

Data Mining

Disciplina: Machine Learning

Prof. Carlos Eduardo Martins Relvas

Coordenação:

Prof. Dr. Adolpho Walter
Pimazzi Canton

Profa. Dra. Alessandra de
Ávila Montini

Currículo

- Bacharel em Estatística, Universidade de São Paulo.
- Mestre em Estatística, Universidade de São Paulo.
- Doutorado em Ciência da Computação, 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, desde 2015. Equipe de Data Science, responsável por toda a parte de modelagem da empresa, desde modelos de crédito e até mesmo identificar motivos de atendimento.

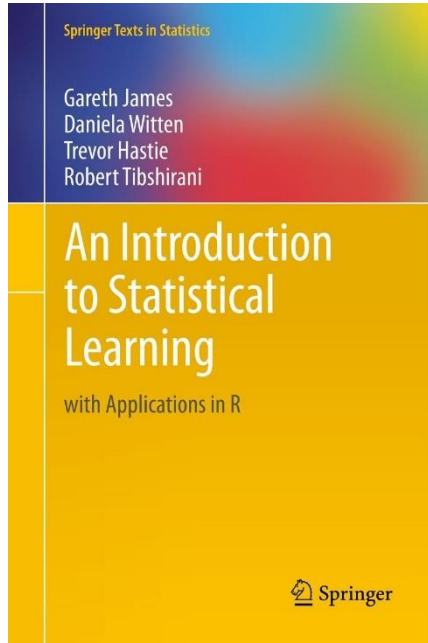
Agenda

- **Introdução Machine Learning.**
- **Regressão Linear.**
- **Árvore de decisão.**

Plano de Aulas

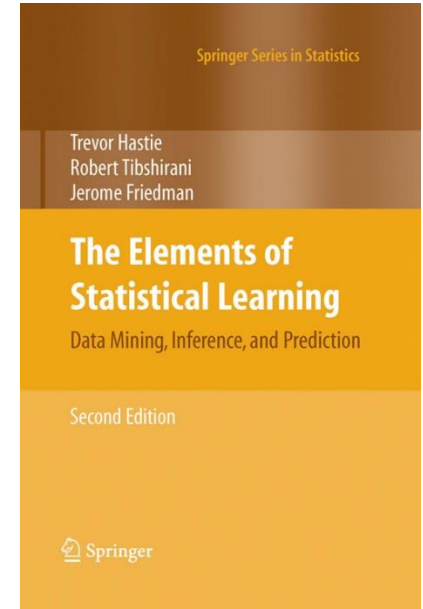
- Introdução ao R
- Regressão Linear e Regressão logística
- Etapas de um projeto de machine learning, Métricas e Validação
- Regularização
- Métodos não lineares (GAM, splines, etc)
- Métodos baseados em árvores (Árvore de decisão, Bagging, Random Forest e Boosting).
- Python (sklearn)
- Support Vector Machine (SVM)
- Rede Neural
- Técnicas de combinação de modelos (Ensemble)
- Técnicas de imputação, Feature engineering, Feature Selection
- Métodos não supervisionados
- Sistema de recomendação
- Spark e Map Reduce

Livros



- The elements of Statistical Learning,
Hastie, Tibshirani e Friedman,
Second Edition

- An Introduction to Statistical Learning,
Hastie, Tibshirani, James e Witten



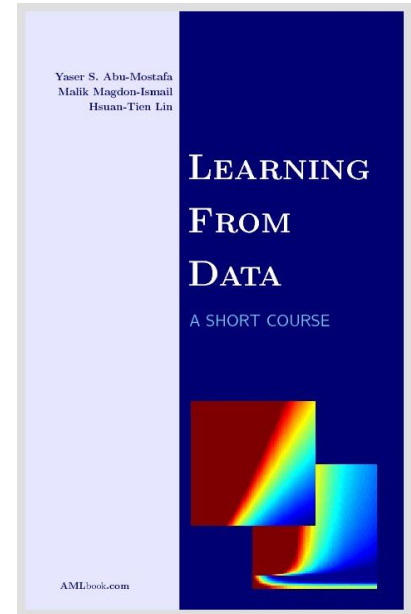
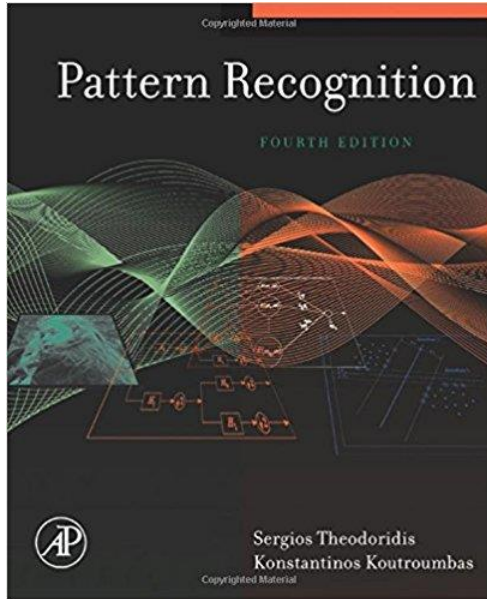
https://statweb.stanford.edu/~tibs/ElemStatLearn/printings/ESLII_print10.pdf

<http://www-bcf.usc.edu/~gareth/ISL/ISLR%20Sixth%20Printing.pdf>

Livros

- Learning from data,
Abu-Mostafa, Magdon-Ismail e Liu

- Pattern Recognition,
Theodoridis e Koutroumbas



O que eu espero no final do curso?

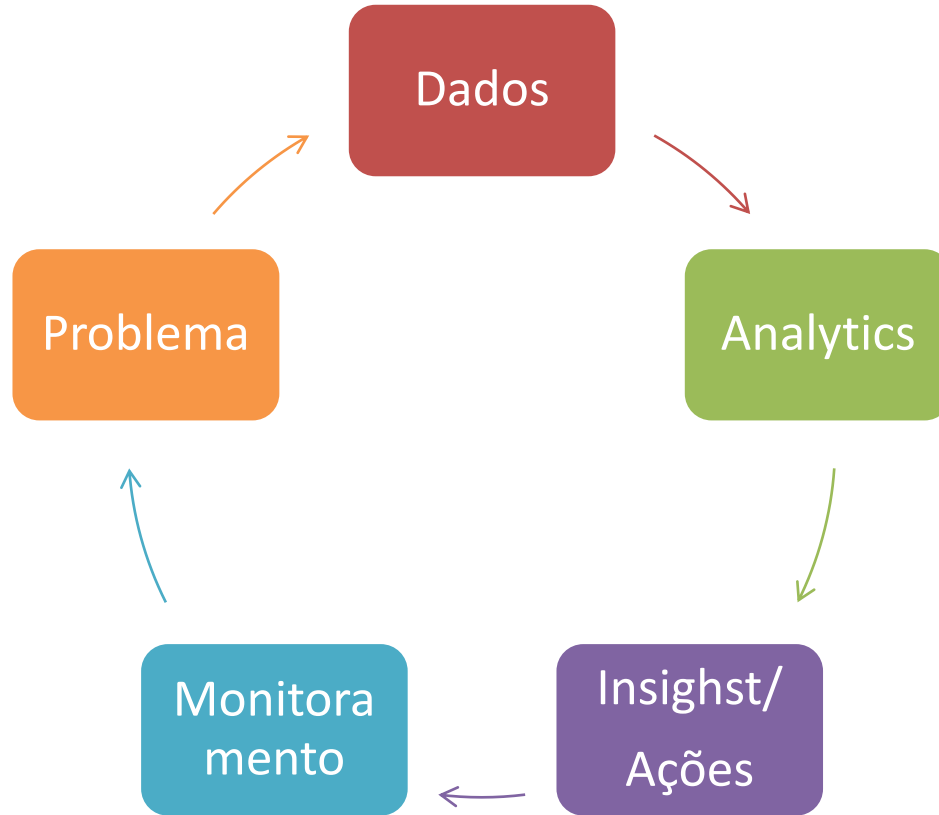
- Definir a variável resposta de forma adequada.
- Saber qual métrica utilizar e como validar seus resultados.
- Saber a melhor forma de tratar dados missing.
- Saber como proceder com o feature engineering e feature selection.
- Utilizar e otimizar o algoritmo mais eficiente.
- Saber usar o modelo preditivo criado.
- Onde executar seu projeto de machine learning.

Formato das Aulas

Cada aula será composta basicamente por 3 etapas:

- Parte teórica.
- Aplicação prática utilizando alguma das ferramentas estudadas.
- Exercícios práticos

Analytics – Soluções orientadas a dados



Mas o que mudou?

- A quantidade de dados gerados por empresas e pessoas aumenta a cada dia.
- Hoje temos tecnologias (Hadoop, Spark) a disposição para analisar esta imensidão de dados.
- Os valores gerados por decisões baseadas em dados se tornaram mais visíveis, ocasionando uma decisão em cadeia.

