Changelog: 14 Oct 2016, 30 Oct 2106

Basic Concepts

Induction, Generalization, Evaluation

Lecture 2: Part 1

Marina Santini

Acknowledgements
Pictures retrieved from the web

Lecture 2: Required Reading

- Handout
- Daumé III (2015: 8-10; 19-24)
- Witten et al. (2011): Ch 2

Outline

- Induction
- Generalization
- Splitting the Data
- Evaluation Measures
- Modelling
- Parameters
- Inductive Bias

Supervised Learning

- also called Inductive Learning
- Inductive learning is the general theory behind supervised learning

What is inductive learning?

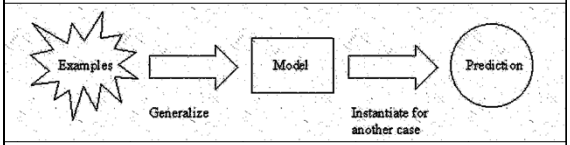
- input (x)
- output (y)
- estimate the function (f)

More specifically:
"the problem is to generalize from the samples and the mapping to be useful to estimate the output for new samples in the future. In practice it is almost always too hard to estimate the function, so we are looking for very good approximations of the function."

Jason Browlee
ML Mastery Blog

Induction

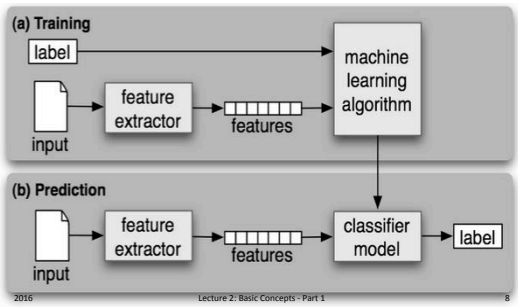
Induction is the process of reaching a general conclusion from specific examples



Inductive Machine Learning (Simplified)

- The goal of inductive machine learning is to take some training data and use it to induce a model.
- This model will be evaluated on the test data
- The machine learning algorithm has succeeded if its performance on the test data is high.

Induction pipeline (supervised classification)



Example: Predict the class of an unseen iris flower

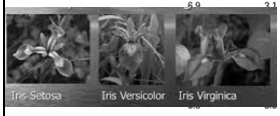
Sepal length –Sepal width –Petal length –Petal width -Type
5.2 3.7 1.7 0.3 ???

Sepal length (cm)	Sepal width (cm)	Petal length (cm)	Petal width (cm)	Type
5.1	3.5	1.4	0.2	*Iris setosa*
4.9	3.0	1.4	0.2	*Iris setosa*
4.7	3.2	1.3	0.2	*Iris setosa*
4.6	3.1	1.5	0.2	*Iris setosa*
5.0	3.6	1.4	0.2	*Iris setosa*
7.0	3.2	4.7	1.4	*Iris versicolor*
6.4	3.2	4.5	1.5	*Iris versicolor*
6.9	3.1	4.9	1.5	*Iris versicolor*
5.5	2.3	4.0	1.3	*Iris versicolor*
6.5	2.8	4.6	1.5	*Iris versicolor*
6.3	3.3	6.0	2.5	*Iris virginica*
5.8	2.7	5.1	1.9	*Iris virginica*
7.1	3.0	5.9	2.1	*Iris virginica*
6.3	2.9	5.6	1.8	*Iris virginica*
6.5	3.0	5.8	2.2	*Iris virginica*

Require us to ***generalize*** from the training data

Data: The iris dataset

Sepal length (cm)	Sepal width (cm)	Petal length (cm)	Petal width (cm)	Type
5.1	3.5	1.4	0.2	*Iris setosa*
4.9	3.0	1.4	0.2	*Iris setosa*
4.7	3.2	1.3	0.2	*Iris setosa*
4.6	3.1	1.5	0.2	*Iris setosa*
5.0	3.6	1.4	0.2	*Iris setosa*
7.0	3.2	4.7	1.4	*Iris versicolor*
6.4	3.2	4.5	1.5	*Iris versicolor*
6.9	3.1	4.9	1.5	*Iris versicolor*
5.5	2.3	4.0	1.3	*Iris versicolor*
6.5	2.8	4.6	1.5	*Iris versicolor*
6.3	3.3	6.0	2.5	*Iris virginica*
5.8	2.7	5.1	1.9	*Iris virginica*
7.1	3.0	5.9	2.1	*Iris virginica*
6.3	2.9	5.6	1.8	*Iris virginica*
6.5	3.0	5.8	2.2	*Iris virginica*



Three components:
1. Class label (aka "label", denoted y)
2. Features (aka "attributes")
3. Feature values (aka "attribute values", denoted x) ⇒ Features can be binary, nominal or continuous

- **A labeled dataset is a collection of (x,y) pairs**

Task

- Predict the class for this "test" example:

Sepal length – Sepal width – Petal length – Petal width - Type

5.2 3.7 1.7 0.3 ???

