Introduction to Machine Learning via Statistical Learning Theory

Problem set 1: Theoretical Problems

Problem 1: VC Dimension Guarantee on Learning

We saw in class that for a hypothesis class $\mathcal{H}$ with VC-dimension D, for any source distribution $\mathcal{D}$, with probability at least $1-\delta$ over the choice $S \sim \mathcal{D}^m$ of a random training set of size $m > D$, for all $h \in \mathcal{H}$:

$$\left|R(h) - \hat{R}_S(h)\right| \leq 2\sqrt{2\frac{D\log\frac{2em}{D} + \log\frac{2}{\delta}}{m}} \tag{1}$$

Consider the ERM learning rule for the hypothesis class $\mathcal{H}$:

$$\text{ERM}_{\mathcal{H}}(S) = \hat{h}_S = \arg\min_{h\in\mathcal{H}} \hat{R}_S(h)$$

(a) Prove that for any source distribution $\mathcal{D}$, with probability at least $1-\delta$ over the choice $S \sim \mathcal{D}^m$ of a random training set of size $m > D$:

$$R(\hat{h}_S) \leq \inf_{h\in\mathcal{H}} \left(R(h) + 4\sqrt{2\frac{D\log\frac{2em}{D} + \log\frac{2}{\delta}}{m}} \right) \tag{2}$$

(b) Assume there exists some good hypothesis $h^* \in \mathcal{H}$ with low $R(h^*) = R^*$, but of course we do not know which hypothesis is the good hypothesis. Obtain an upper bound on the "sample complexity" required for the ERM learning rule to return a hypothesis with error $R^* + \epsilon$, i.e. a hypothesis almost as good as the one we know exists. That is, find an expression for $m_0(D, \epsilon)$, such that for any source distribution $\mathcal{D}$ for which there exists $h^* \in \mathcal{H}$ with $R(h^*) = R^*$, for any $\epsilon > 0$, with probability at least $1-\delta$ over a training set of size $m > m_0(D, \epsilon)$:

$$R(\hat{h}_S) \leq R^* + \epsilon \tag{3}$$

Optional Problem $1\frac{1}{2}$: A More Optimistic Bound

(Will be added later)

Problem 2: Can't Beat the VC Dimension

We saw that for the cardinality based bound, the dependence of the sample complexity on the cardinality was tight for some classes, but for other classes, which included many very similar hypothesis, the bound did not capture the true complexity of the hypothesis class. We will now show that, up to constants and a log factors, the VC dimension does always capture the true complexity of a hypothesis class. That is: for any class with VC dimension D, we saw above that having $\Omega(D \log D)$ samples is enough for learning. We will see now that learning is *not* possible with $o(D)$ samples.

(a) First consider the generalization ability of the ERM learning rule. Show that for *any* hypothesis class $\mathcal{H}$ with VC dimension D, there exists a source distribution $\mathcal{D}$, such that for any sample S of $m < \frac{D}{2}$ training example (i.e. with probability one over $S \sim \mathcal{D}^m$), the minimum empirical error is zero, i.e. $\hat{R}_S(\hat{h}) = 0$, but the generalization error of the ERM is half, i.e. $R(\hat{h}) = 1/2$. We can conclude that if we have less then $D/2$ samples, even if we get zero training error, we cannot have any meaningful guarantee on the generalization error.

Note that, as you might have encountered here, there might actually be many different hypothesis minimizing the empirical error $\hat{R}_S(h)$. Since the ERM learning rule does not give us any guidance for choosing between these different learning rules, to be able to say that the learning rule is "good", it must be that *any* hypothesis minimizing $\hat{R}_S(h)$ is good. Conversely, to show that the learning rule is bad, it is enough to show that there exists *some* hypothesis h minimizing $\hat{R}_S(h)$ that is not good.

(b) Show that for *any* hypothesis class $\mathcal{H}$ with VC dimension D, there exists a source distribution $\mathcal{D}$, such that there is some $h \in \mathcal{H}$ with $R(h) = 0$, but such that for any sample S of $m < (1-\epsilon)D$ training example, there exists a hypothesis $\hat{h}$ minimizing the empirical error:

$$\hat{R}_S(\hat{h}) = \min_{h \in \mathcal{H}} \hat{R}_S(h) \tag{4}$$

but such that

$$R(\hat{h}) \geq \epsilon = \epsilon + \min_{h \in \mathcal{H}} R(h). \tag{5}$$

We can conclude that even if there is a good hypothesis in $\mathcal{H}$, with less than D samples, the ERM learning rule might not find a hypothesis with low generalization error. If a hypothesis class has infinite VC dimension, then there exists some source distribution with $R(h) = 0$ for some $h \in \mathcal{H}$, but such that for any sample of any size, there exists an empirical error minimizer with generalization error arbitrarily close to one.

