Week 1 Notes: Practical Machine Learning

Process:

question -> input data -> features -> algorithm -> parameters -> evaluation

Features matter!

Properties of good features

- Lead to data compression
- Retain relevant information
- Are created based on expert application knowledge

Common mistakes

- Trying to automate feature selection
- Not paying attention to data-specific quirks
- Throwing away information unnecessarily

Algorithms matter less than you'd think

Issues to consider:

Interpretable

Simple

Accurate

Fast

Scalable

In sample and out sample error

In Sample Error: The error rate you get on the same data set you used to build your predictor. Sometimes called resubstitution error.

Out of Sample Error: The error rate you get on a new data set. Sometimes called generalization error.

Key ideas

1. Out of sample error is what you care about
2. In sample error < out of sample error
3. The reason is overfitting

- Matching your algorithm to the data you have

- Data have two parts

- Signal

- Noise

- The goal of a predictor is to find signal

- You can always design a perfect in-sample predictor

- You capture both signal + noise when you do that

- Predictor won't perform as well on new samples

Prediction Study Design

1. Define your error rate
2. Split data into:

- Training, Testing, Validation (optional)

3. On the training set pick features

- Use cross-validation

4. On the training set pick prediction function

- Use cross-validation

6. If no validation

- Apply 1x to test set

7. If validation

- Apply to test set and refine
- Apply 1x to validation

http://www2.research.att.com/~volinsky/papers/ASAStatComp.pdf

Rules of thumb for prediction study design

- If you have a large sample size
- 60% training
- 20% test
- 20% validation
- If you have a medium sample size
- 60% training
- 40% testing
- If you have a small sample size
- Do cross validation
- Report caveat of small sample size

Basic terms

In general, **Positive** = identified and **negative** = rejected. Therefore:

True positive = correctly identified

False positive = incorrectly identified

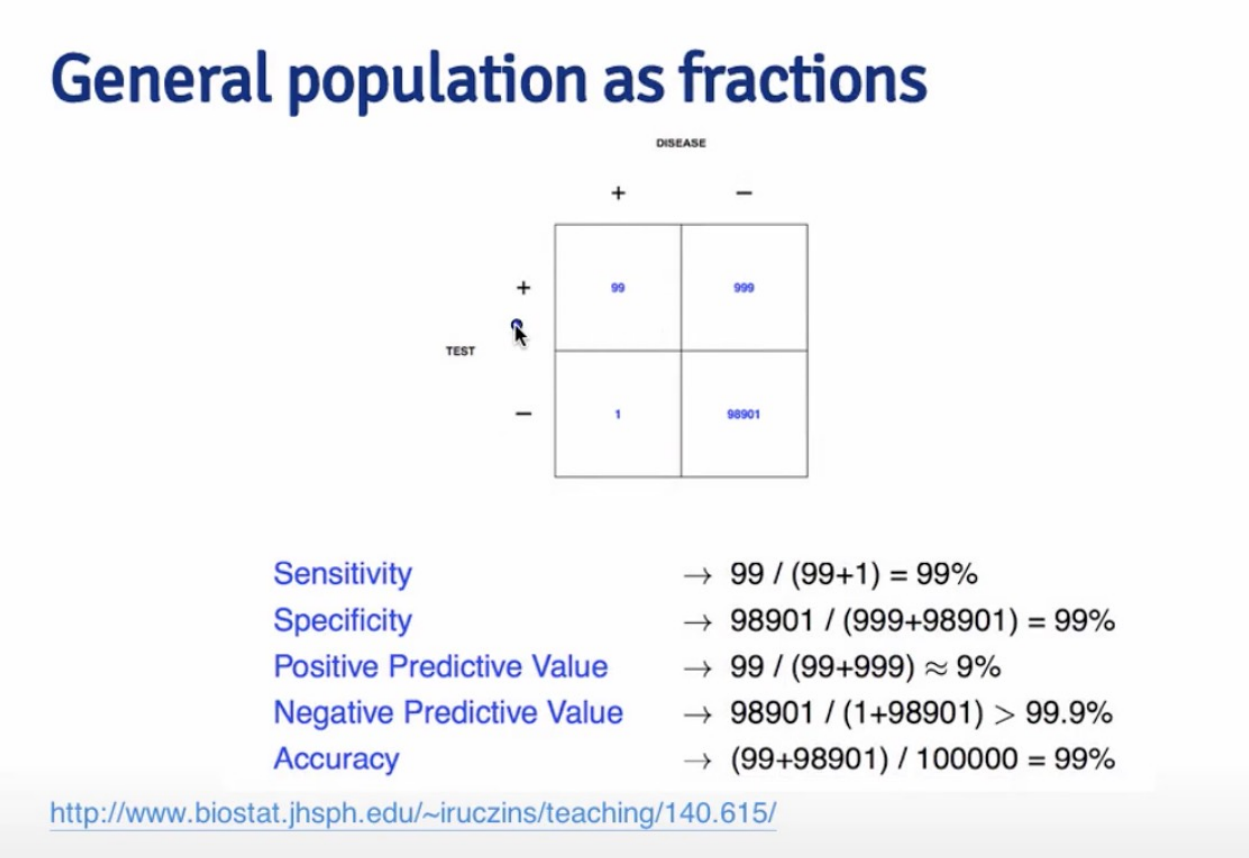
True negative = correctly rejected

False negative = incorrectly rejected

** Summary **

For continuous data

Mean squared error (MSE):



