Data Mining Evaluation

Simple Evaluation Process

Process:

- Have two large datasets – ‘training’ data set and a ‘testing’ dataset
- Run a wide variety of algorithms on the training dataset
- Test each model on the test dataset to see which performs best

More Usual Evaluation Process

In many cases 3 data sets are needed:

- Training data set for selecting the learning algorithm
- Validation data set for setting parameters on the chosen learning algorithms
- Testing data set for determining the accuracy

Independent and Representative Data

As long as the training and testing datasets

- are independent and
- representative of the underlying problem

the performance predicted on the test dataset will match reality.

Recommendation

The training set should be balanced with equal numbers of each of the outcomes. A smaller, balanced sample is preferable to a larger one with a very low proportion of rare outcomes

Electrical Load Forecasting

Want to predict future demand for power as far in advance as possible

With accurate predictions can fine tune:

- operating reserves
- maintenance scheduling, and
- fuel inventory management

Data collected over 15 years

Major holidays, such as Thanksgiving, Christmas and New year's Day show significant variation from normal loads

From "Data Mining: Practical Machine Learning Tools and Techniques" by Witten, Frank, Hall & Pal

Electrical Load Forecasting

Data went back 15 years but:

- Only 15 Christmas and Thanksgiving Days
- Only 4 Feb. 29th and presidential elections

Diagnosis of Electromechanical Failures

Preventative maintenance of electromechanical devices such as motors and generators can forestall failures that disrupt industrial processes

Had data with 600 faults, each comprising a set of measurements along with an expert's diagnosis, representing 20 years of experience

Half were unsatisfactory for various reasons and had to be discarded, the remainder used for training examples

From "Data Mining: Practical Machine Learning Tools and Techniques" by Witten, Frank, Hall & Pal

Rare Class Values

Target variables might represent something relatively rare:

- Prospects responding to a direct mail offer
- Credit card holders committing fraud
- In a month, newspaper subscribers canceling their subscription

Maximizing Training

Once error rate is measured, re-bundle all three datasets and train again – but don't re-measure the error rate!

Resubstitution Error Rate

Resubstitution error rate – error rate resulting from testing on the training data

- This error rate will be highly optimistic
- Not a good indicator of what the performance will be on an independent test dataset