(c) We saw that the ERM learning rule cannot learn with less then D samples. We will now proceed to show that no learning rule can. Show that for *any* hypothesis class $\mathcal{H}$ with VC dimension D, for any learning rule $A(S)$, for any $0 < \epsilon < \frac{1}{4}$, and for any sample size $m < (1-4\epsilon)D$, there exists a source distribution $\mathcal{D}$, such that there is some $h \in \mathcal{H}$ with $R(h) = 0$, but for any $\epsilon > 0$, with probability at least ϵ over a random sample $S \sim \mathcal{D}^m$,

$$R(A(S)) \geq \epsilon. \tag{6}$$

Hint: use, e.g., the No Free Lunch Theorem, to establish a lower bound on the expectation of $R(A(S))$, and then use Markov's inequality to conclude that with probability $\geq \epsilon$, the error must be at least ϵ.

We can conclude that if a hypothesis class has infinite VC-dimension, then for any learning rule and any sample size m there exists a distribution $\mathcal{D}$ with $R(h) = 0$ for some $h \in \mathcal{H}$, but such that with probability at least 0.2, we will have $R(A(S)) > 0.2$.

Problem 3: MDL with Noise

We considered the ERM learning rule which is appropriate for a prior belief which is uniform over some hypothesis class. We also considered the Minimum Description Length (MDL) learning rule which factors in the strength of our prior belief (expressed as description length), but as presented, is appropriate only when we expect a zero-error hypothesis (the MDL learning rule can also be thought of as "choose the zero-empirical-error hypothesis that a-priori we think is most likely").

If we do want to take into consideration the strength of our prior belief, but also want to account for errors, we must balance the empirical error $\hat{R}_S(h)$ with the prior belief, or complexity, of the hypothesis (where we believe simpler, shorter to describe, hypothesis are more likely).

Recall that for any distribution p over hypotheses in $\mathcal{H}$ and and any source distribution $\mathcal{D}$, with probability at least $1 - \delta$ over $S \sim \mathcal{D}^m$, for all $h \in \mathcal{H}$:

$$R(h) \leq \hat{R}_S(h) + \sqrt{\frac{\log \frac{1}{p(h)} + \log \frac{1}{\delta}}{2m}} \tag{7}$$

With (7) in mind, for a prior distribution $p(\cdot)$, define the following learning rule:

$$\text{MDL-SRM}_p(S) = \tilde{h}_S = \arg\min_h \hat{R}_S(h) + \sqrt{\frac{\log \frac{1}{p(h)}}{2m}} \tag{8}$$

Prove that for any distribution $\mathcal{D}$, if there is some hypothesis $h^* \in \mathcal{H}$ with generalization error $R(h^*) = R^*$, then for any $\epsilon > 0$, with probability at least $1 - \delta$ over a sample $S \sim \mathcal{D}^m$ of size:

$$m > \frac{\log \frac{1}{p(h)} + 4 \log \frac{2}{\delta}}{\epsilon^2} \tag{9}$$

we will have $R(\text{MDL-SRM}_p(S)) \leq R^* + \epsilon$. That is, with enough samples, we can get generalization error which is arbitrarily close to the error possible using any hypothesis.

Hint: Take the following steps:

i. Obtain a bound (with probability at least $1 - \frac{\delta}{2}$) on $\hat{R}_S(h^*)$ in terms of $R(h^*)$ that does not depend on p.

ii. Use the inequality $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ to rewrite (7) in terms of the optimization objective of MDL-SRM_p.

iii. Use the above to bound $R(\tilde{h})$ in terms of $\hat{R}_S(h^*)$ and $p(h^*)$, rather then $\hat{R}_S(\tilde{h})$ and $p(\tilde{h})$.

iv. Now use the bound on $\hat{R}_S(h^*)$ to bound $R(\tilde{h})$ in terms of $R(h^*)$ and $p(h^*)$. Be sure to keep track of the minimum probability at which both bounds used hold.

v. Use the inequality $\sqrt{a} + \sqrt{b} \leq \sqrt{2a + 2b}$ to obtain the desired result.

Optional Problem $3\frac{1}{2}$: A Better Balance Between Description Length and Noise

(Will be added later)

Problem 4: Contradiction Between Problems 2 and 3?

The result of Problem 3 holds also for countable hypothesis classes with infinite VC dimension: for any countable hypothesis class $\mathcal{H}$, we can get arbitrary close to the generalization error of any $h \in \mathcal{H}$. But in Problem 2 we saw that the VC-dimension correctly captures the complexity of a hypothesis classes, and that in some sense, no learning guarantee is possible for classes with infinite VC dimension. We will now try to understand why there is no contradiction here.