Risco de Crédito

Idade	24
Profissão	Engenheiro
Renda Mensal	R\$5.000,00
Gênero	Masculino
Anos no emprego	2
Anos na mesma residência	12
Gastos no cartão de crédito no último mês	R\$1.800,00
Total de investimentos	R\$20.000,00
...	...
Total de crédito anterior	R\$0

Aprovo ou nego um crédito consignado de R\$2.000,00?

Classificar E-mail como Spam

- Dados de 4601 e-mails enviados para o mesmo individuo, todos marcados como spam ou não spam.
- Objetivo: construir um método automático para detecção de spam.
- Variáveis: porcentagem que as 57 palavras ou pontuações mais frequentes aparecem em cada e-mail.

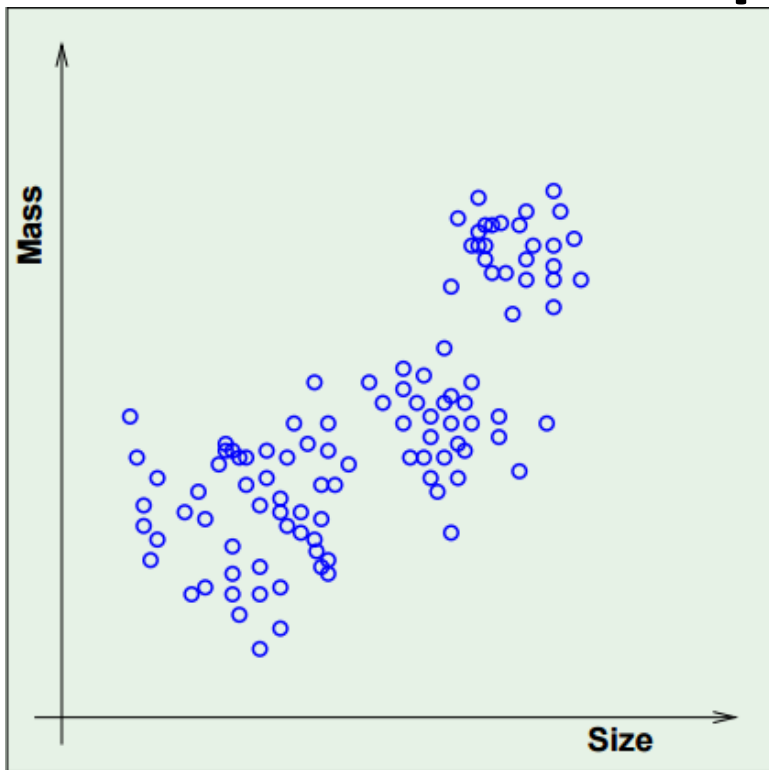
	george	you	hp	free	!	edu	remove
spam	0.00	2.26	0.02	0.52	0.51	0.01	0.28
email	1.27	1.27	0.90	0.07	0.11	0.29	0.01

Identificar os números em um CEP escrito à mão

7210414959
0690159734
9665407401
3134727121
1742351244

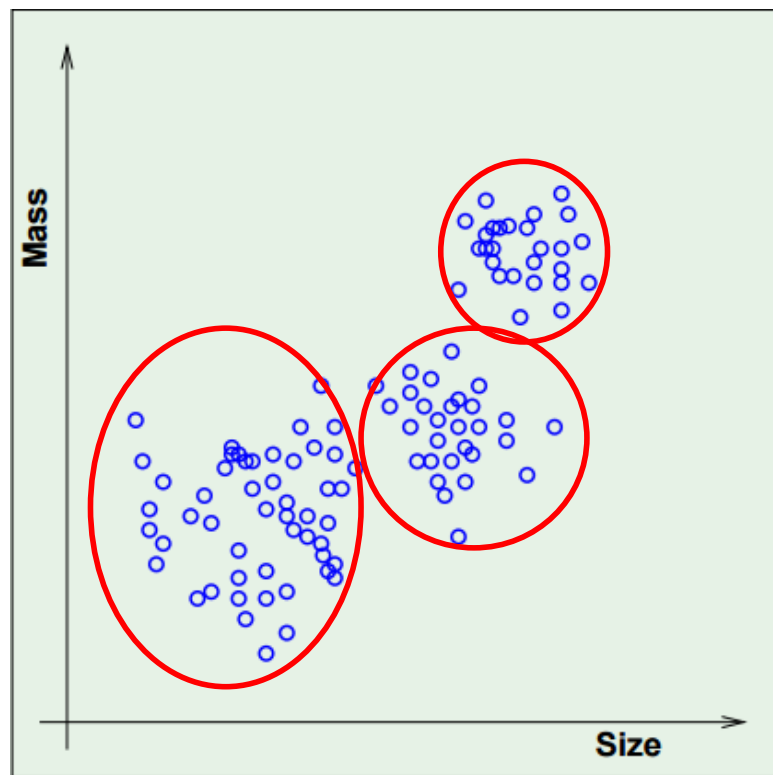
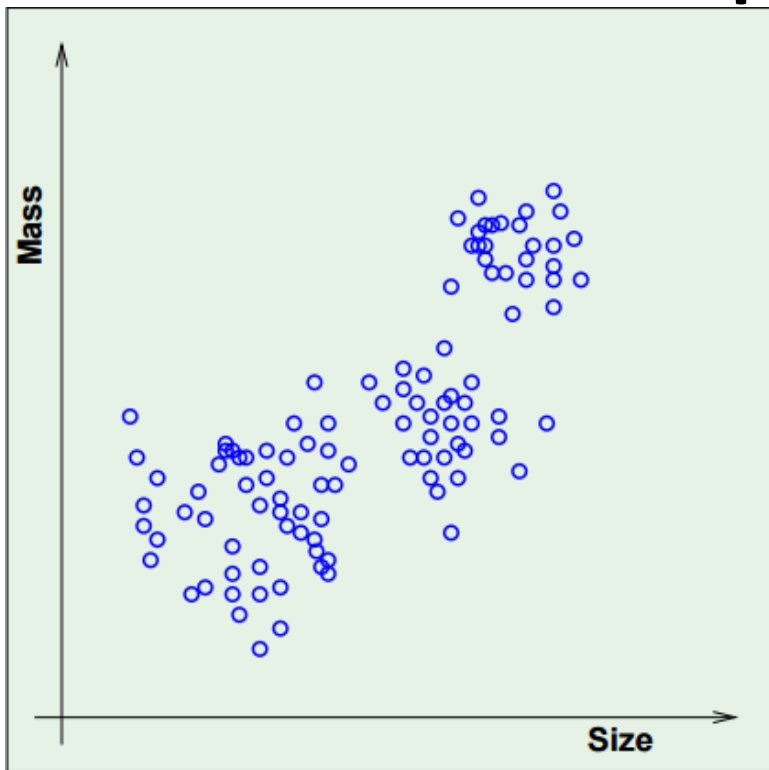
Atualmente os modelos para este tipo de função apresentam uma performance melhor do que o ser humano.

Máquinas de moeda



A máquina recebe uma nova moeda, quanto que ela vale?

Máquinas de moeda



A máquina recebe uma nova moeda, quanto que ela vale?

Identificar rostos de fraudadores recorrentes



Cadastro online por meio de uma selfie e foto dos documentos.

Fraudadores tentam aplicar recorrentemente com diversos documentos falsos. Mas sempre com o mesmo rosto.

Algoritmos de reconhecimento de face identificam que aquele rosto já fez um cadastro anterior.

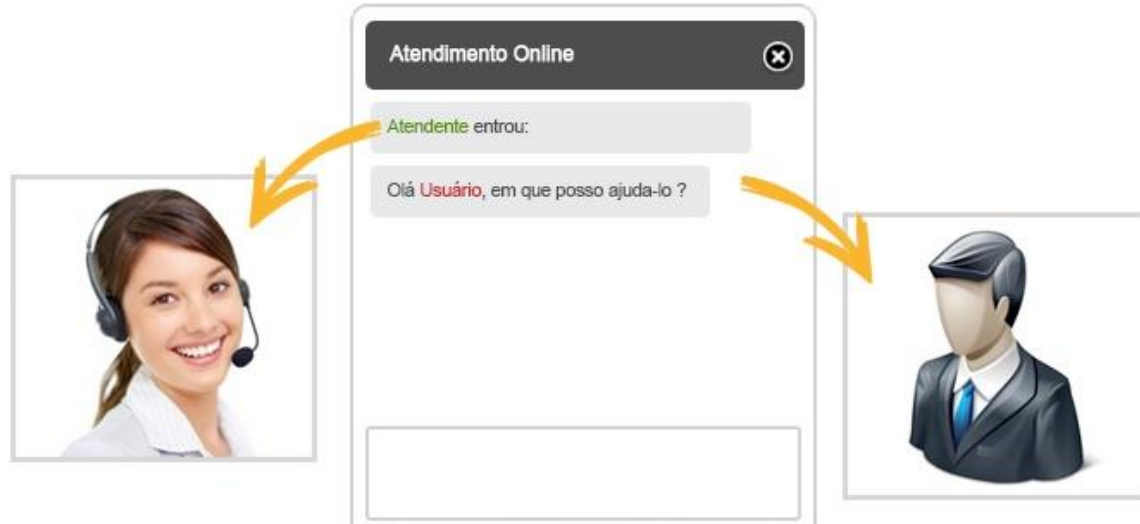
Roteamento de Chats

Usuário entra no chat e faz uma pergunta / dúvida ao atendente.

