Introduction to Machine Learning **Fall Semester**

Homework 2: Nov 11, 2019

Due: Nov 26, 2019 (See the submission guidelines in the course web site)

Theory Questions

1. **Singletons** (Section 3.5, Ex.2 in the course textbook). Let $\mathcal{X}$ be a discrete domain, and let $\mathcal{H}_{Singleton} = \{h_z : z \in \mathcal{X}\} \cup \{h^-\}$, where for each $z \in \mathcal{X}$, h_z is the function defined by $h_z(x) = 1$ if $x = z$ and $h_z(x) = 0$ if $x \neq z$. h^- is simply the all-negative hypothesis, namely, $\forall x \in \mathcal{X}, h^-(x) = 0$.

   The realizability assumption here implies that the true hypothesis f labels negatively all examples in the domain, perhaps except one.

   (a) Describe an algorithm that implements the ERM rule for learning $\mathcal{H}_{Singleton}$ in the realizable setup.

   (b) Show that $\mathcal{H}_{Singleton}$ is PAC learnable. Provide an upper bound on the sample complexity.

2. **PAC in Expectation**. Consider learning in the realizable case. We say an hypothesis class $\mathcal{H}$ is ***PAC learnable in expectation*** if there exists a learning algorithm A and a function $N(a) : (0, 1) \to \mathbb{N}$ such that $\forall a \in (0, 1)$ and for any distribution P, given a sample set S, such that $|S| \geq N(a)$ it holds that,

   $$\mathbb{E}\left[e_P(A(S))\right] \leq a$$

   Show that $\mathcal{H}$ is PAC learnable *if and only if* $\mathcal{H}$ is PAC learnable in expectation (Hint: Use Markov's inequality and refer to derivations between equations 3.13-3.14 in the lecture scribes about VC).

3. **Union Of Intervals.** Determine the VC-dimension of the subsets of the real line formed by the union of k intervals (see question 1 of the programming assignment for a formal definition of $\mathcal{H}$).

4. **Right-angle Triangles.** Determine the VC-dimension of the hypotheses class $\mathcal{H}$, of axis-aligned right-angle triangles in the plane, with the right angle in the lower left corner.

5. **Structural Risk Minimization.** Let $\mathcal{H}$ be a countable hypothesis class, that is, $\mathcal{H}$ can be written as $\mathcal{H} = \bigcup_{i \in \mathbb{N}} \{h_i\}$. Let $w : \mathcal{H} \to [0, 1]$ be a function such that $\sum_{h \in \mathcal{H}} w(h) \leq 1$. We refer to w as a *weight function* over the hypotheses which reflects the prior for each hypothesis.

(a) Show that with probability $1 - \delta$ over the choice $S \sim P\ (|S| = n)$

$$\forall h \in \mathcal{H},\ e_P(h) \leq e_S(h) + \sqrt{\frac{1}{2n} \ln \frac{2}{\delta \cdot w(h)}}$$

Hint: use the uniform convergence property for each hypothesis class $\{h_i\}$ of size 1.

(b) Denote $h_{srm} = \underset{h \in \mathcal{H}}{argmin}\ e_S(h) + \sqrt{\frac{1}{2n} \ln \frac{2}{\delta \cdot w(h)}}$. Find a high probability upper bound for $|e_P(h_{srm}) - \underset{h \in \mathcal{H}}{min}\ e_P(h)|$.

Programming Assignment

1. **Union Of Intervals.** In this question, we will study the hypothesis class of a finite union of disjoint intervals, and the properties of the ERM algorithm for this class.

