Evaluation of Machine Learning Models

Shrey Gupta

Applied Machine Learning (HOUSECS 59-01), Duke University

September 19, 2018

Evaluation

- How can we evaluate the performance of the models we create?
- Various performance measures for regression, classification, and clustering.
  - Depending on various “goals” and “priorities”, different measures used.

Regression

- Metrics: mean absolute error, mean squared error, and R^2 value.

Mean Absolute Error

- —

Mean Squared Error

- —

R^2 Value

- —

Classification

- Metrics: misclassification error (and classification accuracy), precision, recall, F1-score, confusion matrices, and ROC curves (and associated AUC).

Misclassification Error

$$error = 1 - accuracy$$

$$accuracy = \frac{true\ positives + true\ negatives}{positives + negatives}$$

- Most commonly used metric for classification.
- Insightful when the number of positive points is approximately equal to the number of negative points.

Precision and Recall

$$precision = \frac{true\ positives}{predicted\ positives}$$

$$recall = \frac{true\ positives}{positives}$$

- Precision: "fraction of relevant instances among the retrieved instances" (Wikipedia).
- Recall: "fraction of relevant instances that have been retrieved over the total amount of relevant instances" (Wikipedia).

F1-score

$$score = 2\left(\frac{precision \times recall}{precision + recall}\right)$$

- Most applications require a balance between precision and recall (as there is a trade-off between the two).

Confusion Matrices

	y = +1	y = -1
$\hat{y} = +1$	*true positives* (TP)	*false positives* (FP)
$\hat{y} = -1$	*false negatives* (FN)	*true negatives* (TN)

- Provides a concise presentation of the predictive power of a model.

ROC Curves

- —

Clustering

- Metrics: purity, Rand measure, and F1-score.

Purity

- —

Rand Measure

$$measure = \frac{true\ \ positives + true\ \ negatives}{positives + negatives}$$

- Similar to accuracy measure for classification, and requires labeled points.

F1-score

$$score = 2\left(\frac{precision \times recall}{precision + recall}\right)$$

- Precision and recall are calculated with labeled points, similar to classification.
- Most applications require a balance between precision and recall (as there is a trade-off between the two).

Training, Testing, and Validation

- —

Cross-validation

- —