument more pithy, an exercise that can often lead to more insight into the problem. Please state any additional assumptions you make while answering the questions. You need to submit a single PDF file on autolab. Please make sure you write legibly for grading.

You can work in groups. However, no written notes can be shared, or taken during group discussions. You may ask clarifying questions on Piazza. However, under no circumstances should you reveal any part of the answer publicly on Piazza or any other public website. The intention of this policy is to facilitate learning, not circumvent it. Any incidents of plagiarism will be handled in accordance with CMU's Policy on Academic Integrity.

⋆: Code of Conduct Declaration

- Did you receive any help whatsoever from anyone in solving this assignment? **No**.
- If you answered *yes*, give full details: ____________________ (e.g. *Jane explained to me what is asked in Question 3.4*)
- Did you give any help whatsoever to anyone in solving this assignment? **No**.
- If you answered *yes*, give full details: ____________________ (e.g. *I pointed Joe to section 2.3 to help him with Question 2*).

1: True or False. (TA:- Abhinav Maurya)

State whether true (with a brief reason) or false (with a contradictory example). Credit will be granted only if your reasons/counterexamples are correct.

(a) If the VC Dimension of a set of classification hypotheses is ∞, then the set of classifiers can achieve 100% training accuracy on any dataset.

☑ True □ False

[*2 points*]

True. That is by definition. Though note here only has to be one classifier that can achieve 100% training accuracy, in compliance with the definition of **shattered**, where it says **there** *exists* **some consistent hypothesis in** H.

(b) There is no actual set of classification hypotheses useful in practical machine learning that has VC Dimension ∞.

□ True ☑ False

[2 points]

False. A 1-Nearest Neighbor (1-NN) would be a counter-example.
Also a support vector machine with Gaussian kernel.

(c) VC Dimension of the set of all decision trees (defined on a given set of real-valued features) has is finite.

☑ True □ False

[2 points]

False. Since a single feature can be tested for multiple times, no matter how many instances we have we can always build a decision adequately complex to contain all possible scenarios.

(d) Since the true risk is bounded by the empirical risk, it is a good idea to minimize the training error as much as possible.

□ True ☑ False

[2 points]

False. It is a good idea to achieve a lower training error on *more* instances. A minimal training error merely moves the center of the confidence interval to zero, but its width could be large, leaving the bounding quite pointless.

(e) PAC learning bounds help you estimate the number of samples needed to reduce the discrepancy between the true risk and empirical risk to an arbitrary constant with high probability.

☑ True □ False

[2 points]

True. This is basically what it does as defined.

(f) If the VC Dimension of a set of classification hypotheses is K, then no algorithm can have a mistake bound that is strictly less than K.

□ True ☑ False

[2 points]

True. K is equal or greater than the optimal mistake bound.

(g) SVMs with a gaussian kernel have VC Dimension equal to $n + 1$ where n is the number of support vectors.

□ True ☑ False

[*2 points*]

False. The VC Dimension is ∞. A Gaussian kernel effectively maps the feature space to a Hilbert space with infinite dimensions. A linear kernel would instead have VC Dimension $n + 1$.

(h) As the degree of the polynomial kernel increases, the VC Dimension of the set of classification hypotheses increases.

☑ True □ False

[*2 points*]

True. The dimension of the Hilbert space also increases, so does its VC Dimension.

(i) VC Dimensions of the sets of classification hypotheses induced by logistic regression and linear SVM (learnt on the same set of features) are different.

□ True ☑ False

[*2 points*]

False. Both classifiers are essentially linear, therefore having the same VC Dimension.

(j) VC Dimension depends on the dataset we use for shattering.

□ True ☑ False

[*2 points*]

False. VC Dimension is defined over an instance space, and has nothing to do with any particular dataset.