   To review, let the sample space be $\mathcal{X} = [0,1]$ and assume we study a binary classification problem, i.e. $\mathcal{Y} = \{0,1\}$. We will try to learn using an hypothesis class that consists of k intervals. More explicitly, let $I = \{[l_1,u_1],\dots,[l_k,u_k]\}$ be k disjoint intervals, such that $0 \leq l_1 \leq u_1 \leq l_2 \leq u_2 \leq \ldots \leq u_k \leq 1$. For each such k disjoint intervals, define the corresponding hypothesis as

   $$h_I(x) = \begin{cases} 1 & \text{if } x \in [l_1,u_1],\dots,[l_k,u_k] \\ 0 & \text{otherwise} \end{cases}$$

   Finally, define $\mathcal{H}_k$ as the hypothesis class that consists of all hypotheses that correspond to k disjoint intervals:

   $$\mathcal{H}_k = \{h_I | I = \{[l_1,u_1],\dots,[l_k,u_k]\},\ 0 \leq l_1 \leq u_1 \leq l_2 \leq u_2 \leq \ldots \leq u_k \leq 1\}$$

   We note that $\mathcal{H}_k \subseteq \mathcal{H}_{k+1}$, since we can always take one of the intervals to be of length 0 by setting its endpoints to be equal. We are given a sample of size $m = \langle x_1, y_1\rangle, \dots, \langle x_n, y_m\rangle$. Assume that the points are sorted, so that $0 \leq x_1 < x_2 < \ldots < x_m \leq 1$.

   Download the file `intervals.py` from Moodle, you will find there a function that implements an ERM algorithm for $\mathcal{H}_k$. Given a sorted list `xs`=$[x_1,\dots,x_m]$, the respective labeling `ys`=$[y_1,\dots,y_m]$ and k, the given function `find_best_interval` returns a list of up to k intervals and their error count on the given sample. These intervals have the smallest empirical error count possible from all choices of k intervals or less.

   Note that in sections (d)-(f) you will need to use this function for large values of m. Execution in these cases could take time (more than 10 minutes for experiment), so plan ahead.

   (a) Assume that the true distribution $P[x,y] = P[y|x] \cdot P[x]$ is: x is distributed uniformly on the interval $[0,1]$, and

   $$P[y=1|x] = \begin{cases} 0.8 & \text{if } x \in [0,0.2] \cup [0.4,0.6] \cup [0.8,1] \\ 0.1 & \text{if } x \in [0.2,0.4] \cup [0.6,0.8] \end{cases}$$

   and $P[y=0|x] = 1 - P[y=1|x]$.

   Write a function that draws m pairs of (x,y) according to the distribution P. Use it to draw a sample of size $m = 100$ and create a plot:

   i. Plot the points and their label (have the y axis in range $-0.1, 1.1$ for clarity of presentation).

   ii. Mark the lines $x = 0.2, 0.4, 0.6, 0.8$ clearly on the plot.

iii. Run the `find_best_interval` function on your sample with $k = 3$, and plot the intervals clearly.

(b) Note that here, we know the true distribution P, so for every given hypothesis $h \in \mathcal{H}_k$, we can calculate $error(h)$ precisely. What is the hypothesis with the smallest error?

(c) Write a function that, given a list of intervals, calculates the error of the respective hypothesis. Then, for $k = 3$, $m = 10, 15, 20, \ldots, 100$, perform the following experiment $T = 100$ times: (i) Draw a sample of size m and run the ERM algorithm on it; (ii) Calculate the empirical error for the returned hypothesis; (iii) Calculate the true error for the returned hypothesis. Plot the average empirical and true errors, averaged across the T runs, as a function of m. Discuss the results. Do the empirical and true error decrease or increase in m? Why?

(d) Draw a data set of $m = 1500$ samples. Find the best ERM hypothesis for $k = 1, 2, \ldots, 20$, and plot the empirical and true errors as a function of k. How does the error behave? Define k^* to be the k with the smallest empirical error for ERM? Does this mean the hypothesis with k^* intervals is a good choice?

(e) Now we will use the principle of structural risk minimization (SRM), to search for a k that gives good test error. Let $\delta = 0.1$:

- Use your results from question 3 in the theoretical part and the VC confidence bound to construct a penalty function.
- Draw a data set of $m = 1500$ samples, run the experiment in (d) again, but now plot two additional lines as a function of k: 1) the penalty for the best ERM hypothesis and 2) the sum of penalty and empirical error.
- What is the best value for k in each case? is it better than the one you chose in (d)?

(f) Here we will use holdout-validation to search for a k that gives good test error. Draw a data set of $m = 1500$ samples and use 20% for a holdout-validation. Choose the best hypothesis based on 3 experiments. Discuss how close this gets you to finding the hypothesis with optimal true error.