TTIC 31020: Intro to ML **Winter 2024**

Homework 2

Due: 6 p.m., January 16th, 2024

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 recommend that you typeset your solutions in LaTeX. 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.*

Notation In all questions, we will use the following notation: let $\mathcal{X}$ and $\mathcal{Y}$ be the instance space and label space respectively. A predictor is a mapping $h : \mathcal{X} \to \mathcal{Y}$, which outputs a label $h(x)$ for data point x.

Discussion We discussed memorization in HW 1, and pointed out how limited it is, and why. We also pointed out that one way of extending memorization is by letting each observed label y_i be associated also with instances x that are close to x_i, if not identical to it.

In this discussion, we will first discuss the Bayes Optimal Predictor (or simply "Bayes Predictor") for a joint distribution over $(\mathcal{X}, \mathcal{Y})$; then define the Parzen Window Predictor, which is the Bayes Optimal Predictor for a certain estimated joint distribution.

Recall that for a joint distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$ and a predictor h we defined its *expected error* as:

$$L_{\mathcal{D}}(h) = \mathbf{P}_{(x,y)\sim\mathcal{D}}\left[h(x) \neq y\right]. \tag{1}$$

A Bayes Optimal Predictor is a mapping $h : \mathcal{X} \to \mathcal{Y}$ with the smallest possible error[1]

$$h_{\text{Bayes}(\mathcal{D})} = \arg\min_h L_{\mathcal{D}}(h), \tag{2}$$

and the Bayes Error of $\mathcal{D}$ is the error $L_{\mathcal{D}}\left(h_{\text{Bayes}(\mathcal{D})}\right)$ of a Bayes Optimal Predictor, i.e. the smallest possible error, $\min_h L_{\mathcal{D}}(h)$. Since the prediction $h(x)$ of each instance x is unconstrained by the prediction on any other $x' \neq x$, the optimal prediction of x (minimizing the probability of error) is the label most frequently associated with x. A Bayes Optimal Predictor is then:

$$h_{\text{Bayes}(\mathcal{D})}(x) = \arg\max_{y\in\mathcal{Y}} \mathbf{P}_{\mathcal{D}}(Y = y|X = x) = \arg\max_{y\in\mathcal{Y}} \mathbf{P}_{\mathcal{D}}(X = x, Y = y). \tag{3}$$

For binary classification problems, with $\mathcal{Y} = \{-1, +1\}$, this can also be written as:

$$h_{\text{Bayes}(\mathcal{D})}(x) = \text{sign}\left(\eta_{\mathcal{D}}(x) - \frac{1}{2}\right) \tag{4}$$

where $\eta_{\mathcal{D}}(x) = \mathbf{P}_{\mathcal{D}}(Y = +1|X = x)$ is the *posterior* probability of the label being positive after observing x.

[1] If you want to be formal and very careful, the minimization is over h that are *measureable* in the σ-algebra over which $\mathcal{D}$ is defined. This can be important since for an abstract space $\mathcal{X}$ (e.g. all possible views, or all possible people) this σ-algebra indicates how we observe $x \in \mathcal{X}$, or what the predictor can depend on. **But in this course we will not discuss or worry about measureability, certainly not in a formal way.**

Comprehension and review questions (Do NOT turn these in) When is the Bayes Error equal to zero? How does $\eta(x)$ look like when the Bayes Error is equal to zero? Can you write an expression for the Bayes Error in terms of $\eta(x)$ (hint: take an expectation w.r.t. x)? Is the Bayes Predictor unique? When is it not unique and what happens then? What happens when some x are outside the support of $\mathcal{D}$ (or rather, its marginal over X)? Note that in this case the conditional $\mathbf{P}(Y|X = x)$, and so $\eta(x)$ are not even defined. Prove the characterization Equation (3). Prove the second equality in Equation (3) by using Bayes Rule and noting that the marginal $\mathbf{P}(x)$ doesn't depend on y. Show that for binary problems, Equation (3) is given by Equation (4).

If we knew, and could directly work with, the true joint distribution $\mathcal{D}(\mathcal{X}, \mathcal{Y})$, we would just use the Bayes Optimal Predictor. When we do not know $\mathcal{D}$, or perhaps when it is too complicated to represent and work with, we can build an estimated $\hat{\mathcal{D}}$ based on a sample $S = \{(x_i, y_i)\}_{i=1}^m$ and then use the Bayes Optimal Predictor w.r.t. $\hat{\mathcal{D}}$.

1. **Parzen Window Predictor**

   Given $S = \{(x_i, y_i)\}_{i=1}^m$, the Parzen Window (a.k.a. Kernel) Density Estimate $\hat{f}(x|Y = y)$ of the conditional densities[2] $f(x|Y = y)$, for each label $y \in \mathcal{Y}$, is defined as:

   $$\hat{f}(x|y) = Z_y \sum_{i \text{ s.t. } y_i = y} K(x, x_i) \tag{5}$$

   where $Z_y \in \mathbb{R}$ is a normalization factor that ensures $\int_x \hat{f}(x|y)\, dx = 1$, and K is the kernel[3]

