

# Homework One, for Thu 1/14

CSE 250B

**Note: Your homework must be typeset and uploaded in PDF format to Gradescope by midnight on the due date.**

1. *Prototype selection.* One way to speed up nearest neighbor classification is to replace the training set by a carefully chosen subset of “prototypes”.

Think of a good strategy for choosing prototypes from the training set, bearing in mind that the ultimate goal is good classification performance. Assume that 1-NN will be used.

Then implement your algorithm, and test it on the MNIST dataset, available at:

<http://yann.lecun.com/exdb/mnist/index.html>

What to turn in:

- (a) A short, high-level description of the idea for prototype selection.  
A few sentences should suffice.
  - (b) Concise and unambiguous pseudocode. (Please do not submit any actual code.)  
Your scheme should take as input a labeled training set as well as a number  $M$ , and should return a subset of the training set of size  $M$ .
  - (c) A table of results showing classification performance on MNIST for a few values of  $M$ , including at the very least  $M = 10000, 5000, 1000$ . In each case, you should compare the performance to that of uniform-random selection (that is, picking  $M$  of the training points at random). For any strategy with randomness, you should do several experiments and give error bars – give all relevant details.
2. *Bayes optimality.* Consider the following setup:
    - Input space  $\mathcal{X} = [-1, 1] \subset \mathbb{R}$ .
    - Input distribution:  $\mu(x) = |x|$ .
    - Label space  $\mathcal{Y} = \{0, 1\}$ .
    - Conditional probability function

$$\eta(x) = \Pr(Y = 1|X = x) = \begin{cases} 0.2 & \text{if } x < -0.5 \\ 0.8 & \text{if } -0.5 \leq x \leq 0.5 \\ 0.4 & \text{if } x > 0.5 \end{cases}$$

- (a) What is the Bayes optimal classifier in this setting? What is the optimal risk  $R^*$ ?
- (b) Suppose we obtain the following training set of four labeled points:

$$(-0.8, 0), (-0.4, 1), (0.2, 1), (0.8, 0).$$

What is the decision boundary of 1-NN using this training set? What is the (true) error rate of this classifier, on the underlying distribution given by  $\mu$  and  $\eta$ ?

- (c) In a binary setting, there are two possible errors:  $0 \rightarrow 1$  (label is 0 but prediction is 1) or  $1 \rightarrow 0$  (label is 1 but prediction is 0). Suppose these errors have different costs,  $c_{01}$  and  $c_{10}$ , respectively. We can then define the cost-sensitive risk of a classifier  $h : \mathcal{X} \rightarrow \{0, 1\}$  as

$$R(h) = c_{01}\Pr(Y = 0, h(X) = 1) + c_{10}\Pr(Y = 1, h(X) = 0).$$

In the example above, what is the classifier that minimizes this cost-sensitive risk, if  $c_{01} = 1$  and  $c_{10} = 0.1$ ?

- (d) Now consider a setting with  $\mathcal{Y} = \{0, 1\}$  and with arbitrary  $\mathcal{X}, \mu, \eta, c_{01}, c_{10}$ . Write down an expression for the classifier with minimum cost-sensitive risk.
3. *Properties of metrics.* Which of the following **distance functions are metrics**? In each case, either prove it is a metric or give a counterexample showing that it isn't.
- (a)  $\ell_1$  distance.
  - (b)  $d_1 + d_2$ , where  $d_1$  and  $d_2$  are each metrics.
  - (c) Let's say  $\Sigma$  is a finite set and  $\mathcal{X} = \Sigma^m$ . The *Hamming distance* on  $\mathcal{X}$  is

$$d(x, y) = \# \text{ of positions on which } x \text{ and } y \text{ differ.}$$

- (d) Squared Euclidean distance on  $\mathbb{R}^m$ , that is,

$$d(x, y) = \sum_{i=1}^m (x_i - y_i)^2.$$

(It might be easiest to consider the case  $m = 1$ .)

- (e) Let  $\mathcal{X}$  be the space of probability distributions over  $m$  outcomes. We can represent any such distribution as a vector of  $m$  nonnegative numbers that sum to 1 (corresponding to the probabilities of each of the outcomes). That is,  $\mathcal{X} = \{p \in \mathbb{R}^m : p_i \geq 0, \sum_i p_i = 1\}$ . A very popular distance function between such probability distributions is the *Kullback-Leibler divergence*:

$$K(p, q) = \sum_{i=1}^m p_i \log \frac{p_i}{q_i}.$$