Educational Data Mining and Applciations: HW#3

By J. H. Wang

Nov. 7, 2023

Homework #3: Classification

- Chap.8:
  - 8.11
  - 8.12
  - 8.16
- Chap.9:
  - 9.4
  - 9.5
- Due: 2 weeks (Nov. 21, 2023)

- 8.11 The harmonic mean is one of several kinds of averages. Chapter 2 discussed how to compute the arithmetic mean, which is what most people typically think of when they compute an average. The harmonic mean, H, of the positive real numbers, x1,x2, …,xn, is defined as:

$$H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

- The *F* measure is the harmonic mean of precision and recall. Use this fact to derive Eq. (8.28) for *F*. In addition, write *F* as a function of true positives, false negatives, and false positives.

Exercises for Chap.8

- 8.12: The data tuples of Fig. 8.25 are sorted by decreasing probability value, as returned by a classifier. For each tuple, compute the values for the number of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN).
  Compute the true positive rate (TPR), and false positive rate (FPR).
  Plot the ROC curve for the data.

[Hint: You should set a number of thresholds t for classifying the probability values p into positive (when p>=t) or negative classes (when p<t) to plot the ROC curve. (for example, t=0, 0.1, 0.2, …, 0.9)]

[… to be continued]

Figure 8.25 Tuples sorted by decreasing score, where the score is the value returned by a probabilistic classifier

Tuple #	Class	Probability
1	P	0.95
2	N	0.85
3	P	0.78
4	P	0.66
5	N	0.60
6	P	0.55
7	N	0.53
8	N	0.52
9	N	0.51
10	P	0.40

- 8.16: Outline methods for addressing the *class imbalance problem*. Suppose a bank wants to develop a classifier that guards against fraudulent credit card transactions. Illustrate how you can induce a quality classifier based on a large set of nonfraudulent examples and a very small set of fraudulent cases.

Exercises for Chap.9

- 9.4: Compare the advantages and disadvantages of *eager classification* (e.g., decision tree, Bayesian, neural network) versus *lazy classification* (e.g., *k*-nearest neighbor).
- 9.5: Write an algorithm for *k-nearest-neighbor classification* given k, the nearest number of neighbors, and n, the number of attributes describing each tuple.

Homework Submission

- For hand-written exercises, please hand in your homework in class (paper version)
  - Remember to write your student ID

- For programming projects, please submit a compressed file to iSchool+
  - It should contain your source codes, sample input and generated output, and documentation on how to compile, install, or configure the environment

Thanks for Your Attention!