Introdução a machine learning com aplicações em R

Aula 5 - Extensão para aprendizado supervisionado

Felipe Barletta

Departamento de Estatística

12 novembro, 2020

Sumário

Etapas em um ajuste de algoritmos de machine learning

1. **Definição dos objetivos:** trata-se do primeiro passo de qualquer análise.
2. **Coleta dos dados:** envolve a coleta de material que o algoritmo utilizará para gerar conhecimento;
3. **Exploração e preparação dos dados:** é exigido um trabalho adicional na preparação desses, recodificando-os de acordo com os *inputs* esperados;
4. **Formação do modelo:** depois dos dados preparados, o pesquisador já é capaz de dizer o que é possível aprender deles, e como;
5. **Avaliação dos modelos:** avaliamos a qualidade do aprendizado, não pode ser pouco (*underfitting*) nem decorar os dados (*overfitting*);
6. **Melhoria do modelo:** se necessário, podemos melhorar o desempenho do modelo através de estratégias avançadas;
7. **Comunicação e automazição:** relatórios com os resultados e colocar em produção o modelo final (*deploy*).

Etapas em um ajuste de algoritmos de machine learning

1. **Definição dos objetivos:** trata-se do primeiro passo de qualquer análise.
2. **Coleta dos dados:** envolve a coleta de material que o algoritmo utilizará para gerar conhecimento;
3. **Exploração e preparação dos dados:** é exigido um trabalho adicional na preparação desses, recodificando-os de acordo com os *inputs* esperados;
4. **Formação do modelo:** depois dos dados preparados, o pesquisador já é capaz de dizer o que é possível aprender deles, e como;
5. **Avaliação dos modelos:** avaliamos a qualidade do aprendizado, não pode ser pouco (*underfitting*) nem decorar os dados (*overfitting*);
6. **Melhoria do modelo:** se necessário, podemos melhorar o desempenho do modelo através de estratégias avançadas;
7. **Comunicação e automazição:** relatórios com os resultados e colocar em produção o modelo final (*deploy*).

Etapas em um ajuste de algoritmos de machine learning

1. **Definição dos objetivos:** trata-se do primeiro passo de qualquer análise.
2. **Coleta dos dados:** envolve a coleta de material que o algoritmo utilizará para gerar conhecimento;
3. **Exploração e preparação dos dados:** é exigido um trabalho adicional na preparação desses, recodificando-os de acordo com os *inputs* esperados;
4. **Formação do modelo:** depois dos dados preparados, o pesquisador já é capaz de dizer o que é possível aprender deles, e como;
5. **Avaliação dos modelos:** avaliamos a qualidade do aprendizado, não pode ser pouco (*underfitting*) nem decorar os dados (*overfitting*);
6. **Melhoria do modelo:** se necessário, podemos melhorar o desempenho do modelo através de estratégias avançadas;
7. **Comunicação e automazição:** relatórios com os resultados e colocar em produção o modelo final (*deploy*).

Etapas em um ajuste de algoritmos de machine learning

1. **Definição dos objetivos:** trata-se do primeiro passo de qualquer análise.
2. **Coleta dos dados:** envolve a coleta de material que o algoritmo utilizará para gerar conhecimento;
3. **Exploração e preparação dos dados:** é exigido um trabalho adicional na preparação desses, recodificando-os de acordo com os *inputs* esperados;
4. **Formação do modelo:** depois dos dados preparados, o pesquisador já é capaz de dizer o que é possível aprender deles, e como;
5. **Avaliação dos modelos:** avaliamos a qualidade do aprendizado, não pode ser pouco (*underfitting*) nem decorar os dados (*overfitting*);
6. **Melhoria do modelo:** se necessário, podemos melhorar o desempenho do modelo através de estratégias avançadas;
7. **Comunicação e automazição:** relatórios com os resultados e colocar em produção o modelo final (*deploy*).

Etapas em um ajuste de algoritmos de machine learning

1. **Definição dos objetivos:** trata-se do primeiro passo de qualquer análise.
2. **Coleta dos dados:** envolve a coleta de material que o algoritmo utilizará para gerar conhecimento;
3. **Exploração e preparação dos dados:** é exigido um trabalho adicional na preparação desses, recodificando-os de acordo com os *inputs* esperados;
4. **Formação do modelo:** depois dos dados preparados, o pesquisador já é capaz de dizer o que é possível aprender deles, e como;
5. **Avaliação dos modelos:** avaliamos a qualidade do aprendizado, não pode ser pouco (*underfitting*) nem decorar os dados (*overfitting*);
6. **Melhoria do modelo:** se necessário, podemos melhorar o desempenho do modelo através de estratégias avançadas;
7. **Comunicação e automazição:** relatórios com os resultados e colocar em produção o modelo final (*deploy*).

Etapas em um ajuste de algoritmos de machine learning

1. **Definição dos objetivos:** trata-se do primeiro passo de qualquer análise.
2. **Coleta dos dados:** envolve a coleta de material que o algoritmo utilizará para gerar conhecimento;
3. **Exploração e preparação dos dados:** é exigido um trabalho adicional na preparação desses, recodificando-os de acordo com os *inputs* esperados;
4. **Formação do modelo:** depois dos dados preparados, o pesquisador já é capaz de dizer o que é possível aprender deles, e como;
5. **Avaliação dos modelos:** avaliamos a qualidade do aprendizado, não pode ser pouco (*underfitting*) nem decorar os dados (*overfitting*);
6. **Melhoria do modelo:** se necessário, podemos melhorar o desempenho do modelo através de estratégias avançadas;
7. **Comunicação e automazição:** relatórios com os resultados e colocar em produção o modelo final (*deploy*).

