Homework Assignment 4b

Due: Friday, April 10, 2024, 11:59 p.m.

Total marks: 15

Question 1. [15 MARKS]

In your assignment 4a coding implementation you measured the accuracy of your classifier using the 0-1 cost. We can write this cost as $1(\hat{y} \neq y)$ for prediction $\hat{y}$ and observed target y. The generalization error (the expected cost) for your binary classifier $f(\mathbf{x})$, across all pairs $(\mathbf{x}, y)$ is

$$\mathrm{GE}(f) \doteq \mathbb{E}[1(f(\boldsymbol{X}) \neq Y)] \tag{1}$$

where $\boldsymbol{X}, Y$ are the random variables with instances $\mathbf{x}, y$ drawn from joint distribution $p(\mathbf{x}, y)$.

(a) [5 MARKS] Assume you are given $\mathcal{D}_{\text{test}} = \{(\tilde{\mathbf{x}}_i, \tilde{y}_i)\}_{i=1}^m$, where we use the tilde notation above these variables to distinguish them from the pairs used in the training set. Write the formula to estimate the $\mathrm{GE}(f)$ using a sample average on $\mathcal{D}_{\text{test}}$.

(b) [5 MARKS] When we talked about squared costs and GE, we found that the GE decomposed into reducible error and irreducible error. We have a similar decomposition for the 0-1 cost for classification, though instead of equality we only have an upper bound

$$\mathbb{E}[1(f(\boldsymbol{X}) \neq Y)] \leq \underbrace{\mathbb{E}[1(f(\boldsymbol{X}) \neq f^*(\boldsymbol{X}))]}_{\text{reducible error}} + \underbrace{\mathbb{E}[1(f^*(\boldsymbol{X}) \neq Y)]}_{\text{irreducible error}} \tag{2}$$

where $f^*(\mathbf{x}) = \arg\max_{y \in \{0,1\}} p(y|\mathbf{x})$ is the optimal predictor that uses the true probabilities $p(y|\mathbf{x})$ (not estimated ones). Imagine you have a huge dataset of billions of samples, and you learn f_1 with logistic regression and f_4 with polynomial logistic regression with $p = 4$. Do you think f_1 or f_4 will have lower reducible error? Explain your answer in a few sentences.

(c) [5 MARKS] Give an example in classification to explain the irreducible error. Make sure your example highlights why the irreducible error is non-zero. Be specific in your example, with a concrete example of targets and the features in $\mathbf{x}$.

Homework policies:

Your assignment should be submitted as one pdf document on eClass. The pdf must be written legibly and scanned or must be typed (e.g., Latex). This .pdf should be named Firstname_LastName_Sol.pdf,

Because assignments are more for learning, and less for evaluation, grading will be based on coarse bins. **The grading is atypical**. For grades between (1) 81-100, we round-up to 100; (2) 61-80, we round-up to 80; (3) 41-60, we round-up to 60; and (4) **0-40, we round down to 0**. The last bin is to discourage quickly throwing together some answers to get some marks. The goal for the assignments is to help you learn the material, and completing less than 50% of the assignment is ineffective for learning.

We will not accept late assignments. There is no late penalty policy. The assignments must be submitted electronically via eClass on time, by 11:59 pm Mountain time on the due date. There is a grace period of 48 hours when assignments will be accepted. No submissions will be accepted after 48 hours after the deadline, and the assignment will be considered as incomplete if not submitted.

All assignments are individual. All the sources used for the problem solution must be acknowledged, e.g. web sites, books, research papers, personal communication with people, etc. Academic honesty is taken seriously; for detailed information see the University of Alberta Code of Student Behaviour.

Good luck!