Data Mining Cheat Sheet

by sherly2zd via cheatography.com/209490/cs/45048/

Evaluating model performance

The "Training Set-Validation Set-Test Set" Approach

The "Training Set-Validation Set-Test Set" Approach:

Useful for selecting one of several models *and obtaining an estimate of the resulting performance* (model *assessment*):

1. Split the available data into a training set, a validation set and a test set:
   - depending on the amount of available data and the number of models to be compared, 50:25:25 or 60:20:20.
2. Fit each model separately on the training set.
3. Evaluate each model separately on the validation set.
4. Choose the model that performs best on the validation set.
5. Estimate the performance of that model on the test set.

At the end, train the selected model again using all data!!!

Performance measures for regression

Performance Measures for Regression Problems

Let i-th **validation error** be $e_i = Y_i - \hat{f}(X_{i1}, \ldots, X_{ip}),\ i = 1, \ldots, n$:

1. **mean absolute error:** $\frac{1}{n}\sum_i |e_i|$
2. **average error:** $\frac{1}{n}\sum_i e_i$
3. **mean absolute percentage error:** $100\% \cdot \frac{1}{n}\sum_i \left|\frac{e_i}{y_i}\right|$
4. **root-mean-squared error:** $\sqrt{\frac{1}{n}\sum_i e_i^2}$
5. **total sum of squared errors:** $\sum_i e_i^2$

Benchmark: The "average predictor" $\hat{f}(X_{i1}, \ldots, X_{ip}) = \overline{y}$, where $\overline{y}$ is the average output over the training set.

Performance measures for classification

Performance Measures for Classification Problems

Consider the following confusion matrix:

		Predicted Class	
		"yes"	"no"
Actual Class	"yes"	n_{11}	n_{12}
	"no"	n_{21}	n_{22}

1. **estimation misclassification rate (= total error rate):** $\frac{n_{12} + n_{21}}{n_{11} + n_{12} + n_{21} + n_{22}}$
2. **accuracy:** 1 - estimation misclassification rate
3. **sensitivity:** $\frac{n_{11}}{n_{11} + n_{12}}$
4. **specificity:** $\frac{n_{22}}{n_{21} + n_{22}}$

(3 and 4: ***if "yes" is the important class***)

Benchmark: The "majority predictor" (majority class in training data)