Loss, Empirical Risk

This problem continues a previous activity. Consider a classification problem where we aim to classify new data points x into one of $c = 5$ classes denoted by $y \in \{A, B, C, D, F\}$.

- *Training* a classifier in machine learning can be thought of as finding a function $f(x)$ that a good job of assigning each data x point to its correct class. This is generally the challenging task in machine learning.
- *Testing* simply refers to applying the function $f(\cdot)$ to a new data point x.

You have a classifier that predicts the final letter grade for this course based on the number of hours studied for the first exam, denoted x:

$$f(x) = \begin{cases} x \geq 10 & A \\ 10 > x \geq 8 & B \\ 8 > x \geq 6 & C \\ 6 > x \geq 4 & D \\ 4 > x \geq 0 & F. \end{cases} \tag{1}$$

Recall that training error or empirical risk is defined as

$$\frac{1}{n} \sum_i \ell(f(x_i), y_i)$$

where $\ell(\cdot)$ is the *loss* function. You have a labeled training dataset from last year, when only 6 students took the course:

$$\mathscr{D} = \{(x_i, y_i)\}_{i=1}^{6} = \{(25.2, B), (2.1, D), (4.7, D), (12.0, A), (8.1, B), (4.9, C)\}$$

The misclassification loss (or 0/1 loss) is defined as $\ell(f(x), y) = \mathbb{1}_{f(x) \neq y}$ where

$$\mathbb{1}_{f(x) \neq y} = \begin{cases} f(x) \neq y & 1 \\ \text{else} & 0 \end{cases}$$

is the indicator function.

1. What is the empirical risk on $\mathscr{D}$ when $\ell(\cdot)$ is the misclassification loss? You may find it helpful to use a table:

Input x	25.2	2.1	4.7	12.0	8.1	4.9
Label y	B	D	D	A	B	C
Classifier Output $f(x)$	A	F	D	A	B	D
Loss $\ell(f(x), y)$	1	1	0	0	0	1

empirical risk = $\frac{1}{6}$ (1+1+0+0+0+1)

= 0.5

2. A *confusion matrix* C is a matrix with entries $C_{i,j}$ defined as the number of times $f(x) = j$ when $y = i$ on the test set. Find the confusion matrix for $f(\cdot)$ with the dataset above.

3. Instead of using letter grades, you decide to use the integers $1,2,3,4,5$, where A maps to 5, B maps to 4, and so on. What is the empirical risk when $\ell(\cdot) = (f(x) - y)^2$, i.e, the squared error? Make a new table with new labels and use the new loss function:

Input x	25.2	2.1	4.7	12.0	8.1	4.9
Label y	4	2	2	5	4	3
Classifier Output $f(x)$	5	1	2	5	4	2
Loss $\ell(f(x), y)$	1	1	0	0	0	1

empirical risk $= \frac{1}{6}(1+1+0+0+0+1) = 0.5$

4. If you change the mapping from letter grades to integers so that A maps to 1, and F maps to 5 (but no other changes). Does this change the squared error? Which is a more reasonable mapping and why? Yes. A to 5 is more reasonable because in the case A=1, A to D is closer than F to D

5. You use the classifier $f(x)$ in eq. (1) above for a new group of students. Without any new information, what percentage do you expect to correctly predict? How does this relate to the empirical risk when using misclassification loss? 50%, given empirical risk of 0.5

6. *Empirical risk minimization* is a technique for designing a classifier. If $\mathscr{F}$ is a set of candidate functions, the empirical risk minimizer is:

$$f = \arg\min_{f \in \mathscr{F}} \sum_{i=1}^{n} \ell(f(x_i), y_i)$$

which, in words, is the function that minimizes the average loss among the candidate functions (on the training data). Let $\mathscr{F}$ be functions of the form

$$f(x) = \begin{cases} x \geq t_a & A \\ t_a > x \geq t_b & B \\ t_b > x \geq t_c & C \\ t_c > x \geq t_d & D \\ t_d > x \geq 0 & F \end{cases}$$

with $t_a \geq t_b \geq t_c \geq t_d > 0$. Does the classifier specified at the start of this problem minimize empirical risk among $\mathscr{F}$ (on $\mathscr{D}$)? Use the 0/1 loss. If not, propose a new classifier that minimizes empirical risk among candidate functions $\mathscr{F}$.

↳ $t_a = 10$, $t_b = 8$, $t_c = 4$, $t_d = 2$ } maximizes empirical risk