**Mohri**

**5.7**

**Q:**

5.7 VC-dimension of canonical hyperplanes. The objective of this problem is derive a bound on the VC-dimension of canonical hyperplanes that does not depend on the dimension of feature space. Let S be a subset of {x: norm of x <= r}. We will show that the VC-dimension d of the set of canonical hyperplanes {x -> sign(w·x): min over w norm of w = 1 and max over all w norm of w <= Λ} verifies d <= r^2 \* Λ^2.

(a) Let {x1, ..., xd} be a set that can be shattered. Show that for all y = (y1, ..., yd) where y is an element of {-1, +1}^d, d <= sum from i=1 to d of yi \* x\_i.

(b) Use randomization over the labels y and Jensen’s inequality to show that d <= Λ \* sqrt(sum from i=1 to d of norm of x\_i squared).

(c) Conclude that d <= r^2 \* Λ^2.

**A:**

5.7 VC-dimension of canonical hyperplanes

(a) By definition of {x1,...,xd}, for all y = (y1,..., yd) ∈ {-1,+1}^d, there exists w such that, ∀i ∈ [d], 1 ≤ yi(w . xi). Summing up these inequalities yields d ≤ w . ∑[i=1]^d yixi ≤ ||w|| \* ∑[i=1]^d |yixi| ≤ Λ \* ∑[i=1]^d |yi\*xi| .

(b) Since this inequality holds for all y ∈ {-1,+1}^d, it also holds on expectation over y1,..., yd drawn i.i.d. according to a uniform distribution over {-1,+1}. In view of the independence assumption, for i ≠ j we have E[yiyj] = E[yi]E[yj]. Thus, since the distribution is uniform, E[yiyj] = 0 if i ≠ j, E[yiyj] = 1 otherwise. This gives d ≤ Λ \* E[ ∑[i=1]^d yixi ] (taking expectations) ≤ Λ \* E[ |∑[i=1]^d yixi| ]^(1/2) (Jensen’s inequality) = Λ \* [∑[i,j=1]^d E[ yi\*yj ](xi . xj) ]^(1/2) = Λ \* [∑[i=1]^d (xi . xi) ]^(1/2) .

Thus, √d ≤ Λ\*r, which completes the proof.

(c) In view of the previous inequality, we can write d ≤ Λ^2\*r^2 = Λr√d.

**6.2k**

**Q:**

Show that the following kernels K are PDS:

(k) For all σ > 0, K(x, y) = exp( sum from i=1 to N of min(|x\_i|, |y\_i| / sigma^2) ) over R^N \* R^N.

Hint: the function (z0, y0) -> integral from 0 to +∞ of 1\_{|e1\_0, |z0 - y0| > t|} dt defined over R x R could be useful for the proof.

**A:**

(k) Observe that for all x, y in R, integral from 0 to +∞ of 1[x in [0,|x|]] + 1[y in [0,|y|]] dt = min{|x|, |y|}.

Thus, (x, y) -> min{|x|,|y|} is a PDS kernel over R x R since for any z1,...,zm in R^N and c1,...,cm in R,

sum from i,j=1 to m of ci\*cj min{|zi|,|zj|}

equals integral from 0 to +∞ of sum from i,j=1 to m and k=1 to N of ci\*cj 1[zi in [0,|zi|]] + 1[zj in [0,|zj|]] dt

equals integral from 0 to +∞ of c1 1[z1 in [0,|z1|]] squared ... dt >= 0.

Thus, (x, y) -> sum from i=1 to N of min{|zi|,|zi|} is a PDS kernel over R^N x R^N as a sum of PDS kernels. Its composition with z -> e^(z^2) with admits a power series with an infinite radius of convergence and non-negative coefficients is thus also PDS, which concludes the proof.

Top of Form

**11.9**

**Q:**

11.9 Leave-one-out error. In general, the computation of the leave-one-out error can be very costly since, for a sample of size m, it requires training the algorithm m times. The objective of this problem is to show that, remarkably, in the case of kernel ridge regression, the leave-one-out error can be computed efficiently by training the algorithm only once.