General population as fractions

		DISEASE	
		+	−
TEST	+	99	999
	−	1	98901

Sensitivity	→ 99 / (99+1) = 99%
Specificity	→ 98901 / (999+98901) = 99%
Positive Predictive Value	→ 99 / (99+999) ≈ 9%
Negative Predictive Value	→ 98901 / (1+98901) > 99.9%
Accuracy	→ (99+98901) / 100000 = 99%

http://www.biostat.jhsph.edu/~iruczins/teaching/140.615/

Figure 1:

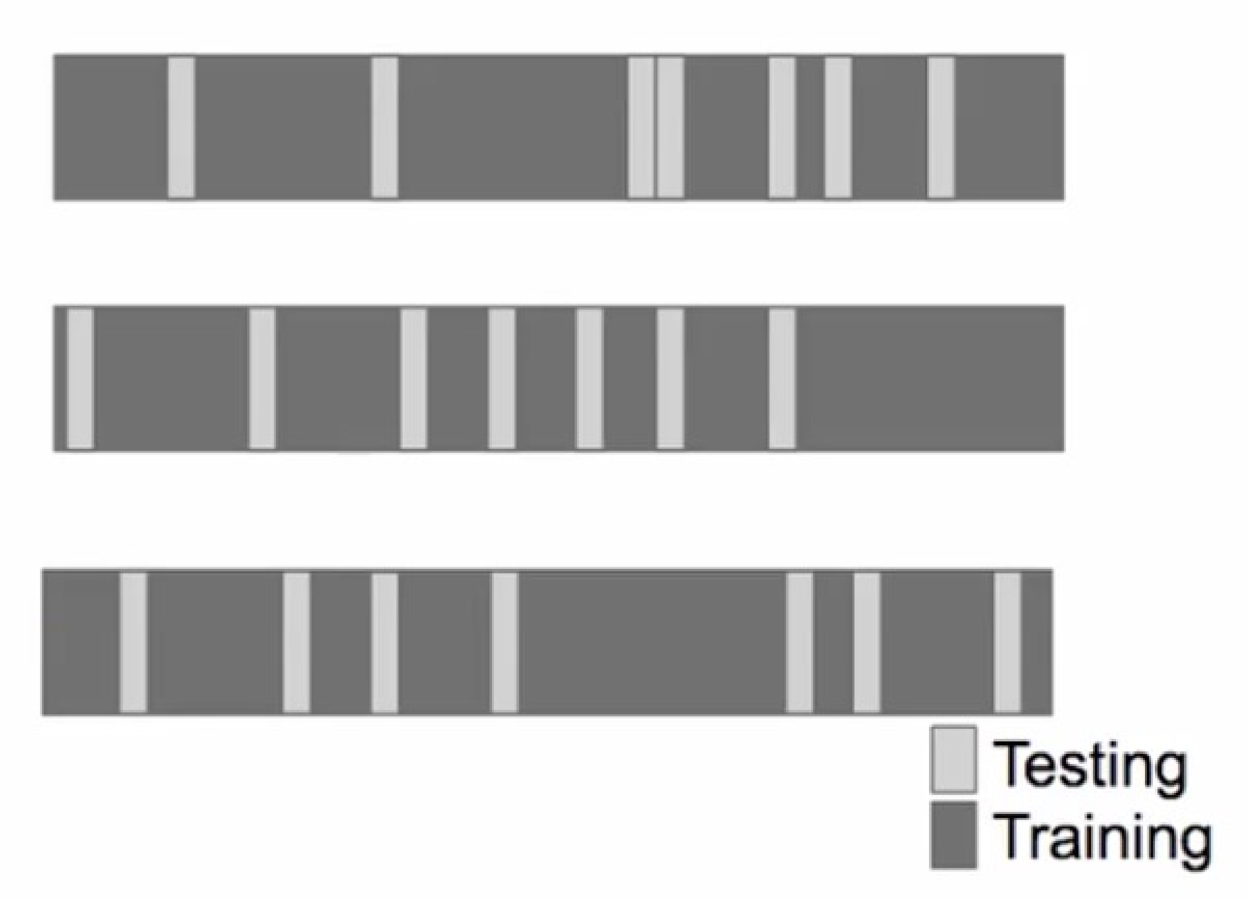


Figure 2:

$$\frac{1}{n}\sum_{i=1}^{n}(Prediction_i - Truth_i)^2$$

Root mean squared error (RMSE):

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Prediction_i - Truth_i)^2}$$

Cross Validation:

Key idea

1. Accuracy on the training set (resubstitution accuracy) is optimistic
2. A better estimate comes from an independent set (test set accuracy)
3. But we can't use the test set when building the model or it becomes part of the training set
4. So we estimate the test set accuracy with the training set.

