EE660 Week5 HW 1139093874 Pin-Hsuan Lee

1. Consider the email spam classification problem of Murphy Problem 8.1. Suppose you intend to use a linear perceptron classifier on that data. In the parts below, unless stated otherwise, assume the dataset of $N = 4601$ samples is split into $N_{Tr} = 3000$ for training and $N_{Test} = 1601$ for testing. Also, for the tolerance δ in the VC generalization bound, use 0.1 (for a certainty of 0.9). The parts below have short answers.

Hint: You may use the relation that if $\mathcal{H}$ is a linear perceptron classifier in D dimensions (D features), $d_{VC}(\mathcal{H}) = D+1$. (This will be proved in Problem 2.)

a) What is the VC dimension of the hypothesis set?

b) Expressing the upper bound on the out-of-sample error as $E_{out}(h_g) \le E_{in}(h_g) + \varepsilon_{\text{vc}}$

For $E_{in}(h_g)$ measured on the training data, use d_{vc} from part (a) to get a value for ε_{vc} .

c) To get a lower ε_{vc}, suppose you reduce the number of features to $D = 10$, and also increase the training set size to 10,000. Now what is ε_{vc}?

d) Suppose that you had control over the number of training samples N_{Tr} (by collecting more email data). How many training samples would ensure a generalization error of $\varepsilon_{\text{vc}} = 0.1$ again with probability 0.9 (the same tolerance $\delta = 0.1$), and using the reduced feature set (10 features)?

e) Instead suppose you use the test set to measure $E_{in}(h_g)$, so let's call it $E_{test}(h_g)$. What is the hypothesis set now? What is its cardinality?

f) Continuing from part (e), use the bound:
$E_{out}(h_g) \le E_{test}(h_g) + \varepsilon$
Use the original feature set and the original test set, so that $N_{Test} = 1601$. Give an appropriate expression for ε and calculate it numerically.

1. $N = 4601 \begin{cases} N_{Tr} = 3000 \\ N_{Test} = 1601 \end{cases}$, $\delta = 0.1$

(a) Since $D = 57$ (57 features), $d_{VC}(\mathcal{H}) = D+1 = \underline{58}$

(b) $\varepsilon_{VC} = \sqrt{\frac{8}{N} \ln \frac{4[(2N)^{d_{VC}}+1]}{\delta}} = \sqrt{\frac{8}{3000} \ln \frac{4[(2\times 3000)^{58}+1]}{0.1}} \approx \sqrt{\frac{8}{3000}\left[\ln(4(2N)^{d_{VC}}) - \ln 0.1\right]}$

$\approx \sqrt{\frac{8}{3000}\left[\ln(4\times(2\times 3000)^{58}) - \ln 0.1\right]} = \sqrt{\frac{8}{3000}\left(\ln 4 + 58\ln 6000 - \ln 0.1\right)} \doteq 1.1642$

$E_{out}(h_g) \le E_{in}(h_g) + 1.1642$

(c) $D = 10 \Rightarrow d_{VC} = 10+1 = 11$, $N = 10000$

$\varepsilon_{VC} \approx \sqrt{\frac{8}{10000}\left[\ln 4 + 11\ln 20000 - \ln 0.1\right]} \doteq 0.3$

(d)

$\varepsilon_{VC} = 0.1 = \sqrt{\frac{8}{N}[\ln 4 + 11\ln(2N) - \ln 0.1]} \Rightarrow \frac{0.01}{8}N = \ln 4 + 11\ln(2N) - \ln 0.1$

$\Rightarrow 1.25\times10^{-3}N = 11\ln(2N) + 3.69$

$\Rightarrow N = 8800\ln(2N) + 2952$

$\Rightarrow e^{N-2952} = 2N^{8800} \Rightarrow N = 117307.23$

$\Rightarrow$ We need to have 117308 samples.

(e) (f)

$E_{out}(hg) \leq E_{test}(hg) + \underbrace{\varepsilon}_{\sqrt{\frac{1}{2N}\ln\frac{2M}{\delta}}}$

$\xrightarrow[N=1601]{M=1} E_{out}(hg) \leq E_{test}(hg) + \sqrt{\frac{1}{2\times1601}\ln\frac{2}{0.1}}$

$\Rightarrow E_{out}(hg) \leq E_{test}(hg) + 0.031$

M = cardinality = 1

hypothesis set: hg

2. AML Exercise 2.4 (page 52). In addition to the hints given in the book, you can solve the problem by following the steps outlined below (on the next page).

Exercise 2.4

Consider the input space $\mathcal{X} = \{1\} \times \mathbb{R}^d$ (including the constant coordinate $x_0 = 1$). Show that the VC dimension of the perceptron (with $d+1$ parameters, counting w_0) is exactly $d+1$ by showing that it is at least $d+1$ and at most $d+1$, as follows.

