Machine Learning

Session 8

7th April 2018

Rama Krishna Bhupathi
ramakris@gmail.com

Agenda

PCA (Principal Component Analysis)

Terminologies (eigenvalues and eigenvectors)

Decision Trees

Variance

Variance is a measure of how a set of values are spread out. Variance is calculated as the average of the squared differences of the values and mean of the values, as per the following equation:

$$s^2 = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}{n-1}$$

Co-Variance

Covariance is a measure of how much two variables change together; it is a measure of the strength of the correlation between two sets of variables. If the covariance of two variables is zero, the variables are uncorrelated. Note that uncorrelated variables are not necessarily independent, as correlation is only a measure of linear dependence.

The covariance of two variables is calculated using the following equation:

$$\operatorname{cov}(X,Y)=\frac{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)\left(Y_i-\bar{y}\right)}{n-1}$$

Eigenvectors and Eigenvalues

A vector is described by a direction and magnitude, or length. An eigenvector of a matrix is a non-zero vector that satisfies the following equation:

$$A\vec{v} = \lambda\vec{v}$$

In the preceding equation, $\vec{v}$ is an eigenvector, A is a square matrix, and λ is a scalar called an eigenvalue.

http://setosa.io/ev/eigenvectors-and-eigenvalues/

Calculate Eigenvectors and Eigenvalues

Code in python

```
from numpy import linalg as LA
import numpy
#x=numpy.array([[1,2,3] ,[4,5,7], [6,9,7]])
x=numpy.array([[1,-2,] , [2,-3]])
w,v=LA.eig(x)
print w
print v
```

Eigenvectors and Eigenvalues

- Eigenvectors and eigenvalues can only be derived from square matrices, and not all square matrices have eigenvectors or eigenvalues.
- If a matrix does have eigenvectors and eigenvalues, it will have a pair for each of its dimensions.
- The principal components of a matrix are the eigenvectors of its covariance matrix, ordered by their corresponding eigenvalues.
- The eigenvector with the greatest eigenvalue is the first principal component; the second principal component is the eigenvector with the second greatest eigenvalue, and so on.

Problems with Data

- Curse of dimensionality, as dimension increases number of samples required increases exponentially.
- The more dimensions ,difficult to acquire such large data or infeasible
- More resources are required to deal with large data (more disk , more RAM).
- Sparseness of data increases with increased data.

PCA.. what is it?

- Also known as Dimension Reduction
- Used to mitigate problems caused by the curse of dimensionality.
- Dimensionality reduction can be used to compress data while minimizing the amount of information that is lost.
- Understanding the structure of data with hundreds of dimensions can be difficult;
- Data with only two or three dimensions can be visualized easily.

PCA.. how does it work?

PCA also known as KLT (Karhunen-Loeve Transform)

- Searches for patterns in High Dimensional Data
- Explores and visualize high dimensional data sets
- Compresses Data and processes Data before being passed to Estimators
- Reduces set of correlated high dimensional variables to lower-dimensional sets of uncorrelated variables called principal components
- Lower dimensional data preserves as much of the variance of the original data as possible

PCA.. how does it work?

- PCA reduces the dimensions of a data set by projecting the data onto a lower-dimensional subspace. eg: a two dimensional data set could be reduced by projecting the points onto a line

- Each instance in the data set would then be represented by a single value rather than a pair of values.

- A three-dimensional dataset could be reduced to two dimensions by projecting the variables onto a plane.
- In general, an n-dimensional dataset can be reduced by projecting the dataset onto a k-dimensional subspace, where k is less than n.

PCA.. how does it work?

PCA can be used to find a set of vectors that span a subspace, which minimizes the sum of the squared errors of the projected data. This projection will retain the greatest proportion of the original data set's variance.

The motivation of PCA is similar; it can project data in a high-dimensional space to a lower-dimensional space that retains as much of the variance as possible. PCA rotates the data set to align with its principal components to maximize the variance contained within the first several principal components.

PCA.. More information

http://setosa.io/ev/principal-component-analysis/

Principal component analysis (PCA) is a technique used to emphasize variation and bring out strong patterns in a dataset. It's often used to make data easy to explore and visualize.

Decision Trees

- Classification Algorithm
- Supervised
- Decision Making
- Helps you build a flow chart as part of decision making

Decision Trees.. example

- You want to build a system to filter out resumes based on historical hiring data
- You have a database of attributes of candidates and which ones you hired and which ones you did not hire

- Based on the data you can predict if a new candidate will get hired or not (who to bring in for an interview)

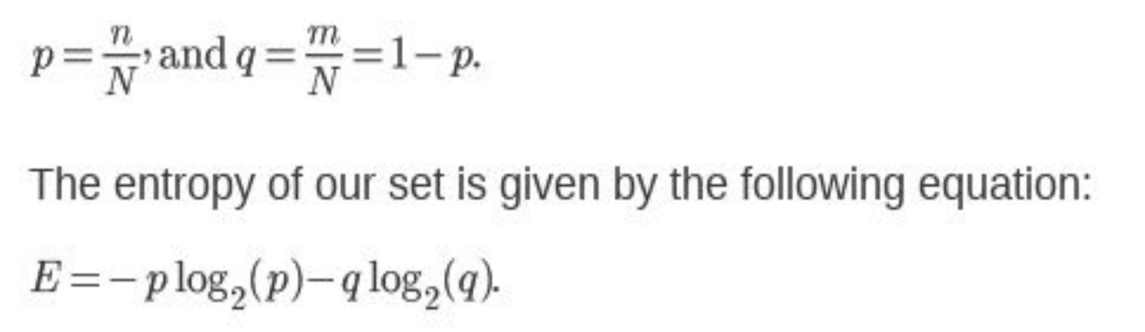
Decision Trees... *how does it work?*

- At each step , find the attribute we can use to partition the data set to minimise the entropy of the data at the next step.
- It is a greedy algorithm -- as it goes down the tree , it just picks the decision that reduce entropy the most at this stage.
- Entropy is represented by gini in the Python Notebooks.
- It may or may not result in an optimal tree.
- Algorithm used is ID3(Iterative Dichotomiser 3)

Entropy in Decision Trees

Entropy is a measure of disorder
Entropy is an indicator of how messy your data is

Let us imagine we have a set of N items. These items fall into two categories, n have Label 1 and m=N-n have Label 2. As we have seen, to get our data a bit more ordered, we want to group them by labels. We introduce the ratio



$$p = \frac{n}{N}, \text{and } q = \frac{m}{N} = 1 - p.$$

The entropy of our set is given by the following equation:

$$E = -p\log_2(p) - q\log_2(q).$$

Decision Trees…
If you have more than 2 variables:

$$E = -\sum_i p_i \log_2 p_i,$$

where the p_i are the ratios of elements of each label in the set. It is quite straightforward!

Advantages of Decision Trees

- Simple to understand and interpret.
- Able to handle both numerical and categorical data
- Requires little data preparation.
- Performs well with large datasets.

Random Forests

- Decision trees are very susceptible to overfitting
- To overcome , we can construct several alternate decision trees and let them vote on the final classification
    - Randomly re-sample the input data or each tree (called bootstrap aggregating or bagging).
    - Randomize a subset of the attributes each step is allowed to choose from