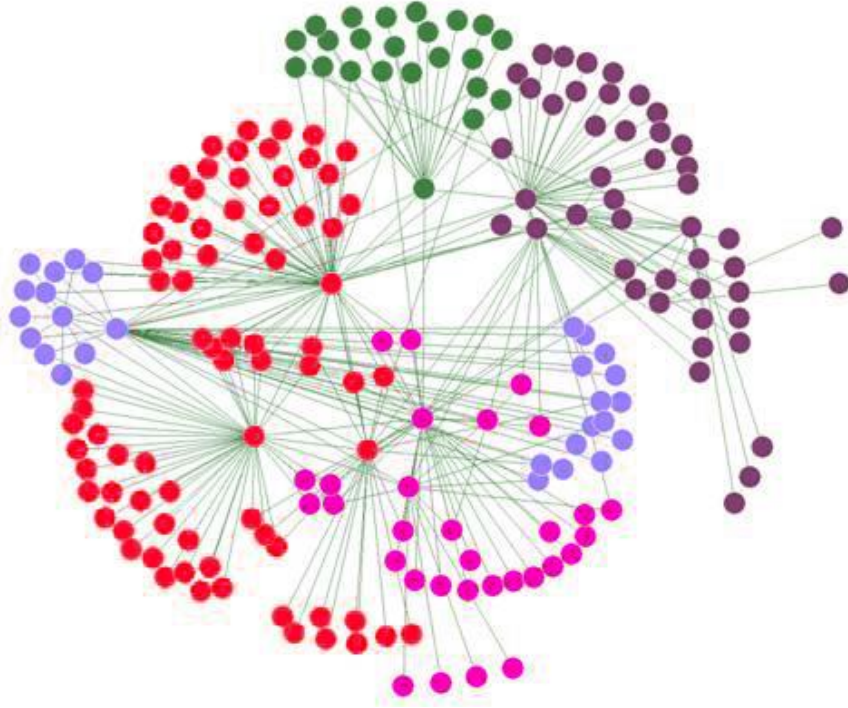
Os atendentes são divididos de acordo com suas especialidades.

Podemos criar um algoritmo para “ler” o que o usuário escreveu e tentar inferir qual o seu problema.

Assim, direcionamos para o atendente especialista neste assunto.



Social Network Analysis



Pessoas se conectam por diversas razões, desde por relações de amizades, familiares, etc.

Será que conseguimos prever características suas baseado nas informações de quem você se relaciona?

“Me diga com quem tu andas que direi quem tu és.”

Machine Learning

O que é machine learning?

Machine Learning

“Campo de estudo que fornece aos computadores a habilidade de aprender sem serem explicitamente programados” – Arthur Samuel

“Um programa de computador aprende com experiência E com respeito a algumas tarefas T e medida de performance P , se sua performance para resolver as tarefas em T , medidas por P , melhora com a experiência E ” – Tom M. Mitchell

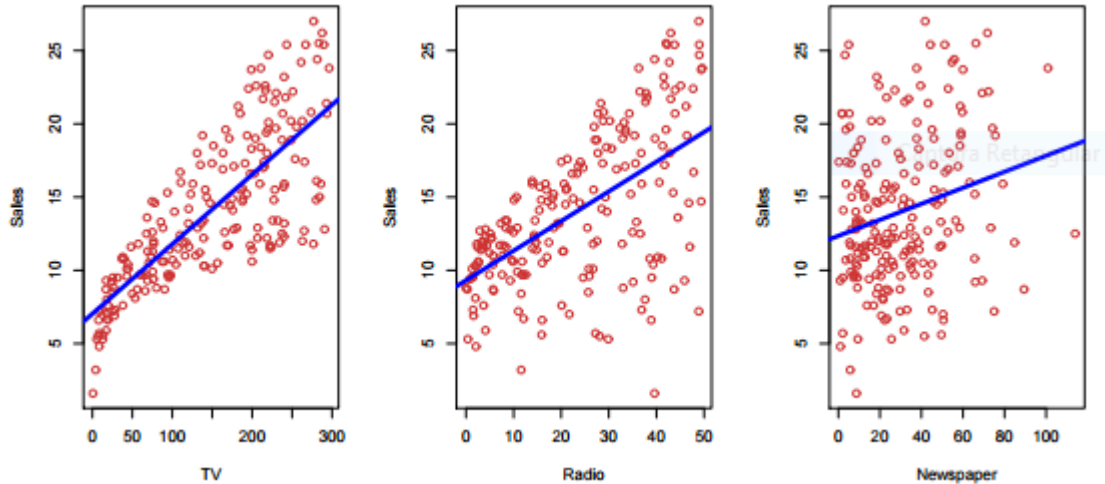


Essências

- Existe um padrão.
- Não podemos fixá-lo matematicamente.
- Temos dados desses padrões.



Machine Learning



Queremos prever as vendas como uma função do gasto em TV, Radio e Jornal.

$$\text{Sales} \approx f(\text{TV}, \text{Radio}, \text{Newspaper})$$

Machine Learning

Vendas (Y) é nossa variável resposta (target, output). Já os gastos em TV (X_1), Radio (X_2) e jornais (X_3), são as variáveis explicativas (features, input).

$$X = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} \quad Y = f(X) + \epsilon$$

Assumimos que o modelo verdadeiro dos dados é dado por $Y = f(X) + \epsilon$ em que ϵ captura erros de medidas e outras discrepâncias.

Machine Learning - Objetivo

O objetivo das técnicas de machine learning é estimar $f(x)$ da **melhor** forma possível sujeito às restrições.

O que é melhor?

Geralmente, melhor significa que as **previsões** estão **próximas** dos valores **observados** em um **novo** conjunto de **dados**.

Machine Learning - Objetivo

O objetivo das técnicas de machine learning é estimar $f(x)$ da **melhor** forma possível sujeito às restrições.

O que é melhor?

Geralmente, melhor significa que as **previsões** estão **próximas** dos valores **observados** em um **novo** conjunto de **dados**.

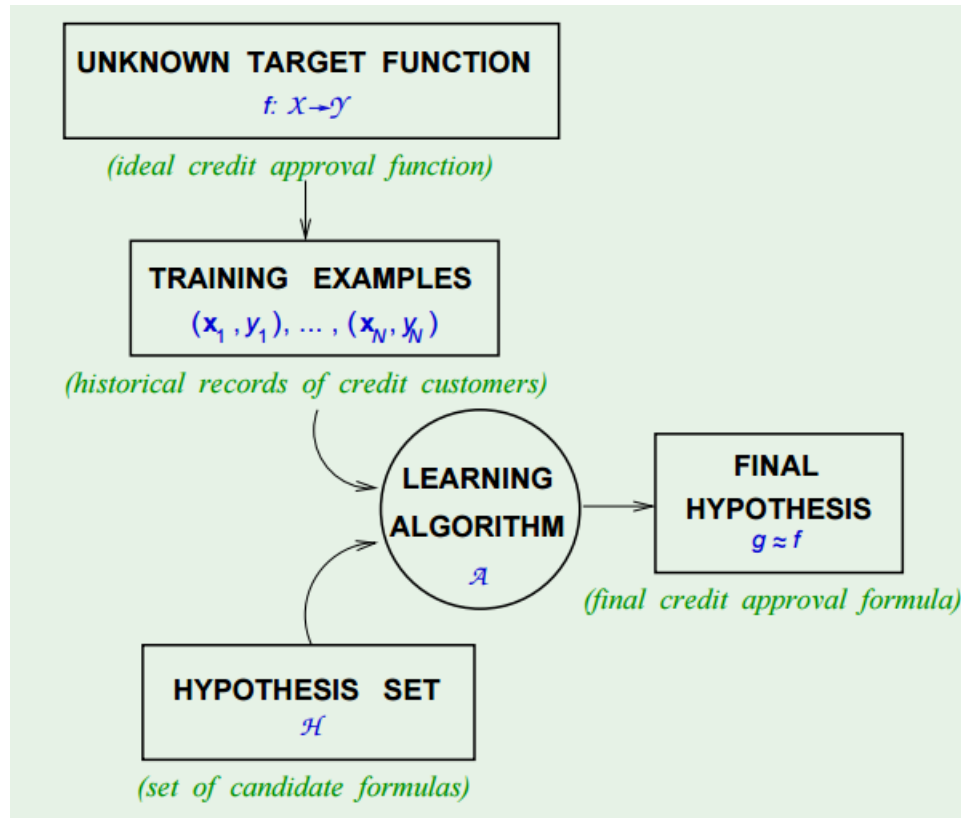
Por que
novo?

Como
medir?