Let S = ((x1, y1), ..., (xm, ym)) denote a training sample of size m and for any i ∈ [m], let Si denote the sample of size m - 1 obtained from S by removing (xi, yi): Si = S - {(xi, yi)}. For any sample T, let hT denote a hypothesis obtained by training T. By definition (see definition 5.2), for the squared loss, the leave-one-out error with respect to S is defined by R̂LOO(KRR) = 1/m ∑(from i=1 to m) (hS(xi) - yi)^2 .

(a) Let Si' = ((x1, y1), ..., (xi, hS(yi)), ..., (xm, ym)). Show that hSi = hSi'. (b) Define yi = y - yi∗ei + hS(xi)ei, that is the vector of labels with the ith component replaced with hS(xi). Prove that for KRR hS(xi) = yi^T(K + λI)^-1Kei. (c) Prove that the leave-one-out error admits the following simple expression in terms of hS: R̂LOO(KRR) = 1/m ∑(from i=1 to m) (hS(xi) - yi)^2 / ei^T(K + λI)^-1Kei . (d) Suppose that the diagonal entries of matrix M = (K + λI)^-1K are all equal to γ. How do the empirical error R̂S of the algorithm and the leave-one-out error R̂LOO relate? Is there any value of γ for which the two errors coincide?

**A:**

11.9 Leave-one-out error

(a) Note that the hypothesis h\_s,i will make zero error on the ith point of S'\_i and is defined as the minimizer with respect to the remainder of the points. Thus, h\_s,i is also the minimizer with respect to the set S'\_i.

(b) Using part (a) and the definition of the KRR hypothesis with respect to the dual variables we have h\_s(xi) = h\_s,i(xi) = α\_s'\_i K\_ei, where α\_s'\_i is the optimal set of dual variable for KRR trained with S'\_i. Noting that the closed-form solution is α = (K + λI)^-1 y\_i proves the equality.

(c) Using part (b) we can write h\_s,i(xi) - y\_i = y'i (K + λI)^-1 K\_ei - y\_i = (y - y\_ei + h\_s,i(xi)e\_i')^T (K + λI)^-1 K\_ei - y\_i = h\_s(xi) - y\_i + (h\_s,i(xi) - y\_i)e\_i' (K + λI)^-1 K\_ei, which implies h\_s,i(xi) - y\_i = (h\_s(xi) - y\_i)/e\_i' (K + λI)^-1 K\_ei. Thus, we can write R\_LOO(λ) = 1/m Σ(i=1)^m (h\_s(xi) - y\_i)^2 / e\_i' (K + λI)^-1 K\_ei.

**Ben-David**

**10.1**

**Q:**

1. Boosting the Confidence: Let A be an algorithm that guarantees the following: There exist some constant δ0 in (0,1) and a function mH : (0,1) -> N such that for every ε in (0,1), if m ≥ mH(ε) then for every distribution D it holds that with probability of at least 1 - δ0, LD(A(S)) ≤ min h in H LD(h) + ε.

Suggest a procedure that relies on A and learns H in the usual agnostic PAC learning model and has a sample complexity of

mH(ε, δ) ≤ k mH(ε) + [2 log(4k/δ) / ε^2],

where

k = [log(δ) / log(δ0)].

Hint: Divide the data into k + 1 chunks, where each of the first k chunks is of size mH(ε) examples. Train the first k chunks using A. Argue that the probability that for all of these chunks we have LD(A(S)) ≤ min h in H LD(h) + ε is at most δk ≤ δ/2. Finally, use the last chunk to choose from the k hypotheses that A generated from the k chunks (by relying on Corollary 4.6).

**A:**

Let ε, δ ∈ (0, 1). Pick k "chunks" of size mH(ε/2). Apply A on each of these chunks, to obtain h1, ..., hk. Note that the probability that min i LD(hi) ≤ min LD(h) + ε/2 is at least 1 - δk > 1 - δ/2. Now, apply an ERM over the class H' = {h1,..., hk} with the training data being the last chunk of size [2 log(4k/δ) / ε^2]. Denote the output hypothesis by h. Using Corollary 4.6, we obtain that with probability at least 1 - δ/2, LD(h) ≤ min i LD(hi) + ε/2. Applying the union bound, we obtain that with probability at least 1 - δ,

LD(h) ≤ min i LD(hi) + ε/2 for all i in [k] ≤ min h in H LD(h) + ε.

