Data Mining

LABDATA

Coordenação:

Prof. Dr. Adolpho Walter Pimazzi Canton

Profa. Dra. Alessandra de Ávila Montini

Disciplina: Machine Learning

Tema da Aula: Unsupervised Learning

Prof. Carlos Eduardo Martins Relvas

Currículo

- Bacharel em Estatística, Universidade de São Paulo.
- Mestre em Estatística, Universidade de São Paulo.
- Itaú, 2010-2015. Principais atividades:
  - Consultoria estatística para várias áreas do banco com foco principal em melhorias no processo de modelagem de risco de crédito.
  - De 2013 a 2015, participação do projeto Big Data do banco usando tecnologia Hadoop e diversas técnicas de machine learning. Desenvolvemos diversos algoritmos em MapReduce usando R e Hadoop streaming, criando uma plataforma de modelagem estatística no Hadoop.
- Nubank, dede 2015. Principais atividades:
  - Equipe de Data Science, responsável por toda a parte de modelagem da empresa, desde modelos de crédito a identificar motivos de atendimento.

Conteúdo da Aula

- Unsupervised Learning
- Clustering
  - Kmeans
  - Hierárquico
- PCA (Principal component analysis)

Unsupervised Learning

Supervised Learning

- Variável resposta (target, output) Y é observada.
- P variáveis explicativas (variáveis independentes, features, covariáveis, inputs).
- Se Y é contínua, temos um problema de regressão.
- Em problemas de classificação, Y assume valores finitos não ordenados.
- Dados: (x_1, y_1), (x_2, y_2), ... , (x_n, y_n). x_i é um vetor de tamanho p.

Supervised Learning

Objetivos:

- Prever o comportamento do fenômeno em novos casos (dados de teste).
- Estudar a relação entre as variáveis explicativas e a resposta.
- Verificar a qualidade das predições.

Supervised Learning






Carros






Motos


Unsupervised Learning

- Não há uma variável resposta. Somente variáveis explicativas.

- Objetivos são mais diversos: encontrar observações que são mais parecidas, encontrar variáveis explicativas que se comportam de maneira parecida, etc.

- Como saber a performance do método?

Unsupervised Learning












Não sabemos o que as imagens acima são!


Agrupamento

K-Means

- Há diversas técnicas de agrupamento (clustering). K-means é a mais popular dentre elas.

- Clustering consiste em criar grupos de observações similares. Mas o que é ser similar?

- A medida de similaridade pode depender de caso a caso. Frequentemente, utiliza-se a distância euclidiana.

K-Means

- O número de grupos é definido a priori. Isto nem sempre é fácil.

- Quando não sabemos a quantidade certa de grupos, costuma-se executar o algoritmo várias vezes variando o tamanho K, número de grupos. Verifica-se qual obteve melhor resultado.

K-Means

- Algoritmo:
  1. Aleatoriamente, para cada observação fixe um grupo, de 1 a K. Estes grupos servem como chute inicial.
  2. Faça iterativamente, até as observações pararem de mudar de grupo:
     a) Para cada cluster (1 a K), compute o centroide do cluster. O centroide é o vetor médio de todas as variáveis.
     b) Para cada observação, calcule a sua distância para todos os centroides.
     c) Mude a observação para o grupo que esteja mais perto do centroide.

K-Means

- O algoritmo do K-Means não garante a convergência para um mínimo global.
- Diferentes inícios levam a diferentes resultados.
- Executamos várias vezes e escolhemos a execução que apresenta melhor performance.
- Performance significa a menor variância dentro de cada cluster.

K-Means

Base simulada com 150 observações e 6 variáveis.

- Gastos no cartão em reais
- Idade
- Renda
- Pagamento de impostos
- Segmento

Objetivo:
Vamos assumir que não temos o segmento. Será que conseguimos segmentar a base com as variáveis explicativas disponíveis?

```
> head(dados)
  Gastos_Cartao Idade Renda Impostos Segmento
1           510    35  1120       60        C
2           490    30  1120       60        C
3           470    32  1040       60        C
4           460    31  1200       60        C
5           500    36  1120       60        C
6           540    39  1360      120        C
```