Sepal length (cm)	Sepal width (cm)	Petal length (cm)	Petal width (cm)	Type
5.1	3.5	1.4	0.2	*Iris setosa*
4.9	3.0	1.4	0.2	*Iris setosa*
4.7	3.2	1.3	0.2	*Iris setosa*
4.6	3.1	1.5	0.2	*Iris setosa*
5.0	3.6	1.4	0.2	*Iris setosa*
7.0	3.2	4.7	1.4	*Iris versicolor*
6.4	3.2	4.5	1.5	*Iris versicolor*
6.9	3.1	4.9	1.5	*Iris versicolor*
5.5	2.3	4.0	1.3	*Iris versicolor*
6.5	2.8	4.6	1.5	*Iris versicolor*
6.3	3.3	6.0	2.5	*Iris virginica*
5.8	2.7	5.1	1.9	*Iris virginica*
7.1	3.0	5.9	2.1	*Iris virginica*
6.3	2.9	5.6	1.8	*Iris virginica*
6.5	3.0	5.8	2.2	*Iris virginica*

Require us to ***generalize*** from the training data

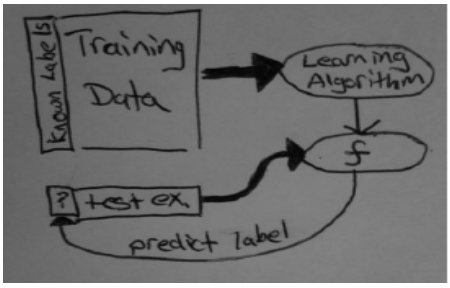
Generalization

- Machine learning is all about finding patterns in data.
- The most central concept in machine learning is **generalization**: how to generalize beyond the examples that have been provided at "training time" to new examples that you see at "test time."

Generalization

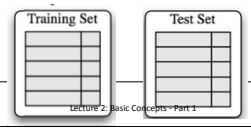
- Learning is generalization not memorization
- Generalization in the context of ML is the ability of a learning system to perform accurately on new/unseen data after having learned from existing examples.

Ex:



Splitting the data (1)

- Training data & Test Data:
  - Common splits in LT: 70/30; 80/20; 90/10
- **TEST DATA MUST BELONG TO THE SAME STATISTICAL DISTRIBUTION AS THE TRAINING DATA**



Face Recognition

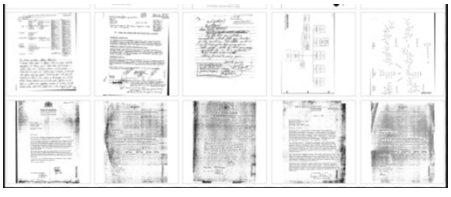
Training examples of a person

Test images

ORL dataset,
AT&T Laboratories, Cambridge UK

Example: Letter vs non-letter classification

Training set



Test set

Cross-Validation: 10-fold

- In 10-fold cross-validation you break you training data up into 10 equally-sized partitions.
- You train a learning algorithm on 9 of them and test it on the remaining 1.
- You do this 10 times, each holding out a different partition as the test data.
- Typical choices for n-fold are 2, 5, 10.
- 10-fold cross validation is the most common.

Leave One Out

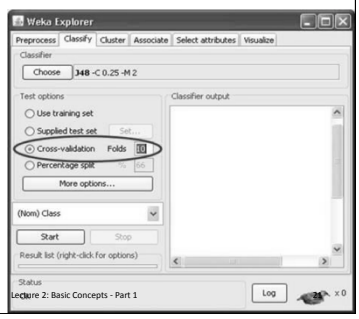
- Leave One Out (or LOO) is a simple cross-validation. Each learning set is created by taking all the samples except one, the test set being the sample left out.

Stratification

- Proportion of each class in the traning set and test sets is the same as the proportion in the original sample.

Weka Evaluation Methods

- Ex: 10-fold cross-validation



True and False Positives and Negatives

- **True positives** are correct labels that are correctly identified.
- **True negatives** are incorrect lables that are correctly identified as incorrect.
- **False positives** (or **Type I errors**) are incorrect labels that are incorrectly identified as correct.
- **False negatives** (or **Type II errors**) are correct labels that are incorrectly identified as incorrect.

Confusion Matrix

- Usually, the rows are the observed/actual class labels and the columns the predicted class labels.
- Each cell contains the number of predictions made by the classifier that fall into that cell.

Predicted

Actual	Positive	Negative
Positive	True Positive	False Positive
Negative	False Negative	True Negative

Truth Table Confusion Matrix

Multi-Class Confusion Matrix

If a classification system has been trained to distinguish between cats, dogs and rabbits, a confusion matrix will summarize the results:

Actual class	Predicted: Cat	Predicted: Dog	Predicted: Rabbit
Cat	5	3	0
Dog	2	3	1
Rabbit	0	2	11

True and False Positives and Negatives & evaluation measures

Given these four numbers , we can define the following measures:

- Accuracy
- Precision
- Recall
- F-Measure

Accuracy

Accuracy measures the percentage of correct results that a classifier has achieved.