(a) To show that $d_{\text{VC}} \geq d+1$, find $d+1$ points in $\mathcal{X}$ that the perceptron can shatter. *[Hint: Construct a nonsingular $(d+1) \times (d+1)$ matrix whose rows represent the $d+1$ points, then use the nonsingularity to argue that the perceptron can shatter these points.]*

(b) To show that $d_{\text{VC}} \leq d+1$, show that no set of $d+2$ points in $\mathcal{X}$ can be shattered by the perceptron. *[Hint: Represent each point in $\mathcal{X}$ as a vector of length $d+1$, then use the fact that any $d+2$ vectors of length $d+1$ have to be linearly dependent. This means that some vector is a linear combination of all the other vectors. Now, if you choose the class of these other vectors carefully, then the classification of the dependent vector will be dictated. Conclude that there is some dichotomy that cannot be implemented, and therefore that for $N \geq d+2$, $m_{\mathcal{H}}(N) < 2^N$.]*

For part (a):

i. Write a point $\underline{x}_i$ as a $d+1$ dimensional vector;

ii. Construct the $(d+1)\times(d+1)$ matrix suggested by the book;

iii. Write $\underline{h}(\underline{\underline{X}})$, the output of the perceptron, as function of $\underline{\underline{X}}$ and the weights $\underline{w}$ (note that $\underline{h}(\underline{\underline{X}})$ is a $d+1$ dimensional vector with elements +1 and -1);

iv. Using the nonsingularity of $\underline{\underline{X}}$, justify how any $\underline{h}(\underline{\underline{X}})$ can be obtained.

For part (b):

i. Write a point $\underline{x}_k$ as a linear combination of the other $d+1$ points;

ii. Write $h(\underline{x}_k)$ (output for the chosen point) and substitute the value of $\underline{x}_k$ by the expression just found on the previous item (**Hint**: use the $\text{sgn}\{\bullet\}$ function);

iii. What part of your expression in (ii) determines the class assignment of each point $\underline{x}_i$, for $i \neq k$?

iv. You have just proven (part (a)) that $\underline{h}(\underline{\underline{X}})$ with $\underline{\underline{X}}_{(d+1)\times(d+1)}$ can be shattered. When we add a $(d+2)^{\text{th}}$ line to $\underline{\underline{X}}$ can it still be shattered? In other words, can you choose the value of $h(\underline{x}_k)$? Justify your answer. **Hint:** you can choose the class label of the other $(d+1)$ points.

(a)

$\underline{x}_i = \begin{bmatrix} x_0 \\ x_{i1} \\ \vdots \\ x_{id} \end{bmatrix}$ $\quad \underline{x}_i \in \mathbb{R}^{d+1}$

$\underline{\underline{X}} = \begin{bmatrix} -\underline{x}_1^T- \\ -\underline{x}_2^T- \\ \vdots \\ -\underline{x}_{d+1}^T- \end{bmatrix}_{(d+1)\times(d+1)} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 1 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ 1 & 0 & 0 & \cdots & 1 \end{bmatrix}_{(d+1)\times(d+1)} \Rightarrow$ $\underline{\underline{X}}$ is invertible.

$\underline{h}(\underline{\underline{X}}) = \text{sign}(\underline{\underline{X}}\,\underline{w}) = \begin{bmatrix} h_1(\underline{\underline{X}}) \\ h_2(\underline{\underline{X}}) \\ \vdots \\ h_{d+1}(\underline{\underline{X}}) \end{bmatrix}_{(d+1)\times 1} \Rightarrow (\underline{\underline{X}}\,\underline{w}) = \underline{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{d+1} \end{bmatrix}_{(d+1)\times 1}$

$\Rightarrow \underline{w} = \underline{\underline{X}}^{-1}\underline{y}$

$\Rightarrow \underline{\underline{X}}$ can be shattered

$\Rightarrow d_{VC} \geq d+1$

(b) $\underline{\underline{X}} = \begin{bmatrix} -\underline{x}_1^T- \\ -\underline{x}_2^T- \\ \\ -\underline{x}_{d+2}^T- \end{bmatrix}_{(d+2)\times(d+1)}$

→ more rows than columns

→ It must contain a dependent row

→ Assume $x_{d+2} = a_1x_1 + a_2x_2 + \cdots + a_{d+1}x_{d+1}$ (some a_i non-zero)

$\underline{h}(\underline{x}_{d+2}) = \text{sign}(\underline{w}^T\underline{x}_{d+2}) \Rightarrow \underline{w}^T\underline{x}_{d+2} = a_1\underline{w}^T\underline{x}_1 + a_2\underline{w}^T\underline{x}_2 + \cdots + a_{d+1}\underline{w}^T\underline{x}_{d+1}$

Assume the sign of a_i is the same with $(\underline{w}^T\underline{x}_i)$'s, which means $\underline{w}^T\underline{x}_{d+2} > 0$, and it's impossible to make $\underline{w}^T\underline{x}_{d+2} < 0$.

$\Rightarrow \underline{x}_{d+2}$ cannot be shattered

$\Rightarrow \underline{\underline{X}}$ cannot be shattered if it contains $d+2$ points

$\Rightarrow d_{VC} \leq d+1$. Q.E.D.