K-Means

Criando bases de desenvolvimento e teste

Variabilidade dos resultados para diferentes pontos iniciais

Processo Iterativo

Avaliação dos resultados

K-Means

Criando bases de desenvolvimento e teste

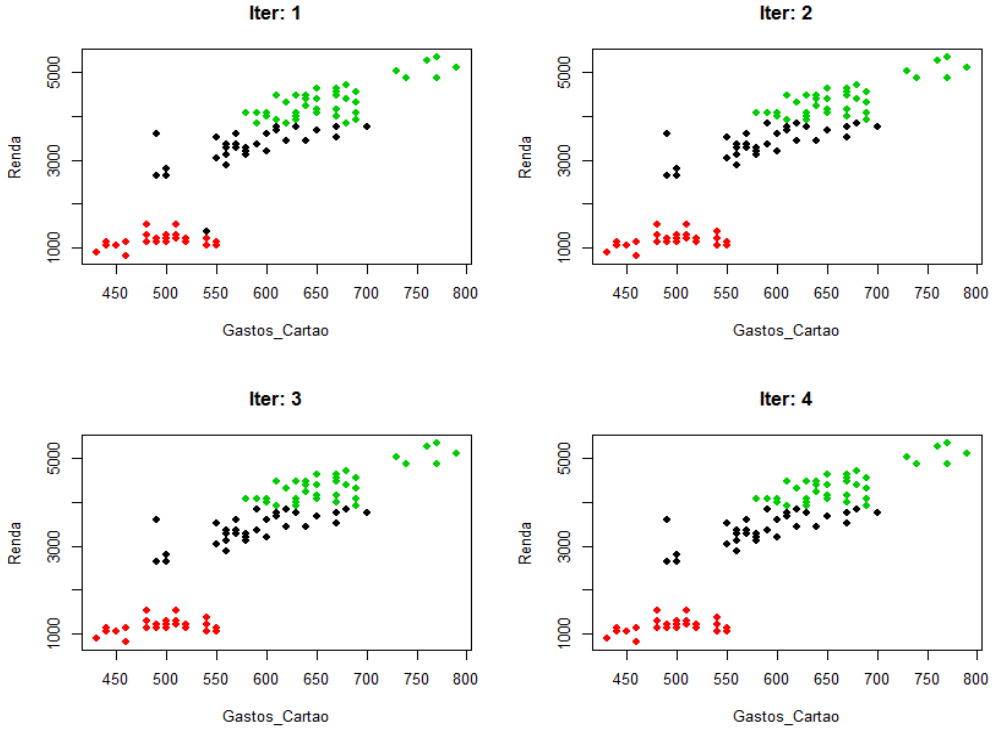
```
> set.seed(432)
> id <- sample(1:nrow(dados), nrow(dados)*0.7)
> dados.des <- dados[id,]
> dados.test <- dados[-id,]
```

Processo Iterativo

```
> par(mfrow=c(2,2))
> for(i in 1:4){
+   set.seed(56)
+   teste <- kmeans(dados.des[, c("Gastos_Cartao", "Renda")],
+                   3, nstart=1, iter.max=i)
+   plot(dados.des[,"Gastos_Cartao"], dados.des[, "Renda"],
+        col=teste$cluster,
+        pch=16, xlab="Gastos_Cartao", ylab="Renda",
+        main=paste0("Iter: ", i))
+ }
Warning messages:
1: did not converge in 1 iteration
2: did not converge in 2 iterations
```