**10.4**

**Q:**

1. In this exercise we discuss the VC-dimension of classes of the form L(B, T). We proved an upper bound of O(dT log(dT)), where d = VCdim(B). Here we wish to prove an almost matching lower bound. However, that will not be the case for all classes B.
2. Note that for every class B and every number T ≥ 1, VCdim(B) ≤ VCdim(L(B, T)). Find a class B for which VCdim(B) = VCdim(L(B, T)) for every T ≥ 1. Hint: Take λ to be a finite set.
3. Let Bd be the class of decision stumps over R^d. Prove that log(d) ≤ VCdim(Bd) ≤ 5 + 2log(d). Hints:
   * For the upper bound, rely on Exercise 11.
   * For the lower bound, assume d = 2^k. Let A be a k x d matrix whose columns are all the d binary vectors in {±1}^k. The rows of A form a set of k vectors in R^d. Show that this set is shattered by decision stumps over R^d.
4. Let T ≥ 1 be any integer. Prove that VCdim(L(Bd, T)) > 0.5T log(d). Hint: Construct a set of T/2\*k instances by taking the rows of the matrix A from the previous question, and the rows of the matrices 2A, 3A, 4A, ..., T/2A. Show that the resulting set is shattered by L(Bd, T).

**A:**

(a) Let X be a finite set of size n. Let B be the class of all functions from X to {0,1}. Then, L(B,T) = B, and both are finite. Hence, for any T ≥ 1,

VCdim(B) = VCdim(L(B,T)) = log2 n.

(b) Denote by B the class of decision stumps in Rd. Formally, B = {hj,b,θ : j ∈ [d], b ∈ {-1,1}, θ ∈ R}, where hj,b,θ(x) = b · sign(θ - xj). For each j ∈ [d], let Bj = {hj,b,θ : b ∈ {-1,1}, θ ∈ R}, where hj,b,θ(x) = b · sign(θ - xj). Note that VCdim(Bj) = 2. Clearly, B = union from j=1 to d of Bj. Applying Exercise 11, we conclude that

VCdim(B) ≤ 16 + 2log d.

• Assume w.l.o.g. that d = 2^k for some k ∈ N (otherwise, replace d by 2^log d). Let A ∈ R^d×k be the matrix whose columns range over the (entire) set {0,1}^k. For each i ∈ [k], let xi = Ai,. We claim that the set C = {x1,...,xk} is shattered. Let I ⊆ [k]. We show that we can label the instances in I positively while the instances [k]\I are labeled negatively. By our construction, there exists an index j such that Aij = xi,j = 1 iff i ∈ I. Then, hi,1,1^2(xi) = 1 iff i ∈ I.

(c) Following the hint, for each i ∈ [Tk/2], let xi = i/kAi,\_. We claim that the set C = {xi : i ∈ [Tk/2]} is shattered by L(Bd,T). Let I ⊆ [Tk/2]. Then I = I₁ ∪ ... ∪ IT/2, where each It is a subset of {(t − 1)k + 1,...,tk}. For each t ∈ [T/2], let jt be the corresponding column of A (i.e., Aij = 1 iff (t − 1)k + 1 ≤ i ∈ It). Let

h(x) = sign(hj₁,1,1^2 + hj₂,1,3^2 + hj₃,1,3^2 + hj₄,1,5^2 + ... + hjT/2-1,1,T^2 + hjT/2,1,T/2-1^2 + hjT/2,1,T/2-1^2).

Then h(xi) = 1 iff i ∈ I. Finally, observe that h ∈ L(Bd,T).

**Lemma 10.3**

10.3.1 The VC-Dimension of L(B,T)

The following lemma tells us that the VC-dimension of L(B,T) is upper bounded by O(VCdim(B)T) (the O notation ignores constants and logarithmic factors).

LEMMA 10.3 Let B be a base class and let L(B,T) be as defined in Equation (10.4). Assume that both T and VCdim(B) are at least 3. Then,

VCdim(L(B,T)) ≤ T(VCdim(B) + 1)(3log(T(VCdim(B) + 1)) + 2).