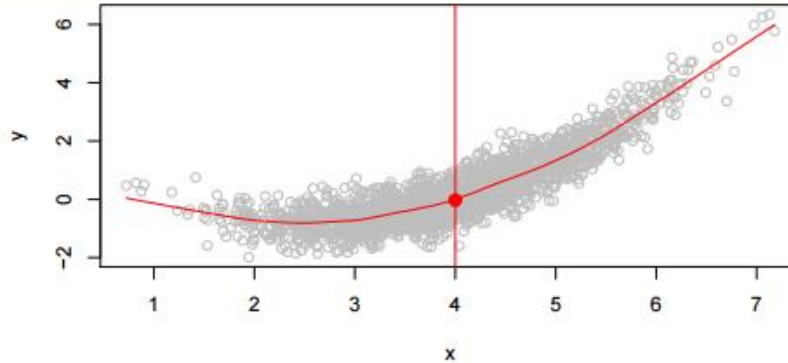
O que podemos fazer com $f(x)$?

- Previsões para novos valores de X .
- Entender quais dos componentes (X_1 , X_2 e X_3) são mais importantes.
- Entender como cada um dos componentes estão relacionados com a variável resposta (dependendo das técnicas utilizadas de estimação).

Machine Learning



Ideias de estimação



Como podemos estimar $f(x)$ para o problema acima quando $x = 4$? Um bom valor será dado por

$$f(4) = E(Y|X = 4)$$

Que representa o valor esperado de Y dado que X é igual a 4. $E(Y|X=4)$ é chamado de função de regressão.

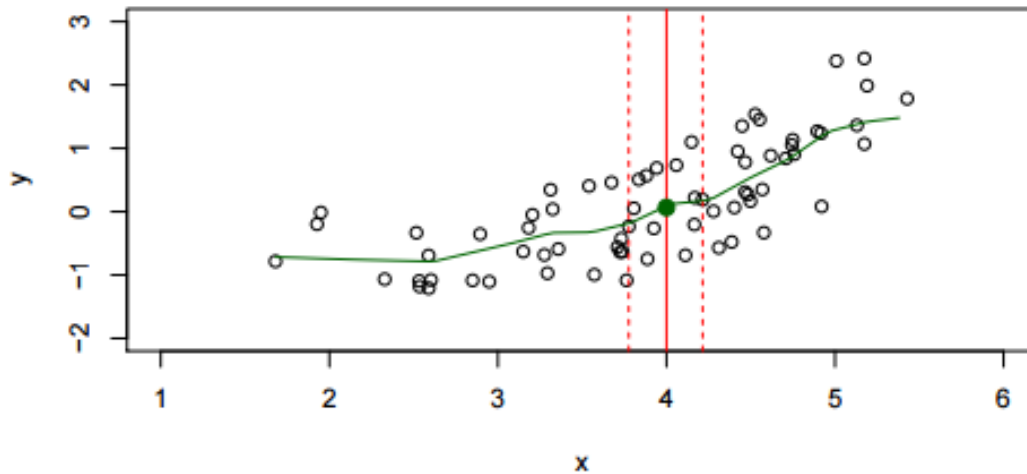
Ideias de estimação

- O valor previsto de $f(x)$ é denominado $\hat{f}(x)$.
- Assim, o erro é dado por $\epsilon = Y - \hat{f}(x)$. Repare que mesmo que conhecermos a real função $f(x)$, ainda teremos erros positivos, visto que para um mesmo X , temos uma distribuição de possíveis valores de Y .
- Em muitos casos, temos poucos ou nenhum valor de Y quando X é igual a 4. O que fazemos neste caso? Relaxamos a suposição e

$$\hat{f}(x) = \text{Ave}(Y|X \in \mathcal{N}(x))$$

em $\mathcal{N}(x)$ representa a região de vizinhança de X .

Método do vizinho mais próximo



- Método do vizinho mais próximo funciona muito bem quando p (número de variáveis) é menor igual a 4.
- Em dimensões grandes, não funciona tão bem devido ao *curse of dimensionality*. Os vizinhos mais próximos tendem a ficar muito longe uns dos outros.

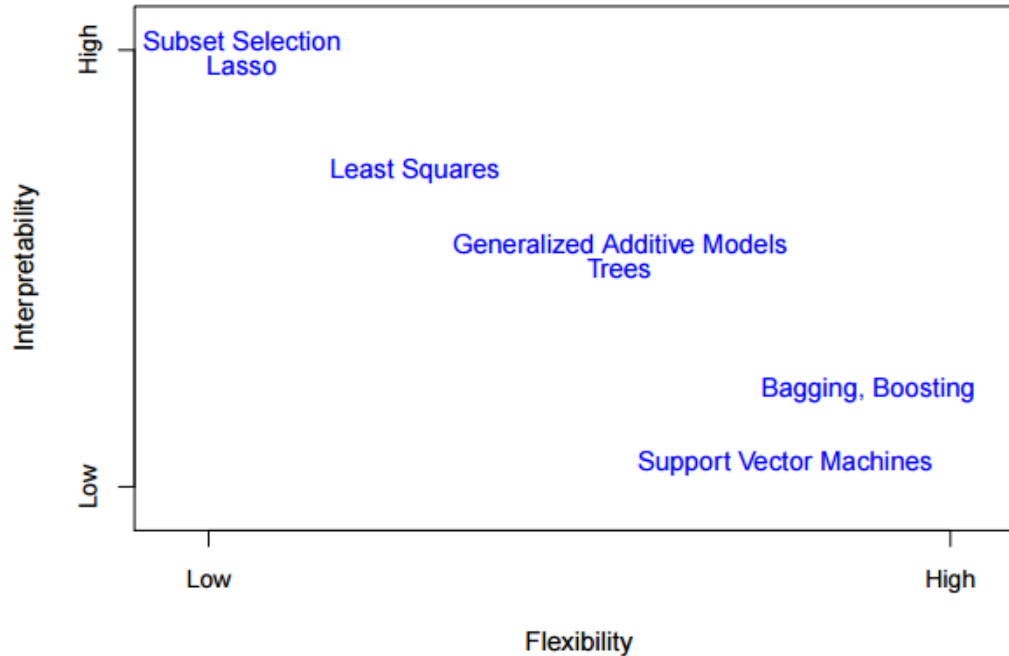
Por que tantos algoritmos de machine learning?

Quais os tradeoffs?

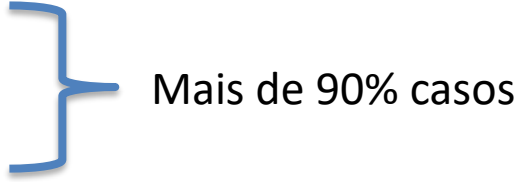
- Não há um algoritmo melhor sempre!!
- Cada algoritmo tende a funcionar melhor em situações específicas.
- Tradeoff clássico entre Interpretabilidade x Acurácia.
- Como evitar overfitting?
- Sempre buscamos parcimônia.

Por que tantos algoritmos de machine learning?

Quais os tradeoffs?



Tipos de Machine Learning

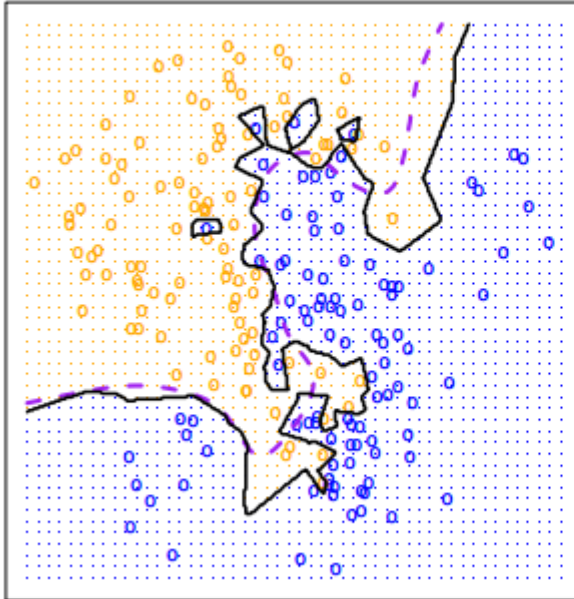
- Supervised Learning
 - Unsupervised Learning
 - Semi-supervised Learning
 - Outros: Reinforcement Learning, sistemas de recomendação.
- 
- Mais de 90% casos

Supervised Learning

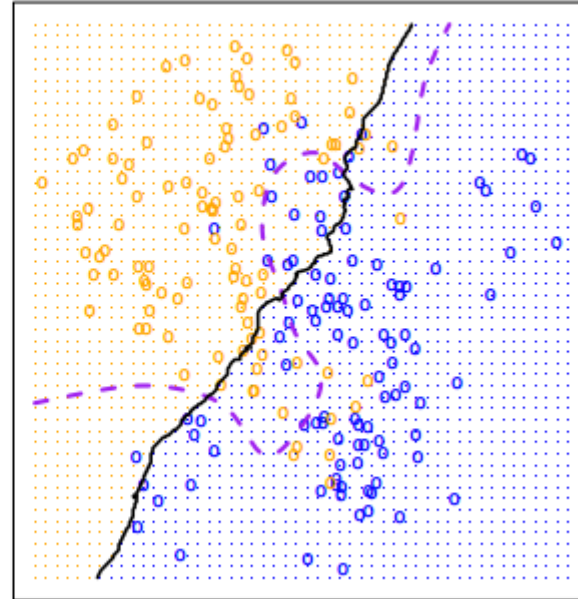
- Variável resposta (target, output) Y é observada.
- P variáveis explicativas (variáveis independentes, features, covariáveis, inputs).
- Se Y é contínua, temos um problema de regressão.
- Em problemas de classificação, Y assume valores finitos não ordenados.
- Dados: $(x_1, y_1), (x_2, y_2), \dots,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

Método de vizinho mais próximo para classificação

KNN: $K=1$



KNN: $K=100$



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!



