MITx | MicroMasters™
in Supply Chain Management

MIT Center for Transportation & Logistics

CTL.SC4x – Technology and Systems

Key Concepts Document

This document contains the Key Concepts for the SC4x course, week 6.

These are meant to complement, not replace, the lesson videos and slides. They are intended to be references for you to use going forward and are based on the assumption that you have learned the concepts and completed the practice problems.

The first draft was created by Dr. Alexis Bateman in the Winter of 2017.

This is a draft of the material, so please post any suggestions, corrections, or recommendations to the Discussion Forum under the topic thread "Key Concept Documents Improvements."

Thanks,

Chris Caplice, Eva Ponce and the SC4x Teaching Community
Winter 2017 v1

Introduction to Machine Learning

Summary

In this lesson we explore machine learning. This includes identifying when we need machine learning instead of other techniques such as regression. We break down the different classes of machine learning algorithms. In addition, we identify how to use machine-learning approaches to make inferences about new data.

Review of Regression

Linear regression uses the value of one or more variables to make a prediction about the value of an outcome variable. Input variables are called independent variables and the output variable is known as the dependent variable.

- Linear regression output includes coefficients for each independent variable.
    - This is a measure of how much an independent variable contributes to the prediction of the dependent variable.
    - The output also includes metrics to be able to assess how the model fits the data. The better fit of the model, the better you are able to make accurate predictions about new data.
- Using coefficients calculated from historic data, a regression model can be used to make predictions about the value of the outcome variable for new records.

Overview of Machine Learning Algorithms

Machine learning algorithms are primarily use to make predictions or learn about new, unlabeled data. There are several classes of algorithms:

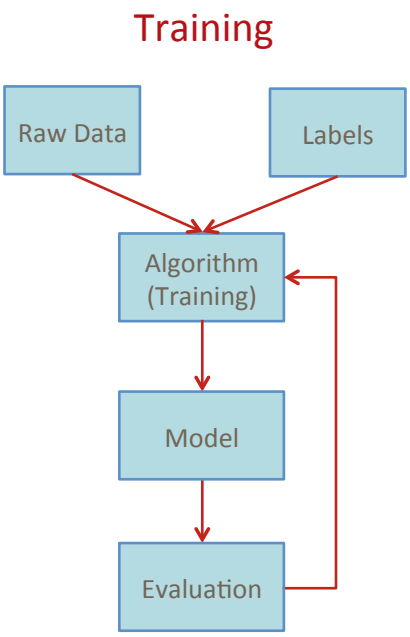
- **Classification**: assigning records to pre-defined discrete groups
- **Clustering**: splitting records into discrete groups based on similarity; groups are not known a priori
- **Regression**: predicting value of a continuous or discrete variable
- **Associate learning**: observing which values appear together frequently

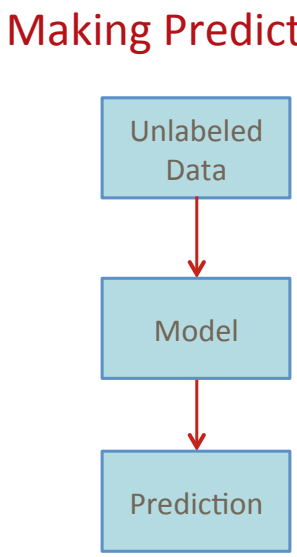
Supervised vs. Unsupervised Machine Learning

Supervised learning uses outcome variables, known as labels, for each record to identify patterns in the input variables or features related to the outcome variable.

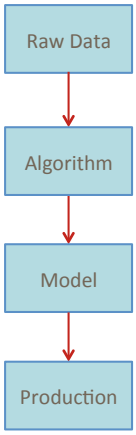
- Correct answer, label is known in the training data
- Label is applied by a person or already exists
- Labeled data are used to train an algorithm using feedback
- Apply or test the trained model on new, unseen data to predict the label

Learning Flow





Learning Flow



In **unsupervised learning**, the outcome variable values are unknown, therefore relationships among the input variables are used to identify patterns of clusters of records.

- Finds previously unknown patterns in the data without labels or guidance
- No training/testing/validating process because correct answer is unknown

Model Quality

Machine learning models should be trained on an unbiased set of data that is representative of the variance in the overall dataset. Bias quantifies the lack of ability of a model to capture underlying trend in the data. More complex models decrease bias but tend to increase variance. Variance quantified a model's sensitivity to small changes in the underlying dataset.