K-Means

Processo Iterativo



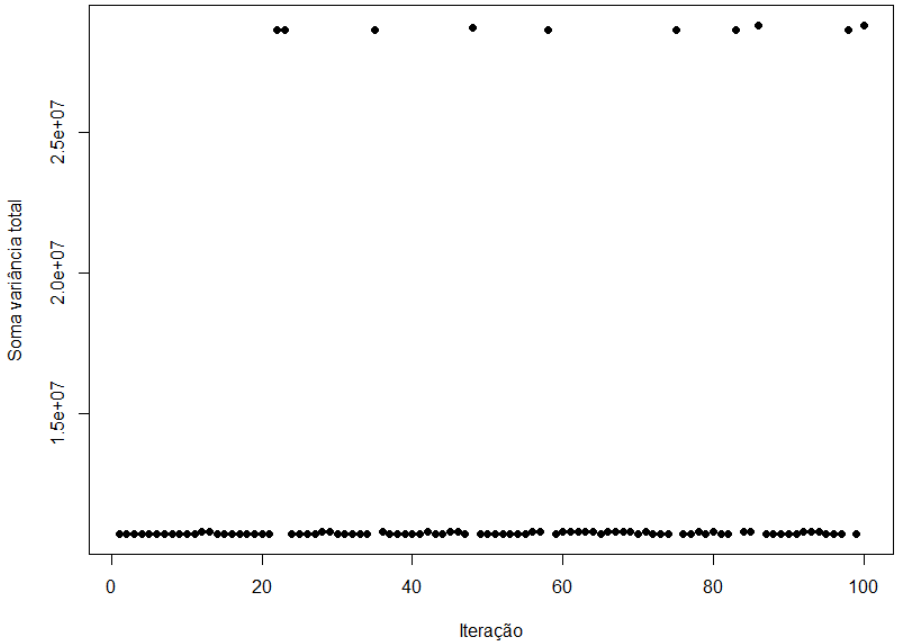
K-Means

Variabilidade dos Resultados para diferentes chutes iniciais

```
> grupos <- array()
> for(i in 1:100){
+   set.seed(i)
+   grupos[i] <- kmeans(dados.des[, c("Gastos_Cartao", "Idade",
+                                     "Renda", "Impostos")],
+                       3, nstart=1)$tot.with
+ }
> plot(grupos, pch=16, xlab="Iteração", ylab="Soma variância total")
```

K-Means

Variabilidade dos Resultados para diferentes chutes iniciais



K-Means

Avaliação dos resultados

```
> set.seed(31)
> agrupamento = kmeans(dados.des[, c("Gastos_Cartao", "Idade",
+                                    "Renda", "Impostos")],
+                      3, nstart=100)
>
> table(dados.des$Segmento, agrupamento$cluster)

    1  2  3
A  33  4  0
B   2 34  0
C   0  0 32
```

K-Means

Exercícios: http://archive.ics.uci.edu/ml/datasets/User+Knowledge+Modeling#

6 variáveis de 258 estudantes cursando uma matéria de engenharia (“students.txt”). Variáveis:

STG - The degree of study time for goal object materails
SCG - The degree of repetition number of user for goal object materails
STR - The degree of study time of user for related objects with goal object
LPR - The exam performance of user for related objects with goal object
PEG - The exam performance of user for goal objects
UNS - The knowledge level of user (Very Low, Low, Middle, High)

K-Means

1.) Construa as bases de desenvolvimento (80%) e validação (20%). Use seed de 121.

2.) Faça o gráfico de dispersão entre LPR e PEG, eixos x e y respectivamente. Discuta com seus colegas a relação entre estas duas notas.

3.) Use o algoritmo de K-means para criar grupos de estudantes parecidos e com isso tentar inferir o nível dos estudantes. Compare com a variável resposta (UNS). Quantos grupos devemos usar?

4.) Use todas as variáveis, exceto UNS, para tentar melhorar o agrupamento. Há melhoras?

Cluster Hierárquico

- Cluster hierárquico consiste em criar clusters de forma hierárquica. Isto é, agrupando clusters menores ou desagrupando clusters grandes.

- São divididos em:
  - Aglomerativos: inicialmente, cada observação é um cluster e os clusters são agrupados de acordo com algum critério de similaridade.
  - Divisivos: inicialmente todas observações formam um único cluster que é dividido sequencialmente.

- O cluster hierárquico mais utilizado é conhecido como método de Ward.

Método de Ward

- Inicialmente cada observação forma um único cluster.
- Iterativamente, agrupamos os dois clusters mais próximos até termos apenas um único cluster com todas as observações.
- Como calculamos a distância entre clusters?
- Teremos sempre apenas um cluster no final?

Método de Ward

- No método de Ward, a distância entre dois clusters (i e j) , quaisquer que sejam seus tamanhos, é definida como:

- $d_{ij} = \frac{n_i n_j}{n_i + n_j} ||c_i - c_j||$

- Em que n_i representa o número de observações no i-ésimo cluster, n_j o número de observações no j-ésimo cluster, c_i o centróide do i-ésimo cluster e , c_j o centróide do j-ésimo cluster.