Supervised Learning

- Regressão Linear, Logística, Naive Bayes, Árvores de decisão, Bagging, Boosting, Random Forest, Redes Neurais, Support Vector Machine, Método do vizinho mais próximo, ...

Unsupervised Learning

- K-means, cluster hierarquico, support vector machine*, redes neurais*, componentes principais, ...

Semi-supervised Learning

- É caro obter a variável resposta de muitas observações.
- Temos alguns dados com variáveis respostas e outros não.
- Semi-supervised learning se encontra entre métodos supervisionados e não supervisionados.

Semi-supervised Learning

Imagens sem resposta!



Carro



Moto

Diferenças entre machine learning e statistical learning

- Há muita intersecção entre os dois, mas:
 - Machine learning foca em performance preditiva.
 - Statistical learning foca nos modelos, na interpretabilidade dos resultados e na incerteza.
- O uso de cada caso depende do problema específico e dos objetivos.

Recomendações

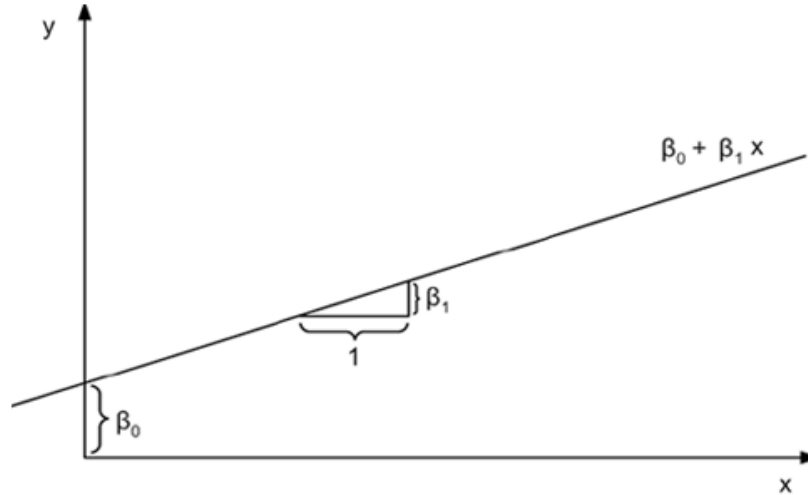
- O que buscamos sempre é a generalização dos resultados. Valide seus resultados!
- Particione seus dados em treino e teste, no mínimo. Algumas técnicas requerem uma base de validação também.
- Entenda os métodos, saiba como eles funcionam!
- Não basta entender as técnicas, você precisa também entender o que está por trás dos dados.

Regressão Linear

- Assumimos que a dependência de Y é linear com X_1, X_2, \dots, X_p .
- A verdadeira relação quase nunca é linear.
- “Essentially, all models are wrong, but some are useful” – George Box.
- Regressão linear se adapta muito bem em diversas aplicações, sendo amplamente utilizada na prática.

Regressão Linear – p=1

- Temos a seguinte previsão: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X_1$.



- β_0 é o intercepto da reta
- β_1 é o coeficiente angular da reta

Regressão Linear – caso geral

- Temos a seguinte previsão:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p.$$

- Como estimamos esses parâmetros?
- Minimizamos a soma de quadrados dos resíduos.
- $RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2$

Regressão Linear

Base simulada com 150 observações e 5 variáveis.

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

Objetivo:

Prever os gastos no cartão de crédito para uma nova observação

```
> 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

Regressão Linear

Descritiva das variáveis

Criação das bases de desenvolvimento e teste

Regressão dos gastos do cartão com apenas uma variável

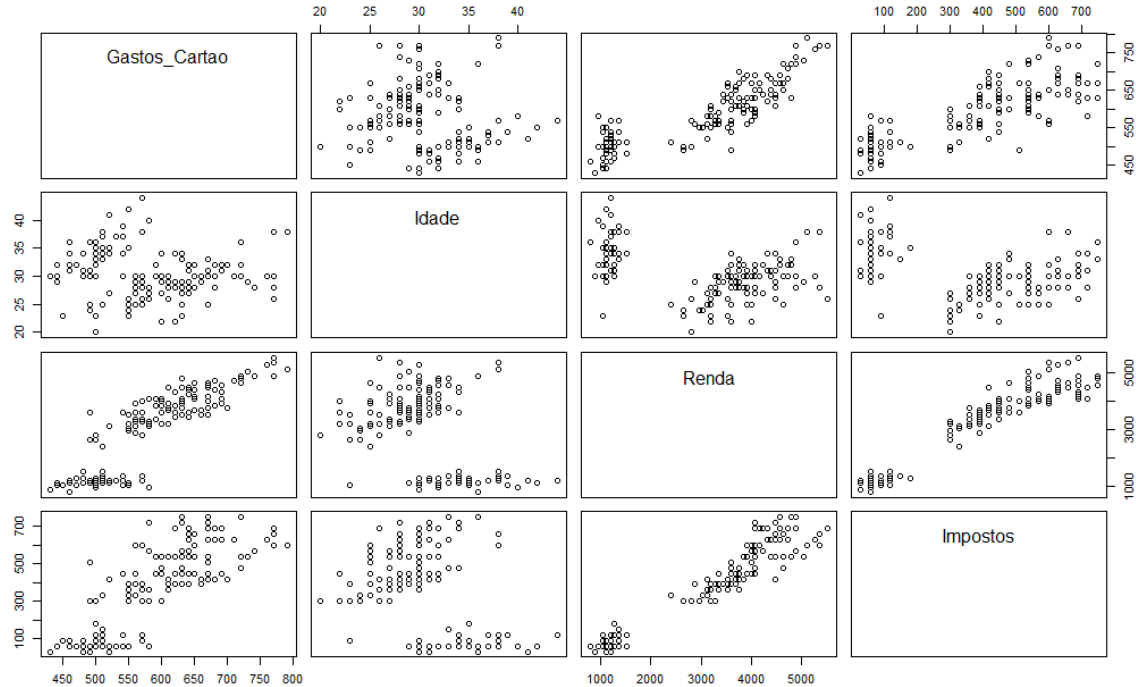
Avaliação dos Resultados

Regressão dos gastos do cartão com várias variáveis

Avaliação dos Resultados

Regressão Linear

Descritiva das variáveis



Regressão Linear

Criação das 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,]
```

Regressão Linear

Regressão dos gastos de cartão pela Renda

```
> fit <- lm(Gastos_Cartao~Renda, data = dados.des)
> summary(fit)
```

call:

```
lm(formula = Gastos_Cartao ~ Renda, data = dados.des)
```

Residuals:

Min	1Q	Median	3Q	Max
-123.68	-26.78	-5.07	24.66	95.04

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	4.212e+02	9.352e+00	45.03	<2e-16 ***
Renda	5.348e-02	2.766e-03	19.33	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 38.93 on 103 degrees of freedom

Multiple R-squared: 0.784, Adjusted R-squared: 0.7819

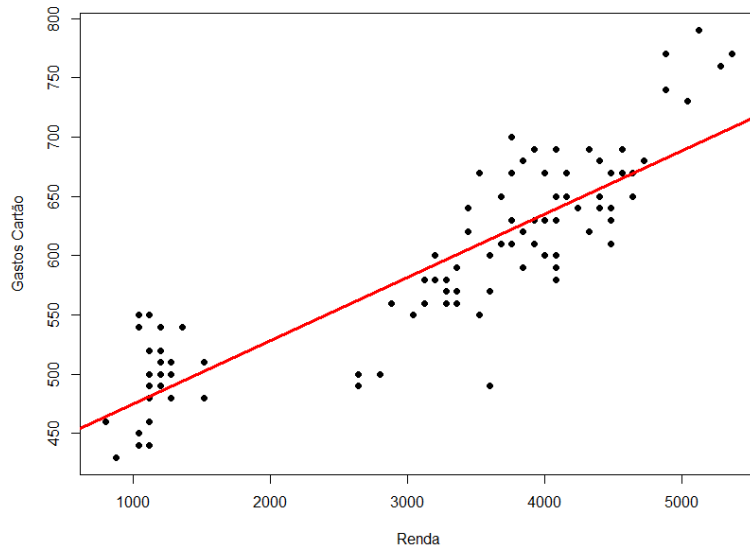
F-statistic: 373.8 on 1 and 103 DF, p-value: < 2.2e-16

```
> fit.val <- predict(fit, newdata=dados.test)
```