For $N \geq d+2$, since it cannot be shattered, $m_H(N) < 2^N$.

3. AML Problem 2.13 (a), (b).

Problem 2.13

(a) Let $\mathcal{H} = \{h_1, h_2, \ldots, h_M\}$ with some finite M. Prove that $d_{\text{VC}}(\mathcal{H}) \leq \log_2 M$.

(b) For hypothesis sets $\mathcal{H}_1$, $\mathcal{H}_2$, $\cdots$, $\mathcal{H}_K$ with finite VC dimensions $d_{\text{VC}}(\mathcal{H}_k)$, derive and prove the tightest upper and lower bound that you can get on $d_{\text{VC}}\left(\cap_{k=1}^{K} \mathcal{H}_k\right)$.

(a) By definition, $m_{\mathcal{H}}(d_{vc}) = 2^d$. $(d_{vc}(\mathcal{H}) = d)$

$$m_{\mathcal{H}}(d) = 2^d = \max_{x_1, x_2, \cdots, x_d \in \mathcal{X}} \left|\mathcal{H}(x_1, x_2, \cdots, x_d)\right|$$
$$= \max_{x_1, x_2, \cdots, x_d \in \mathcal{X}} \left|\{h(x_1), h(x_2), \cdots, h(x_d) : h \in \mathcal{H}\}\right|$$
$$\leq |\mathcal{H}| = M$$

$\Rightarrow 2^d \leq M$

$\Rightarrow d_{vc}(\mathcal{H}) \leq \log_2 M$ Q.E.D

(b) ① In the worst situation, $\cap_{k=1}^{K} \mathcal{H}_k = \{h\} \Rightarrow M = 1 \Rightarrow d_{vc}(\mathcal{H}) = 0$

Therefore, $d_{vc}(\cap_{k=1}^{K} \mathcal{H}_k) \geq 0$

② Assume $d_{vc}(\cap_{k=1}^{K} \mathcal{H}_k) > \min_{1 \leq k \leq K} d_{vc}(\mathcal{H}_k) = d$, which means $\cap_{k=1}^{K} \mathcal{H}$ can shatter $d+1$ points.

Let $x_1, x_2, \cdots, x_{d+1}$ be these points, we can write

$$\cap_{k=1}^{K} \mathcal{H}_k(x_1, x_2, \cdots, x_{d+1}) = \{h(x_1), h(x_2), \cdots, h(x_{d+1}) : h \in \cap_{k=1}^{K} \mathcal{H}_k\}$$
$$\subset \{h(x_1), h(x_2), \cdots, h(x_{d+1}) : h \in \mathcal{H}_k\} = \mathcal{H}_k(x_1, x_2, \cdots, x_{d+1})$$
for all $k = 1, 2, \cdots K$

$\Rightarrow$ it gives us the inequality:

$$2^{d+1} \leq |\{h(x_1), \cdots, h(x_{d+1}) : h \in \mathcal{H}_k\}| \leq 2^{d+1} \Rightarrow |\{h(x_1), h(x_2), \cdots, h(x_{d+1}) : h \in \mathcal{H}_k\}| = 2^{d+1}$$

$\Rightarrow$ Any $\mathcal{H}_k$ can shatter $d+1$ points.

let $\min_{1 \leq k \leq K} d_{vc}(\mathcal{H}_k) = d_{vc}(\mathcal{H}_m) = d \Rightarrow d = d_{vc}(\mathcal{H}_m) \geq d+1 \Rightarrow$ contradiction !!

from ① & ②

$\Rightarrow 0 \leq d_{vc}(\cap_{k=1}^{K} \mathcal{H}_k) \leq \min_{1 \leq k \leq K} d_{vc}(\mathcal{H}_k)$

4. AML Exercise 4.3 (p. 125). part (a) - first question only; part (b) - first question only. [Think about the second question of each part if you like; there isn't necessarily a single answer to each, though.]

Exercise 4.3

Deterministic noise depends on $\mathcal{H}$, as some models approximate f better than others.

(a) Assume $\mathcal{H}$ is fixed and we increase the complexity of f. Will deterministic noise in general go up or down? Is there a higher or lower tendency to overfit?

(b) Assume f is fixed and we decrease the complexity of $\mathcal{H}$. Will deterministic noise in general go up or down? Is there a higher or lower tendency to overfit? *[Hint: There is a race between two factors that affect overfitting in opposite ways, but one wins.]*

(a) $\mathcal{H}$ is fixed and we increase the complexity of f.
More complex model is more susceptible to noise than the simpler one.
As the complexity of f increases, the deterministic noise will go up in general.

(b) f is fixed and we decrease the complexity of $\mathcal{H}$.
Since deterministic noise is the difference between $\mathcal{H}$ and f.
When the order of $\mathcal{H}$ is high, it can fit f better such that the deterministic noise is lower. Therefore, if we decrease the complexity of $\mathcal{H}$, deterministic noise will go up in general.