- Além disso, $\left|\left|c_i - c_j\right|\right| = \sqrt{\sum_{l=1}^{p}(c_{il} - c_{jl})^2}$

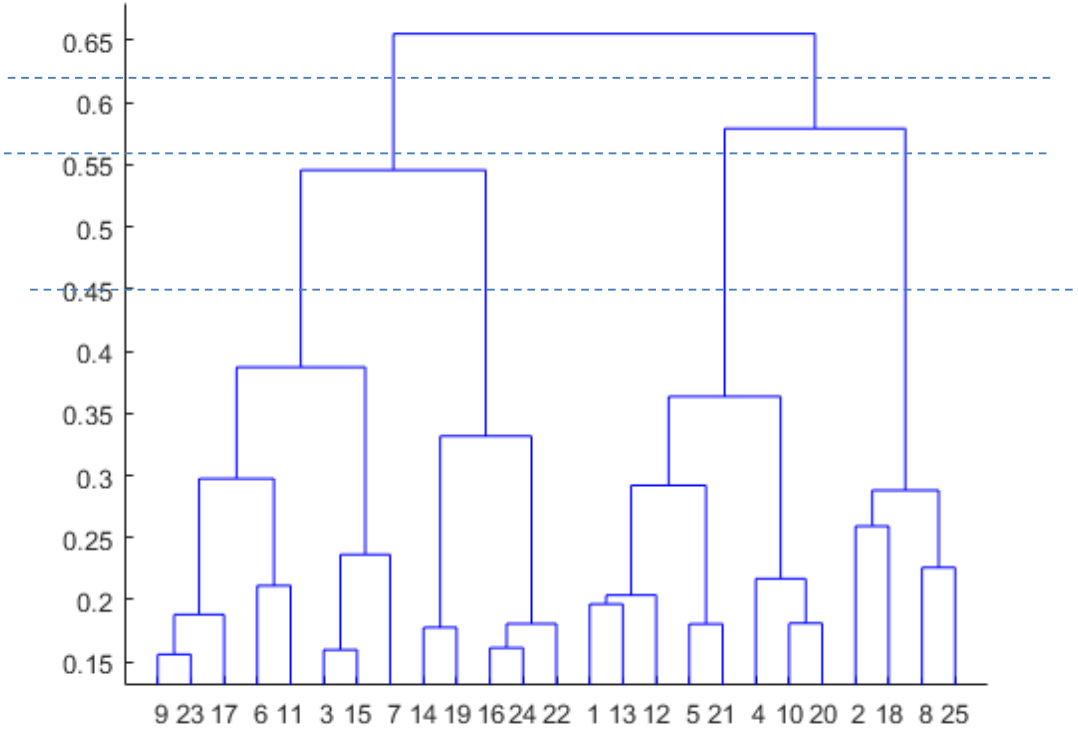
Método de Ward

- Teremos sempre um cluster no final?

- Não!! Como o processo é iterativa, podemos escolher qualquer número de clusters. Em geral, utilizamos uma ferramente chamada dendograma para a escolha do número ótimo de clusters.

- O dendograma é um mapa da clusterização, em que podemos ver o ’esforço’ necessário para unir dois clusters e assim poder escolher qual clusterização nos atende melhor.

Método de Ward



Método de Ward

Base simulada com 150 observações e 6 variáveis.

- Gastos no cartão em reais
- Idade
- Renda
- Pagamento de impostos
- Segmento

Objetivo:
Vamos assumir que não temos o segmento. Será que conseguimos segmentar a base com as variáveis explicativas disponíveis?

```
> head(dados)
  Gastos_Cartao Idade Renda Impostos Segmento
1           510    35  1120       60        C
2           490    30  1120       60        C
3           470    32  1040       60        C
4           460    31  1200       60        C
5           500    36  1120       60        C
6           540    39  1360      120        C
```

Método de Ward

Exercícios: http://archive.ics.uci.edu/ml/datasets/User+Knowledge+Modeling#

6 variáveis de 258 estudantes cursando uma matéria de engenharia (“students.txt”). Variáveis:

STG - The degree of study time for goal object materails
SCG - The degree of repetition number of user for goal object materails
STR - The degree of study time of user for related objects with goal object
LPR - The exam performance of user for related objects with goal object
PEG - The exam performance of user for goal objects
UNS - The knowledge level of user (Very Low, Low, Middle, High)