Regressão Linear

Avaliação dos resultados

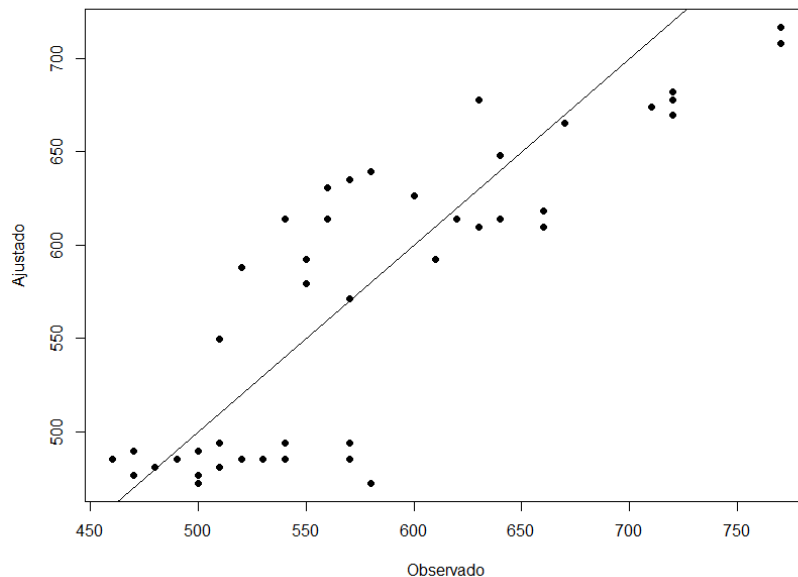
```
> plot(dados.des[, "Renda"], dados.des[, "Gastos_Cartao"], pch=16,  
+       xlab="Renda", ylab="Gastos Cartão")  
> abline(a=fit$coefficients[1], b=fit$coefficients[2], col="red", lwd=3)
```



Regressão Linear

Avaliação dos resultados

```
> plot(dados.test[, "Gastos_Cartao"], fit.val, pch=16,  
+       xlab="Observado", ylab="Ajustado")  
> abline(a=0,b=1)
```



Regressão Linear

Regressão dos gastos do cartão com várias variáveis

```
> fit <- lm(Gastos_Cartao~Idade+Renda+Impostos, data = dados.des)
> summary(fit)
```

Call:
lm(formula = Gastos_Cartao ~ Idade + Renda + Impostos, data = dados.des)

Residuals:

	Min	1Q	Median	3Q	Max
	-85.04	-19.79	0.94	19.39	80.30

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	180.636271	32.815886	5.505	2.82e-07	***
Idade	6.620551	0.874784	7.568	1.82e-11	***
Renda	0.081902	0.008772	9.337	2.59e-15	***
Impostos	-0.129322	0.052156	-2.480	0.0148	*

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 31.38 on 101 degrees of freedom
Multiple R-squared: 0.8623, Adjusted R-squared: 0.8582
F-statistic: 210.9 on 3 and 101 DF, p-value: < 2.2e-16

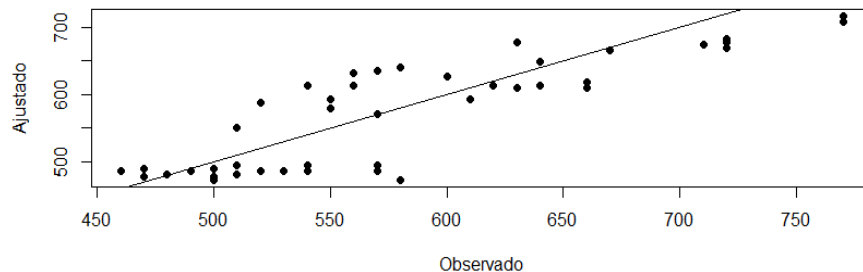
```
> fit.val2 <- predict(fit, newdata=dados.test)
```


Regressão Linear

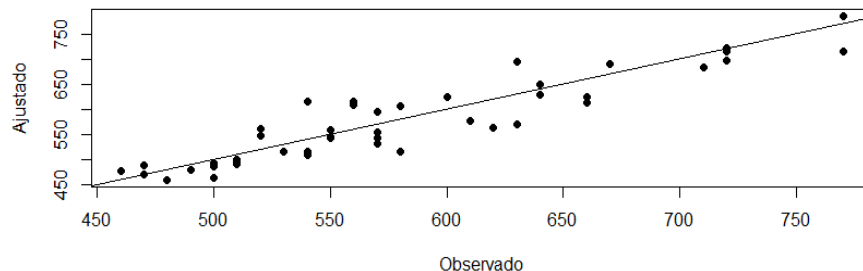
Comparando Resultados

```
> par(mfrow=c(2,1))
> plot(dados.test[, "Gastos_Cartao"], fit.val, pch=16,
+      xlab="Observado", ylab="Ajustado", main="Modelo 1")
> abline(a=0,b=1)
> plot(dados.test[, "Gastos_Cartao"], fit.val2, pch=16,
+      xlab="observado", ylab="Ajustado", main="Modelo 2")
> abline(a=0,b=1)
```

Modelo 1



Modelo 2



Regressão Linear

Comparando Resultados

```
> error.test2 <- dados.test[, "Gastos_Cartao"]-fit.val2
> error.test <- dados.test[, "Gastos_Cartao"]-fit.val
>
> MSE2.test <- sum((error.test2)^2)/nrow(dados.test)
> MAE2.test <- sum(abs(error.test2)/nrow(dados.test))
> MSE1.test <- sum((error.test)^2)/nrow(dados.test)
> MAE1.test <- sum(abs(error.test)/nrow(dados.test))
> MSE1.test
[1] 2037.861
> MSE2.test
[1] 1073.848
> MAE1.test
[1] 38.21734
> MAE2.test
[1] 26.78094
```

Regressão Linear

Exercícios:

<http://archive.ics.uci.edu/ml/datasets/Wholesale+customers>

Base de dados (“wholesales_customers_data.csv”) de um supermercado português com o faturamento anual em diversas categorias (produtos frescos, leite, mercearia, congelados, produtos de limpeza, guloseimas), além da região e o tipo de loja de 440 lojas.

1.) Altere a codificação das variáveis região (Region) e tipo de loja (Channel)

Channel==1 → Horeca (Hotel/Restaurant/Café)

Channel==2 → Retail

Region==1 → Lisboa

Region==2 → Porto

Region==3 → Outro

Regressão Linear

2.) Construa um modelo de regressão linear prevendo o faturamento de leite com outra variável explicativa contínua. Escolha a melhor variável realizando com base na análise descritiva. Interprete os resultados.

Separe seus dados em desenvolvimento (80%) e teste (20%). Use seed de 121.

Construa o histograma dos resíduos para as duas bases.

Calcule o erro quadrático médio e absoluto. Interprete-os

3.) Construa um modelo de regressão linear prevendo o faturamento de leite em relação a todas variáveis explicativas disponíveis.

Interprete os resultados.

Construa o histograma dos resíduos para as duas bases.

Calcule o erro quadrático médio e absoluto. Interprete-os.

Regressão Linear

4.) Acrescente as variáveis de região e tipo de loja no modelo. Elas acrescentam algo em previsibilidade?

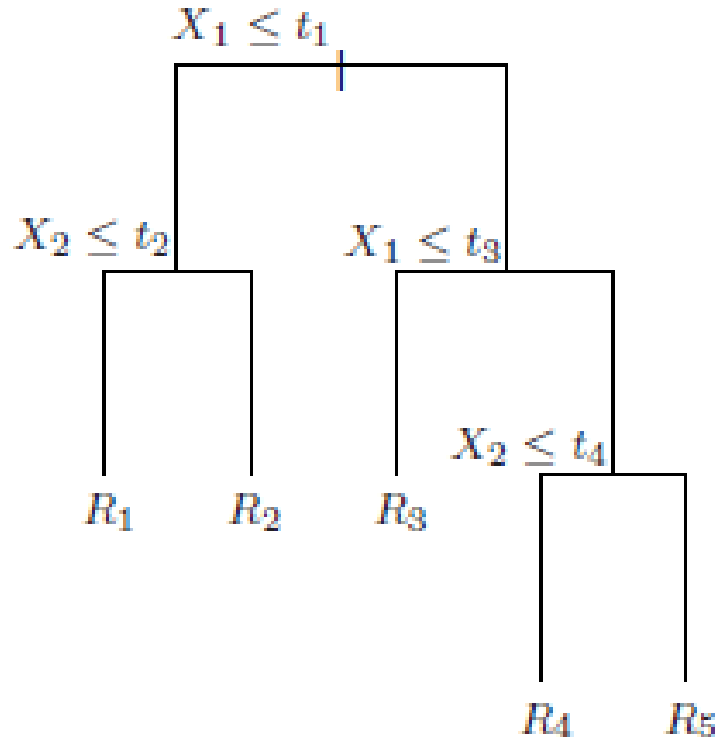
Interprete os resultados.