***Approach*:**

1. Use the training set

2. Split it into training/test sets

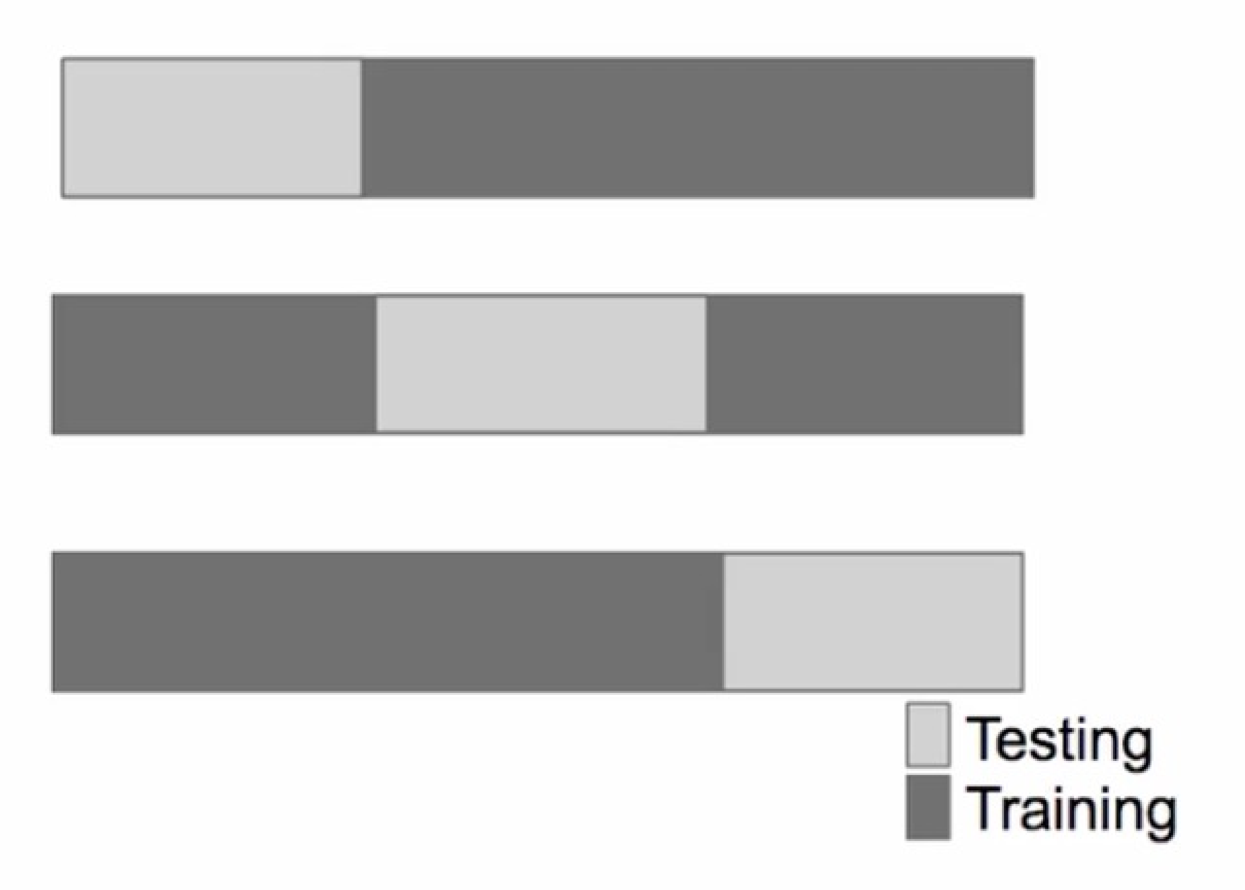


Figure 3:

3. Build a model on the training set
4. Evaluate on the test set
5. Repeat and average the estimated errors

***Used for*:**

1. Picking variables to include in a model
2. Picking the type of prediction function to use
3. Picking the parameters in the prediction function
4. Comparing different predictors

K-Fold

Considerations

- For time series data, data must be used in "chunks"
- For k-fold cross validation
- Larger k = less bias, more variance

- Smaller k = more bias, less variance
- Random sampling must be done *without replacement*
- Random sampling with replacement is the *bootstrap*
- Underestimates of the error
- Can be corrected, but it is complicated (0.632 Bootstrap)
- If you cross-validate to pick predictors estimate you must estimate errors on independent data.