Massachusetts Institute of Technology
6.435 Theory of Learning and System Identification

Prof. Dahleh, Prof. Mitter	Homework 2
Out 2/25	Due F, 3/9

1 [Generalized Glivenko-Cantelli] Consider the set of events characterized by:

$$A_\alpha = \{z | L(z, \alpha) = 1\}, \alpha \in \Lambda.$$

Show that the frequencies will converge to their probabilities *uniformly* over the set of events described above if and only if the set of functions $L(z, \alpha), \alpha \in \Lambda$ has a finite VC dimension.

2 [Markov Chains] Consider a discrete Markov chain of order n. Assume the chain has a single recurrent aperiodic class. Suppose we make the observations $x_1, x_2, \cdots, x_\ell$.

1. Compute the ML estimate of the transition probability matrix.
2. What is the limiting behavior of such estimates? You don't need to prove your claim, just quote the appropriate results.
3. Define the KL distance between two chains with the same set of states and with probability functions P and Q, as:

$$D(P||Q) = \lim_{n\to\infty} \frac{1}{n} D(P_{X_1,\cdots,X_n} || Q_{X_1,\cdots,X_n}).$$

Show that $D(P||Q) = D(P_{X_i|X_{i-1}} || Q_{X_i|X_{i-1}})$, where the conditional KL distance is computed relative to the stationary distribution of the first chain. Recall the definition:

$$D(P_{Y|X} || Q_{Y|X}) = \mathbf{E}_{P_X P_{Y|X}} \left[\log \frac{P_{Y|X}}{Q_{Y|X}} \right].$$

[Hint: Use the chain rule $D(P_{X,Y} || Q_{X,Y}) = D(P_X || Q_X) + D(P_{Y|X} || Q_{Y|X})$.]

4. Show that if P_α is a parametrization of a class of transition probabilities, then the ML estimate from this class has the uniform convergence property. Show that the ML estimate converges to the $\min_\alpha D(P||P_\alpha)$. [Hint: Imitate the lecture for the finite range case.]

3 [VC Dimension and Parametrization]

(a) Consider the class of one-dimensional functions:

$$y = \theta\left(\sum_{j=1}^{n} |a_j x^j| \cdot \operatorname{sgn}(x) + a_0 \right), \qquad a_j, x \in \mathbf{R}.$$

What is the VC dimension of this class? How does it relate to the number of parameters describing the function?

(b) Consider the class of functions:

$$y = \theta\Big(\sin(\beta x)\Big), \qquad \beta \in (0, \infty), \quad x \in (0, 2\pi).$$

Show that this one-parameter class of functions has infinite VC dimension.

4 [Convex Polytopes] Let $\mathcal{C}$ be the set of two-dimensional convex polygones with finite but arbitrary number of faces. Consider the class of two-dimensional functions defined as the interiors of polygones in $\mathcal{C}$:

$$y = I_A(x) \qquad x \in \mathbf{R}^2, \quad A \in \mathcal{C}.$$

(a) Show that the VC dimension of this class is infinite.

(b) Assume the true model is the interior of some $A \in \mathcal{C}$. Consider the algorithm which starts with $S_0 = \emptyset$ and $\hat{A}_0 = \emptyset$, and computes the convex hull of positive samples:

$$\begin{aligned} S_k &= \begin{cases} S_{k-1} \cup \{x_k\}, & \text{if } y_k = 1 \text{ and } x_k \notin \hat{A}_{k-1}, \\ S_{k-1}, & \text{otherwise;} \end{cases} \\ \hat{A}_k &= \mathsf{ConvexHull}(S_k). \end{aligned}$$

Show that if the data is sampled uniformly in a rectangle containing A, then the algorithm converges:

$$\mathbf{P}\big(A \vartriangle \hat{A}_k\big) \to 0, \qquad k \to \infty,$$

where $X \vartriangle Y$ is the symmeteric difference of X and Y, i.e. $X \vartriangle Y = (X \cap Y^c) \cup (X^c \cap Y)$. (In this case, you can show that $\hat{A}_k \subset A$, and thus $A \vartriangle \hat{A}_k = A \setminus \hat{A}_k$.)

(c) Is there a discrepancy between the implications of parts (a) and (b)? Justify your answer.

5 [optional] [Δ-Margin Separating Hyperplane]

Consider the class of N-dimensional functions:

$$y = \begin{cases} 1, & \text{if } w'x - b \geq \Delta, \\ 0, & \text{if } w'x - b \leq -\Delta, \end{cases} \qquad w, b, x \in \mathbf{R}^N, \quad |w| = 1.$$

Note that the class defines hyperplanes separating the space into two halves, but has no specification for points lying within a margin Δ of each half-space. If we think of the function as a classifier, such points cannot be labeled, and are considered misclassified.

Show that, for $N = 2$, the VC dimension of the Δ-margin separating hyperplane is bounded from above by:

$$\min\left\{N, \frac{R^2}{\Delta^2}\right\} + 1,$$

where R is the radius of the smallest ball containing all the data points. (The result can be generalized to all N).