Construa o histograma dos resíduos para as duas bases.

Calcule o erro quadrático médio e absoluto. Interprete-os.

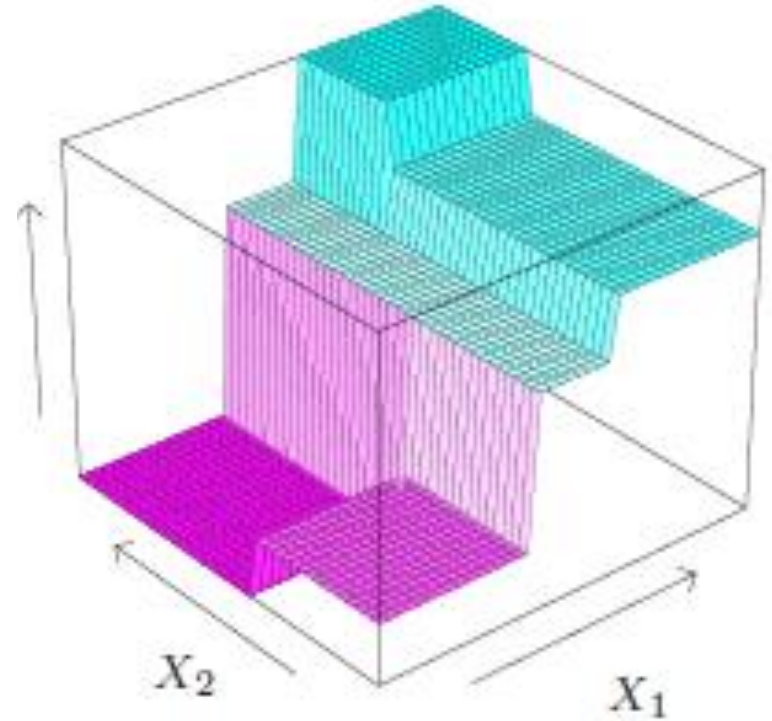
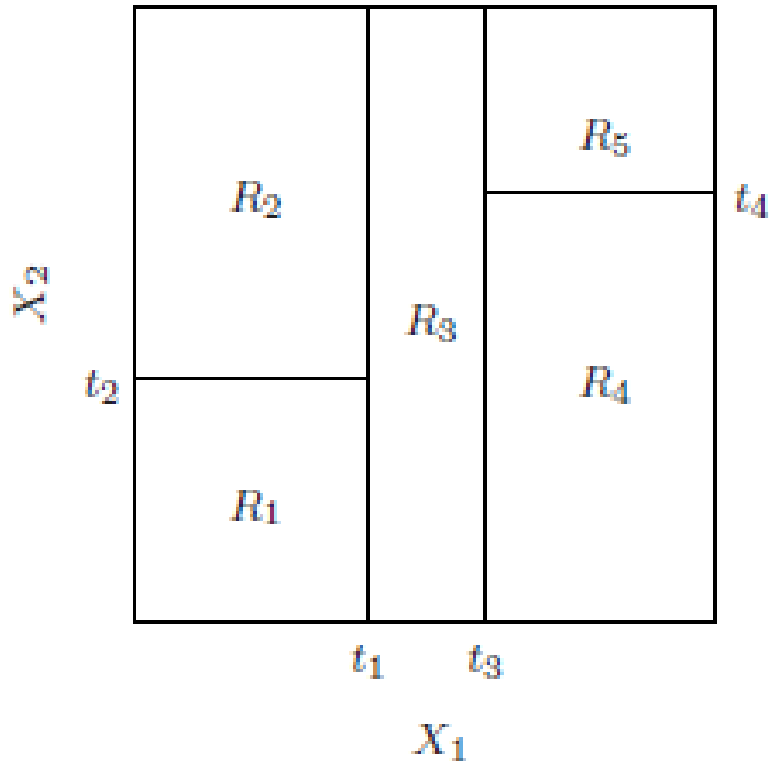
Árvore de decisão

Árvore de decisão



- Método bastante popular por ser facilmente interpretado.
- Popular na área médica, pois imitam a forma de pensar dos médicos.
- Pode ser feita tanto para classificação quanto para regressão.
- Consiste em dividir o seu espaço em retângulos.

Árvore de decisão



Árvore de decisão

Como criamos a árvore? Como proceder? O que ela deve fazer?

Árvore de decisão

Em regressão, para cada quebra da árvore, criamos duas regiões, R_1 e R_2 . Temos que escolher a variável X_j e o ponto s que minimizam:

$$\min_{j,s} \left(\sum_{i \in R_1} (y_i - \overline{R_1})^2 + \sum_{i \in R_2} (y_i - \overline{R_2})^2 \right)$$

Para classificação, alteramos a medida de erro utilizada na minimização.

Árvore de decisão

Como paramos de crescer a árvore?

O sugerido é crescer a árvore até o máximo (número de observação em cada nó final menor do que 5) e depois podar a árvore.

Seja N_m o número de observação no m -ésimo nó final, $|T|$ o número de nós finais e seja Q_m o erro quadrático médio do m -ésimo nó final.

Assim, a ideia é encontrar a sub-árvore que minimize a seguinte medida.

$$C_\alpha(T) = \sum_{m=1}^{|T|} N_m Q_m(T) + \alpha |T|.$$

O parâmetro α controla o tradeoff entre o tamanho da árvore e a performance preditiva.

Árvore de decisão

Laboratório R - Base de Spam

- Base com 4.601 e-mails. Porcentual em que 54 palavras ou pontuações aparecem em cada e-mail. Além disso, temos o tamanho médio das palavras, tamanho da maior palavra e quantidade de palavras.

Objetivo:

Criar um detector automático de SPAM que verificará cada novo e-mail.

- Disponível em: <https://archive.ics.uci.edu/ml/datasets/Spambase>

Árvore de decisão

Laboratório R – Base de Spam

Carregando base e
definindo nomes das
colunas

Criando bases de
desenvolvimento,
validação e teste

Treinando e visualizando
uma árvore de decisão

Otimizando a
profundidade

Avaliando os resultados

Árvore de decisão

Carregando base e definindo nomes das colunas

```
> dados = read.table("spambase.data", sep=";", header=F)
> nomes = c("word_freq_make", "word_freq_address", "word_freq_all", "word_freq_3d",
+           "word_freq_our", "word_freq_over", "word_freq_remove", "word_freq_internet",
+           "word_freq_order", "word_freq_mail", "word_freq_receive", "word_freq_will",
+           "word_freq_people", "word_freq_report", "word_freq_addresses", "word_freq_free",
+           "word_freq_business", "word_freq_email", "word_freq_you", "word_freq_credit",
+           "word_freq_your", "word_freq_font", "word_freq_000", "word_freq_money",
+           "word_freq_hp", "word_freq_hpl", "word_freq_george", "word_freq_650",
+           "word_freq_lab", "word_freq_labs", "word_freq_telnet", "word_freq_857",
+           "word_freq_data", "word_freq_415", "word_freq_85", "word_freq_technology",
+           "word_freq_1999", "word_freq_parts", "word_freq_pm", "word_freq_direct",
+           "word_freq_cs", "word_freq_meeting", "word_freq_original", "word_freq_project",
+           "word_freq_re", "word_freq_edu", "word_freq_table", "word_freq_conference",
+           "char_freq_pvir", "char_freq_par", "char_freq_bra", "char_freq_exc",
+           "char_freq_dolar", "char_freq_num", "capital_run_length_average",
+           "capital_run_length_longest", "capital_run_length_total", "SPAM")
> names(dados) = nomes
```

Árvore de decisão

Criando bases de desenvolvimento, validação e teste

```
> set.seed(432)
> id <- sample(1:nrow(dados), nrow(dados)*0.8)
> id.des <- sample(id, nrow(dados)*0.7)
> id.val <- id[!(id %in% id.des)]
> dados.des <- dados[id.des,]
> dados.val <- dados[id.val,]
> dados.test <- dados[-id,]
```