Proof: Denote d = VCdim(B). Let C' = {x1,...,xm} be a set that is shattered by L(B,T). Each labeling of C' by h ∈ L(B,T) is obtained by first choosing h1,...,hr ∈ B and then applying a halfspace hypothesis over the vector (h1(x),...,hr(x)). By Sauer's lemma, there are at most (em/d)^d different dichotomies (i.e., labelings) induced by B over C'. Therefore, we need to choose T hypotheses, out of at most (em/d)^d different hypotheses. There are at most (em/d)^T ways to do it. Next, for each such choice, we apply a linear predictor, which yields at most (em/T)^T dichotomies. Therefore, the overall number of dichotomies we can construct is upper bounded by

(em/d)^d\*(em/T)^T ≤ m^(d+1)T,

where we used the assumption that both d and T are at least 3. Since we assume that C' is shattered, we must have that the preceding is at least 2^m, which yields

2^m ≤ m^(d+1)T.

Therefore,

m ≤ log(m)^(d + 1)T / log(2).

Lemma A.1 in Chapter A tells us that a necessary condition for the above to hold is that

m ≤ 2^(d + 1)T / log(2) \* (d + 1)T / log(2) ≤ (d + 1)T(3log((d + 1)T) + 2),

which concludes our proof.

**Theorem 10.2**

THEOREM 10.2 Let S be a training set and assume that at each iteration of AdaBoost, the weak learner returns a hypothesis for which ε ≤ 1/2 - γ. Then, the training error of the output hypothesis of AdaBoost is at most

L\_S(h\_s) = 1/m \* ∑\_(i=1)^m I\_h\_(x\_i,y\_i) ≤ exp(-2γ^2T) .

Proof: For each t, denote f\_t = ∑\_(s=1)^t w\_h\_p. Therefore, the output of AdaBoost is f\_T. In addition, denote

Z\_t = 1/m \* ∑\_(i=1)^m e^(-y\_i f\_t(x\_i)).

Note that for any hypothesis we have that I\_h(x,y) ≤ e^(-hy). Therefore, L\_S(f\_T) ≤ Z\_T, so it suffices to show that Z\_T ≤ e^(-2γ^2T). To upper bound Z\_T we rewrite it as

Z\_T = Z\_0 \* Z\_1/Z\_0 \* Z\_2/Z\_1 ... Z\_T/Z\_(T-1) .

where we used the fact that Z\_0 = 1 because f\_0 = 0. Therefore, it suffices to show that for every round t,

Z\_(t+1)/Z\_t ≤ e^(-2γ^2) .

To do so, we first note that using a simple inductive argument, for all t and i,

D\_(i^(t+1)) = e^(-y\_i f\_t(x\_i)) / ∑\_(j=1)^m e^(-y\_j f\_t(x\_j)).

Hence,

Z\_(t+1)/Z\_t = ∑\_(i=1)^m e^(-y\_i f\_(t+1)(x\_i)) / ∑\_(j=1)^m e^(-y\_j f\_t(x\_j))

= ∑\_(i=1)^m e^(-y\_i f\_t(x\_i)-e^(-y\_i w\_t h\_(t+1)(x\_i))) / ∑\_(j=1)^m e^(-y\_j f\_t(x\_j))

= e^(-w\_t) \* ∑ D\_(i^(t+1)) e^(-y\_i w\_t h\_(t+1)(x\_i))

= e^(-w\_t) \* (1 - ε\_(t+1) + e^(w\_t) ε\_(t+1))

= 1 / √(1/ε\_(t+1) - 1) + √(ε\_(t+1) / 1 - ε\_(t+1) + ε\_(t+1))

= 2√(ε\_(t+1) (1 - ε\_(t+1))).

By our assumption, ε\_(t+1) ≤ 1/2 - γ. Since the function g(a) = a(1 − a) is monotonically increasing in [0, 1/2], we obtain that

2√(ε\_(t+1) (1 - ε\_(t+1))) ≤ 2√((1/2 - γ)(1/2 + γ)) = √(1 - 4γ^2).

Finally, using the inequality 1 - a ≤ e^(-a) we have that √(1 - 4γ^2) ≤ e^(-2γ^2/2) = e^(-2γ^2). This shows that Equation (10.3) holds and thus concludes our proof.

**7.1**

**7.3**

**7.10**

**7.12 (a, b, f)\***