TTIC 31020: Introduction to Machine Learning **Winter 2025**

Homework 3

Due: January 23, 2025

Note *You may discuss these problems in groups. However, you must write up your own solutions and mention the names of the people in your group. Also, please do mention any books, papers or other sources you refer to. We prefer and encourage that you typeset solutions in LaTeX for written assignments. Handwritten solutions that are not neat, organized and with clear and easy-to-read handwriting will not be graded. Please submit your solutions as a PDF document on Canvas.*

Challenge and Optional Questions *Challenge questions are marked with* challenge *and optional questions are marked with* optional *. You can get extra credit for solving* challenge *questions. You are not required to turn in* optional *questions, but we encourage you to solve them for your understanding. Course staff will help with* optional *questions but will prioritize queries on non-*optional *questions.*

1. **Shattering Ellipses in $\mathbb{R}^2$**

   Consider the class:

   $$\mathcal{H} = \left\{ \mathbb{I}\left[\frac{(x_1 - c_1)^2}{a_1^2} + \frac{(x_2 - c_2)^2}{a_2^2} \leq r \right] : c_1, c_2, a_1, a_2, r \in \mathbb{R}, a_1 \neq a_2 \right\} .$$

   (a) Show how to shatter 4 points using this class.

   (b) Derive features $\phi : \mathbb{R}^2 \mapsto \mathbb{R}^5$ to represent this class as linear predictors over these features.

   (c) What is the VC dimension of the class you derived in the previous part? How does this compare to the lower bound on VC dimension from part (a) of the original class?

2. **Shattering Sparse Linear Predictors**

   Consider the class:

   $$\mathcal{H}_k = \{ y = \text{sgn}(\langle w, x \rangle) \quad : \quad w \in \mathbb{R}^d \text{ s.t. } ||w||_0 = k \}$$

   (a) Show how to shatter $\Omega(\log d)$ points with respect to $\mathcal{H}_1$.

   (b) Show how to shatter $\max\{k, \log d\}$ points with respect to $\mathcal{H}_k$.

   (c) challenge optional Show how to shatter $\Omega(k \log(d/k))$ points with respect to $\mathcal{H}_k$.

3. **Realizable Online-to-Batch and Leave-One-Out Cross-Validation**

   In this question we will make a formal connection between the online mistake bound and the sample complexity in the statistical setting. We will also study the Leave-One-Out Cross-Validation (LOOCV) estimate of the error of a learning rule.

   Consider a hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ and an online learning rule $\mathcal{A} : (\mathcal{X} \times \mathcal{Y})^* \to \mathcal{Y}^{\mathcal{X}}$ (recall that a learning rule can be formalized as a mapping from the sequence of examples seen so far to a predictor), and suppose $\mathcal{A}$ enjoys a mistake bound M for sequences realized by $\mathcal{H}$ (that is, $\mathcal{A}$ makes at most M mistakes on any sequence $(x_i, y_i)_i$ s.t. there exists some $h \in \mathcal{H}$ for which $y_i = h(x_i)$ for all i.). We want to use this online rule to get a good learning rule in the *statistical* setting, where we are given a sample $S \sim \mathcal{D}^m$ of examples from some unknown source distribution $\mathcal{D}(\mathcal{X}, \mathcal{Y})$. We will use the following construction of a statistical (batch) rule $\tilde{\mathcal{A}}$ that uses the online rule $\mathcal{A}$:

   That is, $\tilde{\mathcal{A}}$ runs the online learning rule $\mathcal{A}$ on a sequence of (possibly repeated) examples chosen from S, keeping track of the examples the current predictor $\mathcal{A}$ uses in S'. As long as there is an example in S misclassified by the current predictor, this misclassified example is fed as the next example to $\mathcal{A}$.