Accuracy = (TP + TN)/(TP + TN + FP + FN)

Watch out: accuracy is a tricky metric:
it can be biassed towards the most frequent class

Precision, Recall, F-Measure in Classifiation

- **Precision** the number of correctly classified positive examples divided by the number of examples labeled by the system as positive P= *TP/(TP+FP)*
- **Recall** the number of correctly classified positive examples divided by the number of positive examples in the data. R= *TP/(TP+FN)*
- The **F-Measure** (or **F-Score**) a combination of the above.

Sokolova and Lapalme (2009)

Confusion matrix: Sokolova and Lapalme (2009)

"A sytematic analysis of performance measures for classification tasks"

Data class	Classified as *pos*	Classified as *neg*
pos	true positive (*tp*)	false negative (*fn*)
neg	false positive (*fp*)	true negative (*tn*)

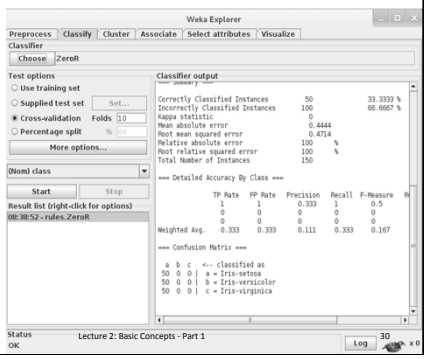
$$\begin{bmatrix} tp & fn \\ fp & tn \end{bmatrix}$$

Summary: Accuracy, Precision, Recall, F-measure

- Accuracy = (TP + TN)/(TP + TN + FP + FN)
- Precision = TP / TP + FP
- Recall = TP / TP + FN
- F-measure (harmonic) =
  2*((precision*recall)/(precision+recall))

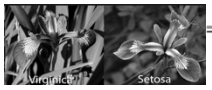
Weka: Output

Example of a classifier output

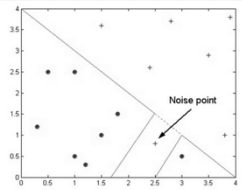


Disturbing factors: Noise

- Unexplained or random variation in the data
- Anomaly



Overfitting due to Noise



Underfitting & Overfitting

- Underfitting: the model has not learned enough from the data and is unable to generalize.
- Overfitting: the model has learned too many idiosyncrasies (noise) and is unable to generalize.

Generalization: Overfitting & Underfitting

- Overfitting occurs when the model fits the training data too well and does not generalize so it performs badly on the test data. Overfitting is often a result of an excessively complicated model.
- Underfitting occurs when the model does **not** fit the data well enough. Underfitting is often a result of an excessively simple model.
- Both overfitting and underfitting lead to **poor predictions** on unseen examples.
- A model that overfits or underfits does not generalize well.

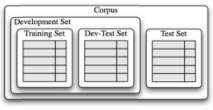
Parameters

Models can have many parameters and finding the best combination of parameters is not trivial.

Hyperparameters

A hyperparameter is a parameter that controls other parameters of the model.

Ex: Testing parameters and hyperpameters via developent set



- Split your data into 70% training data, 10% development data and 20% test data.
- For each possible setting of the hyperparameters:
  - Train a model using that setting on the training data
  - Compute the model error rate on the development data
  - From the above collection of models, choose the one that achieve the lowest error rate on development data.
  - Evaluate that model on the test data to estimate future test performance.

Inductive Bias: Definition

"The inductive bias of a learning algorithm is the set of assumptions that the learner uses to predict outputs given inputs that it has not encountered".

Tom Mitchell, 1980

Inductive Bias

Each learning method has a different inductive bias.

The inductive bias affects the performance of a classifier.

Not everything is learnable

- Noise at feature level
- Noise at class label level
- Features are insufficient
- Class labels are controversial
- Inductive bias not appropriate for the kind of problem we try to learn

Key Point

- Our goal when we choose a machine learning model is the following:
  - the model should perform well on future, unseen data.

Quiz 1: Stratification

What does it mean "stratified" cross-validation?

1. The examples of a class are all in the training set, and the rest of the classes are in the test set.
2. The proportion of each class in the sets ae the same as the proportion in the original sample
3. None of the above.

Quiz 2: Data Splits

Which are recommended splits between training and test data?

1. 80/20
2. 50/50
3. 10/90

Quiz 3: Overfitting

What does ”overfitting” mean?

1. The model has not learned enough from the data and is unable to generalize
2. The proportion of each class in the sets is the same as the proportion in the original sample
3. None of the above.

Quiz 4: Accuracy

Why is accuracy unreliable?

1. Because it can be biassed towards the most frequent class.
2. Because it always guesses wrong.
3. None of the above

The end