Método de Ward

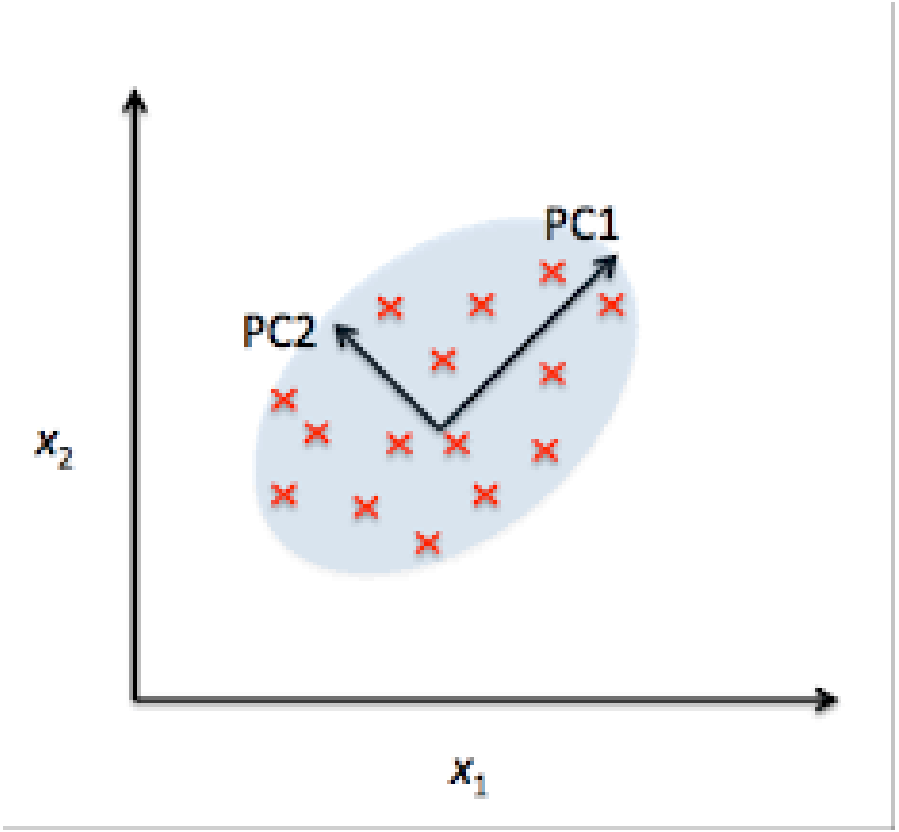
1.) Construa as bases de desenvolvimento (80%) e validação (20%). Use seed de 121.

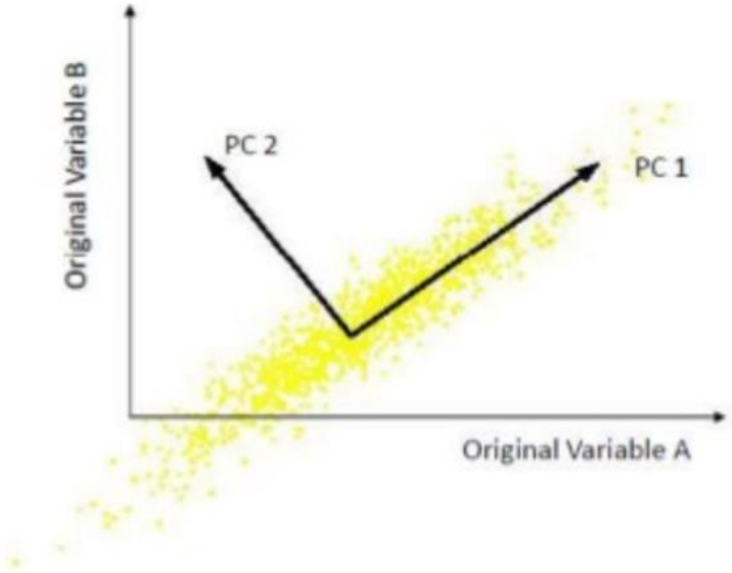
2.) Use o algoritmo de Ward para criar grupos de estudantes parecidos e com isso tentar inferir o nível dos estudantes. Compare com a variável resposta (UNS). Quantos grupos devemos usar?

3.) Use todas as variáveis, exceto UNS, para tentar melhorar o agrupamento. Há melhoras?