2: PAC learning for conjunctions of boolean literals. (TA:- Ying Yang)

Consider a function that takes n-bit binary inputs ($\boldsymbol{x} = (x_1, x_2, \cdots, x_n)$, $x_i \in \{0, 1\}$) and output binary responses. This function is a list of conjunctions of boolean literals, and the list can include x_i or $\neg x_i$, or neither of them, but not both of them, for $i = 1, \cdots, n$. One example of such a function is

$$h(\boldsymbol{x}) = x_1 \wedge \neg x_2 \wedge x_3.$$

We are given data that can be perfectly explained by at least one such function. Suppose we are also given an algorithm that learns a function, which has zero training error on finite data samples, how many training examples do we need to guarantee, with probability at least 95%, that the true error rate of our learned function is $< 5\%$? Use $n = 10$ in your computation.

[*5 points*]

For each bit there can be no corresponding boolean literal, or one with or without logical not. So there are three possibilities in total. The size of the hypotheses space is therefore $|H| = 3^n$.

The minimal number of training examples is given by

$$m \geq \frac{1}{5\%}\left[\ln|H| + \ln\frac{1}{5\%}\right] = 20\left[10\ln 3 + \ln 20\right] = 280 \qquad (1)$$

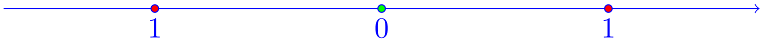
3: VC-dimensions of binary classifiers. (TA:- Ying Yang)

Write the VC-dimensions of the following families of binary classifiers. Explain your results with examples that can or can not be shattered.

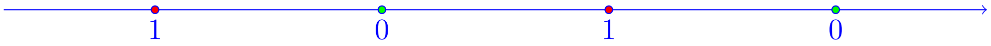
1. $f : \mathbb{R} \to \{0,1\}, f(x) = \begin{cases} 1 \text{ if } x \in [a,b] \\ 0 \text{ otherwise} \end{cases}$ where a and b are any real constants, such that $a < b$.
2. The functions in 1 and the functions that flips the outputs of the functions in 1.
3. $f : \mathbb{R}^2 \to \{0,1\}, f(\boldsymbol{x}) = \begin{cases} 1 \text{ if } ||\boldsymbol{x}||^2 < c \\ 0 \text{ otherwise} \end{cases}$ where $\boldsymbol{x} \in \mathbb{R}^2, c > 0 \in \mathbb{R}$

[*9 points*]

1. VC Dimension = 2. For arbitrary training set with two instances, just pick the minimum neighbor that contains all the points labeled with 1. Here is an example with 3 instances that can not be shattered.



2. VC Dimension = 3. The example above apperently shattered, however, for 4 instances, here is an example that cannot be shattered:



3. VC Dimension = 1. We need to show that there's at least one set of size d, that H can shatter. c.f., Page 10, `http://cs229.stanford.edu/notes/cs229-notes4.pdf`

4: Learning theory of SVMs with quadratic Kernels. (TA:- Ying Yang)

Given a family of support vector machines with a quadratic kernel $k(\boldsymbol{x_1}, \boldsymbol{x_2}) = (\boldsymbol{x_1}^T \boldsymbol{x_2})^2$. The inputs $\boldsymbol{x} \in \mathbb{R}^n$, and the output is binary.

1. What is the VC-dimension of this family?

   [*4 points*]

   This is a homogeneous polynomial kernel of degree 2. So the constant terms, and the linear terms would disappear (deriving from the inhomogeneous case in the lecture), and the quadratic terms and the pairwise terms would remain, which gives a Hilbert space of dimension $n + 0.5n(n-1) = 0.5n(n+1)$, and the VC-dimension would be $0.5n(n+1) + 1$.

2. Now we are given data that can be perfectly classified by one SVM in this family. If we are trying to train an SVM in this family, how many training examples do we need to guarantee that the true error rate of the trained SVM is $< 5\%$ with probability at least 95% ? Use $n = 10$ in your computation.

   [*2 points*]

   The number of examples is given by:

$$m \geq \frac{1}{5\%}\left[4\log_2\frac{2}{5\%} + 8VC(H)\log_2\frac{13}{5\%}\right] = 72307 \tag{2}$$

$$\boxed{\textbf{Total: 40}}$$