Revision of Machine Learning Course @Paris

`MachineLearning`

24/03期末考试，临时抱个佛脚。

General Concepts

- What is machine learning?

  > "Field of study that gives computers the ability to learn without being explicitly programmed".
  >
  > ———— Arthur Samuel

- What is supervised learning? Explain it with an example.
  Supervised learning involves building a statistical model for predicting or estimating, an output based on one or more input.

In this case, all the input are labeled. For example, we can predict one day's weather using the weather data of the last 10 days.

- What is unsupervised learning? Explain it with an example.
  With unsupervised learning, there are input but no supervising output.For example, we have a lot of articles and their features. With unsupervised we can divide them into few groups (maybe the articles of same group share the same author or same idea).

Key words to know

- Qualitative and quantitative variables
  Qualitative variables represent the variables can be described as a numerical values, such as human's age and height.
  Quantitative variables take on values in one of K classes, such as human's gender.
- Trainining set, Test set, Cross Validation
  Training set is the subset which we use to tarin the algorithim. Test set is used to set the performance of this algorithm. Cross Validation is a resampling method, we will add an extra CV set, after tarining our learning algorithm we will choose the parameters work better in CV set to avoid overfitting.
- Linear Regression, Simple and multiple
  The method predicts the relationship between input and output is linear.
  $Y = \beta_0 + \beta_1 X$
  For multiple linear regression, we can increase the feature using as predictors to extend the simple model.
  $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon$
- Classification, KNN, logistic regression
  The output in classification is quantitative variable.
  Logistic function :$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$
  KNN means K-Nearest Neighbors
  Find K points that is closest to the given sample of training data set.Then classfies the test observation to the class that most of its neighbors belong to.
- Tree Based Methos decision tree / entropy / prune

ID3 algo:
entropy is a measure of amount of uncertainty in the data set.
$H(S) = -\Sigma_{x \in X} p(x) log_2 p(x)$
S-current data set
X-set of classes in S
$p(x)$-proportion of the bumber of elements in class x to the number of elements in set S.
Prune: a method to get a subtree from a large tree.Usually we choose the subtree which has lower test error rate.

- SVM
  Support vector machine we choose the boundary that has max sum of distance between boundary and the points that is closest to it.

Algorithim need to know in details

ACP

Steps

1. Calculate the center of data set.
2. Regularization.Minus the center for each observation.
3. Calculate the variance/co-variance matrix.
   $V = \frac{1}{n} X'X$
   X'-transposition of X
4. Calculate the eigen value and its eigen vector.
   By solving the equation :
   $Vu = \lambda u$
5. Calculate the new coordinates of the point by.
   $X_2 = Xu$

R

```
data
Xc <- scale(data,center=TRUE,scale=FALSE)
Mcov=(1/n)*t(Xc)%*%Xc
```

```r
pc=eigen(Mcox)
Xc%*%pc$vectors

#Use FactoMineR
library(FactoMineR)
res.pca <- PCA(data,graph=FALSE,axes=c(1,2))
summary(res.pca)
```

K-means

Steps

1. Randomly split the data set into K subset.
2. Calculate the center of gravity for each set.
3. Calculate the distance between each node and each center, classify the node to the class that is closest to it.
4. Repeat 2 to 3 until nothing changes.
5. You can repeat 1 to 4 for several times and select the result with lowest variance.

R

```r
km.out <- kmens(data,k,nstart=20)
km.out$cluster
#Variance of each cluster
km.out$withinss
#Sum of variance
km.out$tot.withinss
```

CAH

Steps

1. Regard each obversation as a sub set.
2. Calculate the distance between subsets. Merge the two sub set that have the smallest distance.Choose the smallest(single)/greatest(complete)/average(average) variable as the distance of it and other subset.
3. Repeat 2 untill there is only one set.
4. Draw a tree to present its process.Label the node with the samllest value.

R

```r
hc.complete=hclust(dist(x), method="complete")
hc.average=hclust(dist(x), method="average")
hc.single=hclust(dist(x), method="single")
#Draw the plot
par(mfrow=c(1,3))
plot(hc.complete,main="Complete Linkage", xlab="", sub="", cex=.9)
plot(hc.average, main="Average Linkage", xlab="", sub="", cex=.9)
plot(hc.single, main="Single Linkage", xlab="", sub="", cex=.9)
#Do the cluster
cutree(hc.complete, 2)
cutree(hc.average, 2)
cutree(hc.single, 2)
cutree(hc.single, 4)
xsc=scale(x)
plot(hclust(dist(xsc), method="complete"), main="Hierarchical
Clustering with Scaled Features")
```