

Instructions for WEKA:

To create training, validation, and test set:

- Filters > unsupervised > instance > Resample
 - In resample filter:
 - Training Set:
 - InvertSelection: False
 - noReplacement: True
 - randomSeed: 1
 - sampleSizePercent: 70 (80%)
- Save As: “.arrf” file.
- Click “Undo”
- Go back to resample filter:
 - InvertSelection: True
 - Select OK
- Click “Apply”
 - This gives us the 20% that’s not in our training set

Learning Curve in Experimenter:

Weka:

- Experiment > Select Advanced Experiment
- Destination: Choose any destination (Ex. Instances) click and select Output File
- Result Generator: “CrossValidationResultProducer”
 - splitEvaluator: “ClassifierSplitEvaluator”
 - Left Click on the properties for splitEvaluator:
 - Select classifier: meta > Filtered Classifier
 - Because we’re trying to systematically increase the amount of data fed to the classifier.
 - Left click properties on Filtered Classifier:
 - Choose classifier: “”
 - Filter: unsupervised > instance > remove percentage.
- Change the runs:
 - From: To:
- Enable Generator Properties Dropdown:
 - splitEvaluator > classifier > filter > percentage > Select
 - Add percentages from 10 up to 90.
- Datasets > Add new... > Add Training Dataset
- Run Tab > Start
- Analyze Tab > Click Experiment Button
- Analyze Tab:
 - Configure Test:
 - Comparison field: Select dropdown “Percent_incorrect”

- Click “Perform test”
- % is the amount taken out of the data
- Error rate is complement of accuracy

To get data points for the learning curves in knowledge flow:

In Weka KnowledgeFlow:

- ArffLoader to load training data
- ClassValue Picker > TestSetMaker, TrainSetMaker, and CrossValidationFoldMaker
- TestSetMaker & TrainSetMaker > 5 classifiers
 - For Test = Train: To get Training Error
- CrossValidationFoldMaker > 5 classifiers
 - For CV: To get Cross-Validation Error
- 5 classifiers > ClassifierPerformanceEvaluators > TextViewer
 - To get the error rates
- Enter info into Excel and Plot

Model Evaluation:

Experimenter:

- Set Advanced
- Click New
- Choose Destination > Output file location for results
- Result Generator
 - For CV error:
 - CrossValidationResultProducer
 - Left click splitEvaluator
 - Choose meta > FilteredClassifier
 - Left click and choose classifier of interest
 - OK out
 - For Train error:
 - RandomSplitResultProducer
 - Left click:
 - trainPercent = 100.00
 - Left click splitEvaluator
 - Choose CFilteredClassifier
 - Left click and choose classifier of interest
 - OK out
 - Generator Properties > Selected Enabled
 - Expand splitEvaluator > classifier > classifier > Parameters are shown
 - The parameters belong to the classifier you selected
 - Add different parameters in the field

Runs:

- Indicate cross validation folds

Add dataset:

- Select Add new...
- Choose path of dataset

See Results:

- Go to run
- Analyse
 - Click Experiment
 - Select: Comparison field
 - Select Percent_incorrect or Percent_correct
 - Click "Perform Test"

Excel:

- Use output data from Weka to generate line charts for learning curves