- Ideally want low bias and low variance, but there is a tradeoff between the two quantities
- If there is a bias in the training data or if too many features are included in a model, the model is at risk of being overfit. In overfit models, the coefficients, known as parameters, will not be generalizable enough to make good predictions for new records.
- A large and representative sample of the labeled data should be used to train the model, the remainder is used for testing.

Overfitting vs underfitting

- **Underfitting** – model is too simple, high bias and low variance
- **Overfitting** – model is too complex, low bias and high variance
    - Overfitting is a more common pitfall
    - Tempting to make models fit between by adding more features
    - Results in a model that is incapable of generalizing beyond the training data

Learning Objectives

- Be introduced to machine learning
- Become familiar with different types of machine learning algorithms
- Be able to differentiate supervised and unsupervised learning and their processes
- Recognize model quality and the tradeoffs between bias and variance
- Learn how to identify when a model is over or underfit

Machine Learning Algorithms

Summary

In this lesson we are going to dive deeper into machine learning algorithms. Each model has different properties and is best for different types of tasks. We review how to compare them with performance metrics. We need to be able to group records together without labels to inform prediction using unsupervised classification. In addition, we review capability to confidently reduce the number of features included in an analysis without losing information. The lesson also introduces how to compare predictor accuracy and test for sensitivity and specificity.

Dimensionality reduction

Dimensionality reduction is a term for reducing features included in analysis. It is often needed for analysis with many features. Trying to reduce dimensionality randomly or manually leads to poor results

- Results need to be interpreted by humans, should be tractable
- Increasing the number of features included increases the required sample size
- Features should not be included or discarded from analysis based on instinct
    - Dimensionality reduction techniques should be employed, such as principal component analysis.
- Summary statistics are a means of dimensionality reduction

Principal component analysis (PCA)

PCA is a mathematical approach to reduce dimensionality for analysis or visualization. It exploits correlations to transform the data such that the first few dimension or features contain a majority of the information of variance in the dataset. PCA determines which variables are most informative based on the distribution of data and calculates the most informative combinations of existing variables within the dataset. PCA works well for datasets with high dimensionality.

- No information is lost, first few components hold much of the information
- Same premise as linear regression except without a dependent variable
    - Linear regression solution is the first principal component
    - Disregarding the information describing the principal component, PCA calculates the second more informative component, then the third, and so on
- Linear combinations form a set of variables can be used to view the data – new axes
- Components are ranked by importance, so all but the first few can be discarded, leaving only the most important information with very few components
- The coefficients in the table give the proportion of each of the original variables that went into each component
- Relative signs +/- indicate that two variables are positively negatively correlated in that particular component

- The components are difficult to interpret using only the coefficient values, plotting often improves understanding

PC1 = (*a*var1 + b*var2 + c*var3 + …*)
PC2 = (*d*var1 + e*var2 + f*var3 + …*)
PC3 = (*g*var1 + h*var2 + i*var3 + …*)

Clustering

Another way of thinking about dimensionality reduction is how close each point is to other points. The idea is to separate data points into a number of clusters that have less distance between the points internally than to other clusters. Clustering can be helpful to identify groups of records that have similar characteristics to one another. When data is unlabeled, clustering can be used to group records together for deeper inspection. Upon deeper inspection of the records in each cluster, users can understand the patterns that lead to records being grouped together, and also identify reasons for records being grouped separately.

K-means clustering

k-means clustering starts with selecting the number of clusters, k. k cluster-centers are placed randomly in the data space and then the following stages are performed repeatedly until convergence. K-means does not determine the appropriate number of clusters, this is set by the user based on intuition or previous knowledge of the data The algorithm can terminate with multiple solutions depending on initial random positions of cluster-centers and some solutions are better than others.

- Data points are classified by the center to which they are nearest
- The centroid of each cluster is calculated
- Centers are updated to the centroid location

Classifications

- Clustering and PCA allow users to see patterns in the data, which is the best that can be done because there are no labels to guide the analysis
- With supervised learning, the label is included in the learning process:
    - Unsupervised: what features are most important or interesting?
    - Supervised: what features are most informative about the differences between these groups?
- Classification methods: each record falls into some category or class, predict the category of a new record based on values of other features in the record
- Regression methods: one variable depends on some or all of others, predict the value of the dependent variable based on the values of the independent variables

Classification Trees

Classification trees split data to find optimal values for features, used to split data by class. Tree diagrams show the class makeup of each node, and the relative number of data points that reach each node