Análise de Componentes Principais

PCA


PC1
PC2
x_2
x_1


Original Variable B
PC 2
PC 1
Original Variable A

PCA

- A análise de componentes principais (1901 - Principal Component Analysis - PCA) é um método matemático que consiste em uma transformação (combinação linear) das variáveis para criar novas variáveis linearmente não correlacionadas, chamadas de componentes principais.

- A transformação é feita de tal modo que o primeiro componente tenha a maior explicação possível dos dados (responsável pela maior variabilidade dos dados). O segundo componente pelo segundo maior e assim por diante.

PCA

- Hoje, o PCA é utilizado com dois objetivos principais:

➢ Reduzir a dimensionalidade. Como os componentes são ordenados pela explicação dos dados, podemos utilizar apenas os primeiros. É claro que perdemos informação fazendo isso. Quanto?

➢ Como forma de análise descritiva multivaria dos dados.

PCA – Estimação

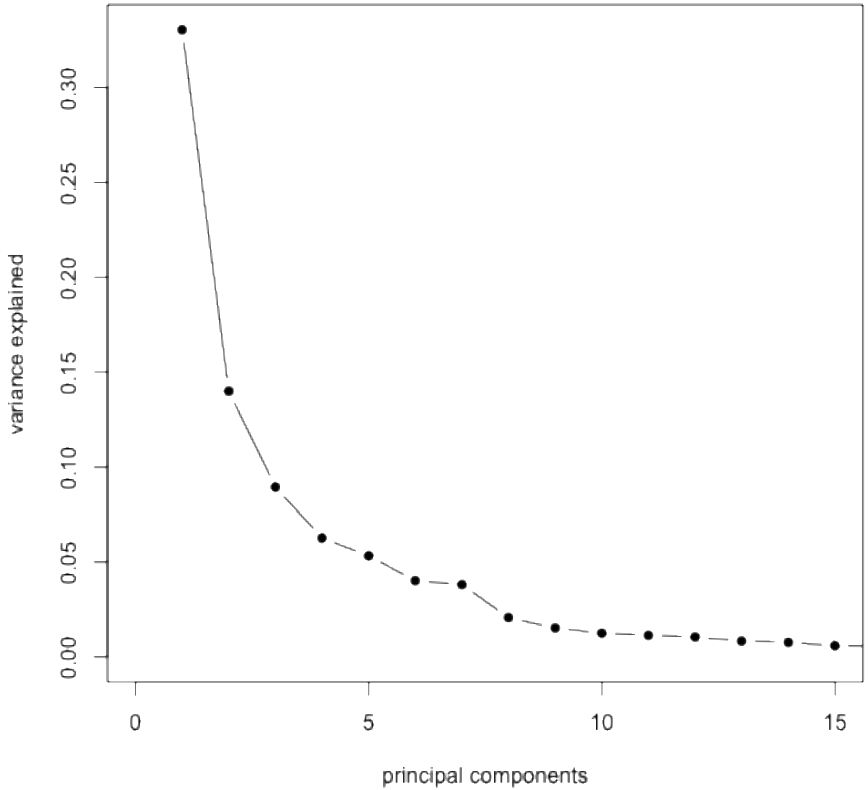
1. Normalize os dados

2. Calcule a matriz de covariância.

3. Encontrar autovalores e autovetores da matriz de covariância.

4. Os autovetores são os pesos dos componentes e os autovalores normalizados representam a proporção de explicação de cada componente.

PCA


variance explained
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0
5
10
15
principal components

PCA – Laboratório

Base de dados com informações sobre um conjunto de vinhos de Portugal.
https://archive.ics.uci.edu/ml/datasets/wine+quality

PCA – Exercício

Base de dados com informações da cidade de Boston.

Construa o PCA e compare o modelo de regressão linear utilizando as variáveis originais e utilizando as componentes principais. Quantas componentes você utilizou?

Referências Bibliográficas

- Hastie, T., Tibshirani, R. & Friedman, J.H. (2001) “The Elements of Statistical Learning”
- Bishop, C.M. (2007) “Pattern Recognition and Machine Learning”
- Mitchell, T.M. (1997) “Machine Learning”
- Abu-Mostafa, Y., Magdon-Ismail, M., Lin, H.T (2012) “Learning from data”
- Theodoridis, S., Koutroumbas, K., (2008) “Pattern Recognition”
- Kuhn, M., Johnson, K., (2013) “Applied Predictive Modeling”