   $$K(x, x_i) = e^{-\rho(x, x_i)^2/\sigma^2} \tag{6}$$

   where $\rho(x, x_i)$ is some shift-invariant distance measure $\rho : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$, e.g. $\rho(x, x_i) = \|x - x_i\|$ with $\mathcal{X} = \mathbb{R}^d$, and σ is a hyper-parameter we need to manually set when using the method. The shift-invariance implies that $\int_x K(x, x_i)\, dx$ is the same constant for all x_i. The Parzen Window estimator $\hat{\mathcal{D}}(\mathcal{X}, \mathcal{Y})$ of the joint distribution is given by the conditional densities in Equation (5) combined with the empirical marginal $\hat{p}(y) = \frac{1}{m} |\{i|y_i = y\}|$. The empirical marginal is simply the count of different labels among the m samples. The Parzen Predictor is then defined as the Bayes Optimal Predictor for $\hat{\mathcal{D}}$. Finally, the Parzen Predictor is specified by the choice of distance measure $\rho(x, x')$ and width parameter σ.

   (a) Show that for binary $\mathcal{Y} = \{-1, +1\}$, the Parzen Predictor is given by:

   $$h(x) = \text{sign}\left(\sum_{i=1}^m y_i K(x, x_i)\right). \tag{7}$$

   (b) Using the kernel $K(x, x') = e^{-\rho(x, x')^2/\sigma^2}$, how does the Parzen predictor h behave in the limit as $\sigma \to \infty$?

   (c) Using the same kernel, how does h behave in the limit as $\sigma \to 0$?

   (d) How is the Parzen predictor related to the Nearest Neighbor predictor [we studied in HW1] when $\sigma \to 0$? Consider, in particular, the situation of "ties", i.e. where there are multiple points in S that are the same distance to x.

[2] If you are worried about whether it is well defined to discuss a density here, you are right. To do so we need to refer to some base measure on $\mathcal{X}$. The choice of this base measure will affect the meaning of the density, and so the resulting estimated distribution. This choice of base measure is not so obvious, especially for an abstract space $\mathcal{X}$, and could be thought of as part of our inductive bias and assumptions about the structure of $\mathcal{X}$. But this discussion is beyond the scope of this course.

[3] Beyond the Gaussian kernel used here, other kernels are also in common use. See, e.g., the Wikipedia page for "Kernel (statistics)". Many common kernels have bounded support, including the original "window", $K(x, x_i) = 1_{\rho(x, x_i) < 1}$. In this question, we use an exponentially decaying kernel to make a connection with the nearest neighbor predictor. **Question to think about:** which answers in this question would change, and how, if we use other kernels?

Discussion In HW1 we showed that the sample complexity of Nearest Neighbor prediction is exponential in the dimension, even in very simple cases. You might want to verify that the same holds also for Parzen Window prediction with any width, or even if the width is selected based on the data, as well as with the k-Nearest Neighbor predictor considered below.

2. **Nearest Neighbor and k-Nearest Neighbor in the Statistical Setting**

   In this question we study Nearest Neighbor prediction based on a sample $S \sim \mathcal{D}^m$ of m i.i.d. samples from some source distribution $\mathcal{D}(X, Y)$.

   (a) Consider a source distribution where $x \in \mathcal{X} = \mathbb{R}$ is uniform on $[-1, 1] \subset \mathbb{R}$, $\mathcal{Y} = \{\pm 1\}$ and $\mathbf{P}_{\mathcal{D}}(Y = +1|x) = 0.5 + 0.3\,\mathrm{sign}(x)$ (i.e. either 0.2 or 0.8). What is the Bayes Optimal Predictor and the Bayes Error?

   (b) For the above source distribution, how does the error $L(h_m)$ of the Nearest Neighbor predictor h_m behave (what does it converge to) as the number of samples increases, i.e. $m \to \infty$? (Hint: when $m \to \infty$, the nearest neighbor of most x in S would have the same sign as x, but how is the label of this Nearest Neighbor distributed?)

   Discussion As you saw, not only might Nearest Neighbor prediction require exponential in the dimension many samples (see also Section 19.1 of [UML]), but even as $m \to \infty$, the Nearest Neighbor Predictor might be much worse than Bayes Optimal Predictor. On the positive side, it is possible to show that $O\left(2^d\right)$-many samples, the Nearest Neighbor error is not more than twice the Bayes error [CH67]. One way to get a better upper bound than twice the Bayes error is by considering the **k-Nearest Neighbor Predictor** h where $h(x)$ is the majority label among the $k > 1$ points in S closest to x. For a theoretical analysis showing how, with $O\left(2^d\right)$-many samples and k increasing, k-Nearest Neighbor classification will approach the Bayes error, see [DGL96]. Both seminal results are conveniently summarized in lecture notes [KS09].

   Experimentation You will investigate k-Nearest Neighbor prediction empirically in the jupyter notebook.

3. **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?

4. **Shattering Sparse Linear Predictors**

   Consider the class:

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

   (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 $\mathcal{H}_k$.

Experimentation:

Look at the Jupyter Notebook, you will work with k-Nearest Neighbor and Decision Tree predictors on synthetic and real-world datasets.

References

[CH67] Thomas M. Cover and Peter E. Hart. Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory*, 13:21–27, 1967.

[DGL96] Luc Devroye, László Györfi, and Gábor Lugosi. *Nearest Neighbor Rules*, pages 61–90. Springer New York, New York, NY, 1996.

[KS09] Sham Kakade and Gregory Shakhnarovich. Lecture notes in large scale learning, 2009.