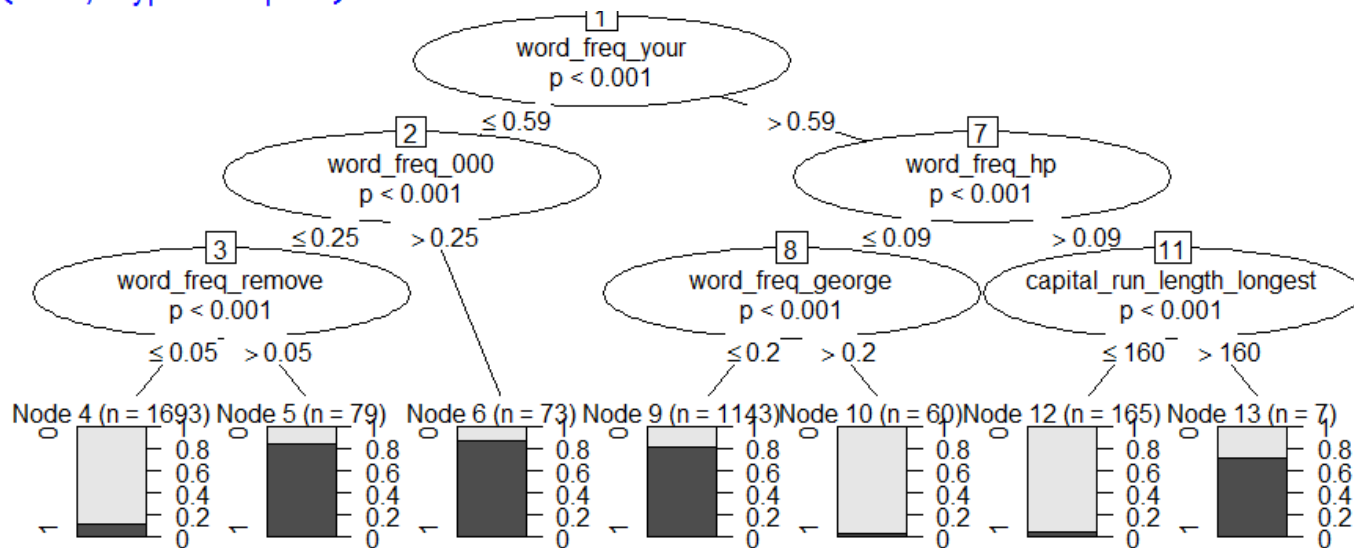
Por que precisamos de uma base de validação neste caso?

- Usamos esta base de validação para otimizarmos alguns parâmetros da árvore de decisão.

Árvore de decisão

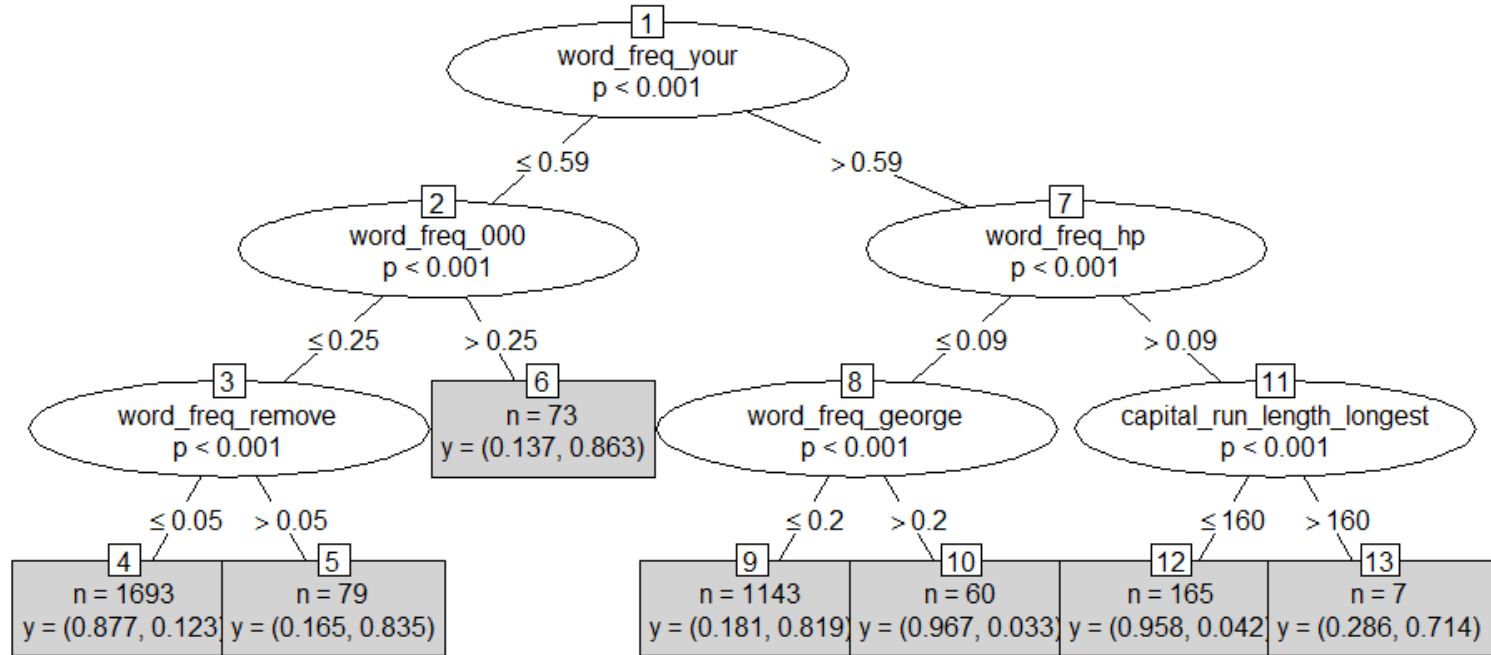
Treinando e visualizando uma árvore de decisão

```
> formula = paste0("as.factor(", nomes[58], ") ~ ", paste0(nomes[1:57], collapse="+"))  
> tree = ctree(as.formula(formula), data=dados.des, controls = ctree_control(maxdepth = 3))  
> plot(tree)  
> plot(tree, type="simple")
```



Árvore de decisão

Treinando e visualizando uma árvore de decisão



Árvore de decisão

Otimizando a profundidade

```
> tree = ctree(as.formula(formula), data=dados.des, controls = ctree_controls(maxdepth = 3))
> fit.val.prof3 = predict(tree, newdata=dados.val)
> sum(ifelse(fit.val.prof3!=dados.val[, "SPAM"],1,0))/nrow(dados.val)
[1] 0.15
> tree10 = ctree(as.formula(formula), data=dados.des, controls = ctree_controls(maxdepth = 10))
> fit.val.prof10 = predict(tree10, newdata=dados.val)
> sum(ifelse(fit.val.prof10!=dados.val[, "SPAM"],1,0))/nrow(dados.val)
[1] 0.09347826
> tree11 = ctree(as.formula(formula), data=dados.des, controls = ctree_controls(maxdepth = 11))
> fit.val.prof11 = predict(tree11, newdata=dados.val)
> sum(ifelse(fit.val.prof11!=dados.val[, "SPAM"],1,0))/nrow(dados.val)
[1] 0.09565217
```

Árvore de decisão

Avaliando os resultados

```
> fit.des = predict(tree10, newdata=dados.des)
> fit.val = predict(tree10, newdata=dados.val)
> fit.test = predict(tree10, newdata=dados.test)
>
> sum(ifelse(fit.des!=dados.des[, "SPAM"],1,0))/nrow(dados.des)
[1] 0.07267081
> sum(ifelse(fit.val!=dados.val[, "SPAM"],1,0))/nrow(dados.val)
[1] 0.09347826
> sum(ifelse(fit.test!=dados.test[, "SPAM"],1,0))/nrow(dados.test)
[1] 0.09120521
```

Árvore de decisão

Exercícios: <http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29>



Breast Cancer Wisconsin (Diagnostic) Data Set

Download: [Data Folder](#), [Data Set Description](#)

Abstract: Diagnostic Wisconsin Breast Cancer Database



Data Set Characteristics:	Multivariate	Number of Instances:	569	Area:	Life
Attribute Characteristics:	Real	Number of Attributes:	32	Date Donated	1995-11-01
Associated Tasks:	Classification	Missing Values?	No	Number of Web Hits:	390512

Árvore de decisão

Base de dados (“cancer.data”) com 699 observações e 10 variáveis de pacientes com tumores. O objetivo é detectar com base em algumas informações dos tumores se é benigno ou maligno.

Remova os dados missing por meio do comando “na.omit”. Poderíamos fazer algo melhor?

Variáveis:

- 1. Sample code number id number
- 2. Clump Thickness 1 - 10
- 3. Uniformity of Cell Size 1 - 10
- 4. Uniformity of Cell Shape 1 – 10
- 5. Marginal Adhesion 1 - 10
- 6. Single Epithelial Cell Size 1 - 10
- 7. Bare Nuclei 1 - 10
- 8. Bland Chromatin 1 - 10
- 9. Normal Nucleoli 1 - 10
- 10. Mitoses 1 - 10
- 11. Class: (2 for benign, 4 for malignant)

Árvore de decisão

- 1.) Crie bases de desenvolvimento (60%), validação (20%) e teste (20%). Utilize seed de 432. Monte a formula para ser usado nos modelos com todas as variáveis explicativas.
- 2.) Construa uma árvore com profundidade 3. Interprete esta árvore.
- 3.) Crie um algoritmo que construa árvores com profundidade 1 à 15. Qual profundidade você escolhe? Como chegou neste critério?
- 4.) Avalie os resultados da “melhor” árvore em todas as bases.

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-Ismael, M., Lin, H.T (2012) “Learning from data”
- Theodoridis, S., Koutroumbas, K., (2008) “Pattern Recognition”
- Kuhn, M., Johnson, K., (2013) “Applied Predictive Modeling”
- Burns, P. (2011) “The R inferno”