- Tree pruning

- Tree pruning removes rules associated with overfitting from the tree
- The new tree misses a few points classified correctly, but contains only meaningful rules, more generalizable to new data

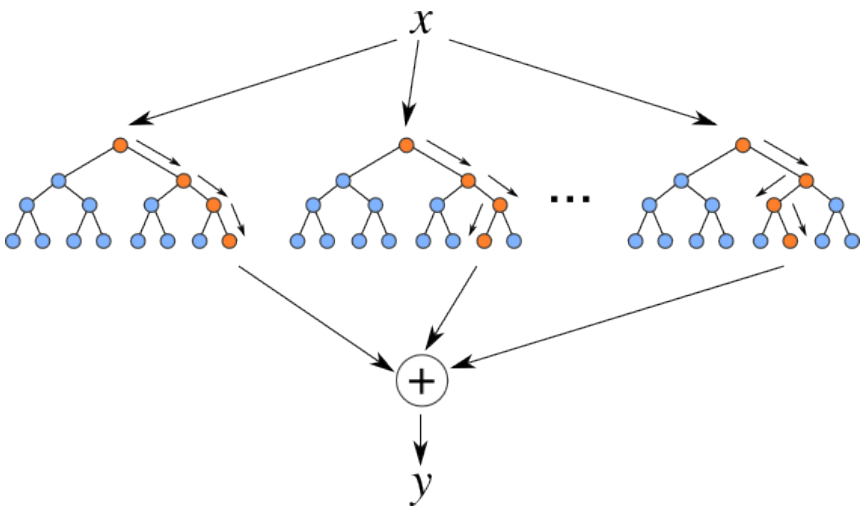
Naïve Bayes classifier

- The Naïve Bayes algorithm considers the value of each feature independently, for each record, and computes the probability that a record falls into each category
- Next, the probabilities associated with each feature are combined for each class according to Bayes' rule to determine the most likely category for each new record
- Almost completely immune to overfitting - Individual points have minimal influence; Very few assumptions are made about the data

Random forest

Random forest is an ensemble classifier that uses multiple different classification trees. Trees are generated using random samples of records in the original training set. Accuracy and information about variable importance is provided with the result.

- No pruning necessary
- Trees can be grown until each node contains very few observations
- Better prediction than classification
- No parameter tuning necessary



Comparing predictor accuracy

Cross - validation

- Models should be good at making classifications of unlabeled data, not describing data that is already classified.
- Randomly divide data into a training set and a test set
  - Hide test set while building the tree
  - Hide training set while calculating accuracy
  - Computed accuracy represents accuracy on unseen data

- Techniques are available to do this multiple times, ensuring each record is in the test set exactly once, e.g. k-folds

Comparing models

- Several standard measure of performance exist, can run multiple models and compare metrics:
  - Accuracy
  - Precision
  - Recall
  - And more
- Application drives which performance metrics are most important for a given task

Sensitivity and specificity

Sensitivity and specificity are statistical measures of the performance of a classification test. Sensitivity measures the proportion of positives are identified. Specificity measures the proportion of negatives that are identified.

Sensitivity	$\frac{a}{a+b}$	**Specificity**	$\frac{d}{c+d}$
Positive Likelihood Ratio	$\frac{\text{Sensitivity}}{1 - \text{Specificity}}$	**Negative Likelihood Ratio**	$\frac{1 - \text{Sensitivity}}{\text{Specificity}}$
Positive Predictive Value	$\frac{a}{a+c}$	**Negative Predictive Value**	$\frac{d}{b+d}$

The ROC Curve

- The Receiver Operating Characteristic (ROC) curve plots the true positive rate (Sensitivity) versus the false positive rate (100 - Specificity) for different cut-off points
- Each point on the curve represents a pair of sensitivity/specificity values corresponding to a particular decision threshold
- A test with perfect discrimination (no overlap in the two distributions) has an ROC curve that passes through the upper left corner (100% sensitivity, 100% specificity)
- The closer the ROC curve is to the upper left corner, the higher the overall accuracy of the test

Learning Objectives

- Be further introduced to machine learning algorithms and how to work with them
- Become familiar with dimensionality reduction and when and how to use it
- Recognize when to use clustering as an approach to dimensionality reduction
- Review different classification methods such as classification tress and random forest
- Learn how to compare predictor accuracy
- Become familiar with sensitivity and specificity as indicators of a binary classification test