Input: Training set $S = \left\{(x_i, y_i)_{i=1}^m\right\}$
Output: Predictor $h : \mathcal{X} \to \mathcal{Y}$ with zero training error, i.e. $h \in \mathcal{H}$ s.t. $\forall\, i, h\left(x_i\right) = y_i$.
Initialize $S' = \varnothing$ (empty sequence of examples) and $h = \mathcal{A}(\varnothing)$ the initial predictor used by the online rule before seeing any examples.
while *there exists* $(x_i, y_i) \in S$ *s.t.* $h\left(x_i\right) \neq y_i$ **do**
  Append (x_i, y_i) to S' and update $h = \mathcal{A}\left(S'\right)$ (i.e. feed (x_i, y_i) to the online rule $\mathcal{A}$).
end
return h

Note that since an example might be misclassified in multiple rounds, it might appear in the sequence S' multiple times, i.e. be fed to $\mathcal{A}$ multiple times. If in some round there are multiple misclassified example, we need to choose one of them—this might be, e.g. the example with the lowest index (although this arbitrary choice doesn't matter).

To analyze $\tilde{\mathcal{A}}$ in the statistical setting, we will use the LOOCV error estimate for a learning rule. Given a training set $S = \left\{(x_i, y_i)_{i=1}^m\right\}$, denote $S_{-i} = \left\{(x_1, y_1), \ldots, (x_{i-1}, y_{i-1}), (x_{i+1}, y_{i+1}), \ldots, (x_m, y_m)\right\}$ the training set with example (x_i, y_i) removed. The LOOCV error rate for a learning rule $\mathcal{F}$ on training set S is obtained by, for each i, running $\mathcal{F}$ on S_{-i} and testing on (x_i, y_i):

$$\mathrm{LOOCV}_S(\mathcal{F}) = \frac{1}{m}\left|\left\{i \middle| \text{ s.t. } \mathcal{F}\left(S_{-i}\right)\left(x_i\right) \neq y_i\right\}\right|. \tag{1}$$

(a) Prove that the LOOCV error on a set of m i.i.d. samples from $\mathcal{D}$ is an unbiased estimator of the expected error rate of $\mathcal{F}$ on $m-1$ i.i.d. samples from $\mathcal{D}$, that is:

$$\mathbb{E}_{S\sim\mathcal{D}^m}\left[\mathrm{LOOCV}_S(\mathcal{F})\right] = \mathbb{E}_{S\sim\mathcal{D}^{m-1}}\left[L_{\mathcal{D}}(\mathcal{F}(S))\right]. \tag{2}$$

(b) Consider running $\tilde{\mathcal{A}}$ on m i.i.d. samples $S \sim \mathcal{D}^m$ from a distribution realizable by $\mathcal{H}$[1]. If $\tilde{\mathcal{A}}$ runs for T iterations, i.e. feeds a total of T examples to $\mathcal{A}$, and so $\mathcal{A}$ is run for T steps, how many mistakes does $\mathcal{A}$ make on examples it is fed? Determine a bound on the number of iteration $\tilde{\mathcal{A}}$ might run.

(c) For a set S of $m+1$ examples, bound $\mathrm{LOOCV}_S(\tilde{\mathcal{A}})$ in terms of the number of iterations $\tilde{\mathcal{A}}$ would run on S.

(d) By combining (a) and (c), determine a bound on the expected error rate of the predictor output by $\tilde{\mathcal{A}}$, $\mathbb{E}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}(\tilde{\mathcal{A}}(S))\right]$, when run on m i.i.d. samples from a distribution realized by $\mathcal{H}$, in terms of the mistake bound M and sample size m. How many samples, as a function of M and ϵ are sufficient to ensure $\mathbb{E}_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}(\tilde{\mathcal{A}}(S))\right] < \epsilon$?

(e) optional Why is it not sufficient to just use the rule $\mathcal{A}$ itself, i.e. run $\mathcal{A}$ on the sequence of examples S and output the final predictor? Give a counterexample where $\mathcal{A}$ has a small mistake bound, but for some large m (using the output of the predictor after m steps on i.i.d. examples from a realizable $\mathcal{D}$) results in a very bad predictor.

Remark. We will use the online-to-batch conversion to analyze the PERCEPTRON algorithm in Homework 4.

[1] Recall that a distribution $\mathcal{D}$ is realizable by $\mathcal{H}$ if $\exists\, h \in \mathcal{H}$ s.t. $L_{\mathcal{D}}(h) = 0$.