Consider the following countable hypothesis class of infinite VC dimension: Let $\mathcal{X}$ be the interval $(0, 1]$. For any integer $r > 0$, consider the hypothesis class:

$$\mathcal{H}_r = \{\text{all binary functions of } \phi_r(x) = \lceil r \cdot x \rceil \}.$$

Our hypothesis class will be an infinite union of such classes. To make things a bit simpler, we consider only resolutions r that are integers power of two:

$$\mathcal{H} = \cup_{q=1}^{\infty} \mathcal{H}_{2^q}.$$

(a) Suggest either a binary description language for $\mathcal{H}$ or a distribution over it. It is OK if multiple descriptions refer to the same function, or if you prefer assigning probability to multiple functions that are actually the same one. But be sure that every hypothesis in $\mathcal{H}$ has a description or positive probability mass.

(b) We first establish that the ERM is not appropriate here, as suggested by Problem 2. Consider a source distribution in which X is uniform, and Y is positive if $X < 0.3473$ but negative otherwise (Hint: we could have chosen any function here). Show that for any sample, and any $\epsilon > 0$, there exists a hypothesis in the class with zero empirical error but with $R(h) > 1 - \epsilon$.

(c) The ERM is not appropriate, but MDL-SRM$_p$ is. Calculate an explicit number m_0 (we are looking for an actual number here, not an expression), such that for with probability at least 0.99 over a sample of size $m > m_0$, we will have MDL-SRM$_p(S) < 0.1$, where MDL-SRM$_p$ uses the description language or prior distribution you suggested above.

(d) Suggest a different description language or prior distribution for the same class $\mathcal{H}$ that would require a much smaller training set size to achieve a generalization error of 0.1. Give an example of a source distribution for which the new description language or prior distribution would require a larger training set size to achieve error 0.1.

(e) But we also established that *no* learning rule can enjoy a strong learning guarantee. For the learning rule MDL-SRM$_p$ using the description language or prior distribution you suggested, for any sample size $m > 0$, descried an explicit source distribution such that three exists $h \in \mathcal{H}$ with $R(h) = 0$, but such that $R(\text{MDL-SRM}_p(S)) > 0.2$ with probability at least 0.2 (It is actually possible to get $R(\text{MDL-SRM}_p(S)) = 0.5$ with probability close to one).

Problem 5: VC Dimension

(a) Consider the hypothesis class $\mathcal{H}_\bullet$ of circles in $\mathbb{R}^2$. That is $\mathcal{H}_\bullet$ contains all functions that are positive inside some circle and negative outside. Calculate the VC dimension of this class. Be sure to show that the VC dimension is not lower OR higher than your answer.

(b) Now consider the hypothesis class $\mathcal{H}_\circ$ of positive *and* negative circles in $\mathbb{R}^2$. That is $\mathcal{H}_\circ$ contains all functions that are positive inside some circle and negative outside, and all functions that are negative inside some circles and positive outside. Show how to shatter four points using this class and thus establish a lower bound of four on the VC dimension.

We now consider the VC dimension of the class $\mathcal{H}_d$ of linear separators in $\mathbb{R}^d$:

$$\mathcal{H}_d = \left\{ x \mapsto \text{sign}(w'x + b) \middle| w \in \mathbb{R}^d, b \in R \right\}$$

Here and throughout $\text{sign}(x)$ is 1 if x is positive and -1 otherwise (i.e. we arbitrarily treat zero as negative).

(c) Consider the set of $d + 1$ points that includes the origin 0 and the d points e_i, where e_i is a vector with one in coordinate i and zeros elsewhere (the unit vector along axis i). Show that these points can be shattered by $\mathcal{H}_d$.

(d) Prove that no set of $d + 2$ points can be shattered by $\mathcal{H}_d$. Hint: Use Radon's Theorem, which states that any set of $d + 2$ points in $\mathbb{R}^d$ can be partitioned into two disjoint sets whose convex hulls intersect.

Conclude that the VC-dimension of $\mathcal{H}_d$ is exactly $d + 1$.

(e) Prove that for any $\mathcal{H}_1 \subseteq \mathcal{H}_2$, the VC-dimension of $\mathcal{H}_1$ is not larger than that of $\mathcal{H}_2$.

(f) Use the above to prove that if for some hypothesis class $\mathcal{H}$, there exists a feature map $\phi : \mathcal{X} \to \mathbb{R}^d$ such that any hypothesis $h \in \mathcal{H}$ can be written as

$$h(x) = \text{sign}\left(\sum_{i=1}^{d} w_i \phi_i(x) + b \right) \tag{10}$$

for some $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$, then the VC-dimension of $\mathcal{H}$ is at most $d + 1$.

(g) Use this to obtain a tight upper bound on the VC-dimension of $\mathcal{H}_\circ$ and conclude that the VC-dimension of this class is indeed four. Note that the bound you can get on $\mathcal{H}_\bullet$ is not tight.