Etapas em um ajuste de algoritmos de machine learning

1. **Definição dos objetivos:** trata-se do primeiro passo de qualquer análise.
2. **Coleta dos dados:** envolve a coleta de material que o algoritmo utilizará para gerar conhecimento;
3. **Exploração e preparação dos dados:** é exigido um trabalho adicional na preparação desses, recodificando-os de acordo com os *inputs* esperados;
4. **Formação do modelo:** depois dos dados preparados, o pesquisador já é capaz de dizer o que é possível aprender deles, e como;
5. **Avaliação dos modelos:** avaliamos a qualidade do aprendizado, não pode ser pouco (*underfitting*) nem decorar os dados (*overfitting*);
6. **Melhoria do modelo:** se necessário, podemos melhorar o desempenho do modelo através de estratégias avançadas;
7. **Comunicação e automazição:** relatórios com os resultados e colocar em produção o modelo final (*deploy*).

Etapas em um ajuste de algoritmos de machine learning

1. **Definição dos objetivos:** trata-se do primeiro passo de qualquer análise.
2. **Coleta dos dados:** envolve a coleta de material que o algoritmo utilizará para gerar conhecimento;
3. **Exploração e preparação dos dados:** é exigido um trabalho adicional na preparação desses, recodificando-os de acordo com os *inputs* esperados;
4. **Formação do modelo:** depois dos dados preparados, o pesquisador já é capaz de dizer o que é possível aprender deles, e como;
5. **Avaliação dos modelos:** avaliamos a qualidade do aprendizado, não pode ser pouco (*underfitting*) nem decorar os dados (*overfitting*);
6. **Melhoria do modelo:** se necessário, podemos melhorar o desempenho do modelo através de estratégias avançadas;
7. **Comunicação e automazição:** relatórios com os resultados e colocar em produção o modelo final (*deploy*).

Melhoria do modelo

1. **Entenda os dados:** explore as características, crie gráficos para conhecer a natureza das variáveis etc.
2. **Decida sobre a validação cruzada:** uma boa estratégia de validação, garante resultados mais confiáveis.
3. **Feature Engineering:** tente aprimorar a predição dos modelos através da engenharia de características.
4. **Combine modelos:** agrupe vários métodos, certificando-se que são correlacionados.

Extreme Gradient Boosting - XGBoost

- Mesma essência do ***GBM***.
- **Computação em paralelo:** por *default*, ele utiliza todos os cores da máquina;
- **Regularização**: controla o *trade-off* entre vício e variância, permitindo selecionar variáveis etc.
- ***Cross validation***: já possui internamente esse recurso. Não necessitando funções externas.
- ***Missing Values***: se existir alguma tendência nos dados faltantes, o algoritmo captará.
- ***Save and Reload***: permite salvar a matriz de dados e o modelo, podendo recarregá-lo em outro momento.

XGBoost - Hiperparâmetros

- `nrounds`: Controla o número de iterações. Para classificação, é similar ao número de árvores.
- `eta`: Controla a taxa de aprendizado. Tipicamente, utilizamos valores entre 0,01 e 0,3.
- `max_depth`: Controla o tamanho de cada árvore. Geralmente, utilizamos árvores menores, para evitar o superajuste(*overfitting*).
- `subsample`: Controla o tamanho da amostra em cada árvore.
- `colsample_bytree`: Controla o número de variáveis apresentadas à árvore

One-hot-Encoding

- *one-hot-encoding* significa em transformar características categóricas em binárias.
- número 1 representa o valor afirmativo e o 0 negativo

id	color
1	red
2	blue
3	green
4	blue

id	color_red	color_blue	color_green
1	1	0	0
2	0	1	0
3	0	0	1
4	0	1	0

Local Interpretable Model-Agnostic Explanations - (LIME)

- Representação interpretável para uma classificação local.
- ***Local***: Refere-se à fidelidade da predição local - ou seja, queremos que a explicação realmente reflita o comportamento do classificador "em torno" da instância que está sendo prevista.
- ***Interpretable***: Cria uma interpretação humana a predição.
- ***Model-agnostic***: Dá um diagnóstico do modelo, independente de qual é o modelo.

$$LIME(x) = \underset{g \in G}{\operatorname{argmin}}\, \Psi(f, g, \pi_x) + \Omega(g)$$

em que G é uma classe de modelo interpretável e π_x é a localidade definida.

SHapley Additive exPlanation - SHAP

- É um método de interpretação de predição individual
- É uma abordagem baseada na teoria dos jogos
- Quanto cada valor da característica contribui para a predição compara a predição média?

$$g(z') = \phi_0 + \sum_{j=1}^{M} \phi_j z'_j$$

em que $\phi_0 = E[\hat{f}(x)]$ e $\phi_j \in \mathbb{R}$, então:

$$I_j = \sum_{i=1}^{n} |\phi_j^{(i)}|$$

Linear Models (LMs)

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad ; \boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2)$$

$$\boldsymbol{X} = [\mathbf{1} : x_1 : x_2 : ... : x_p]$$

$$\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)'$$

$$Y_i = \mu_i + \epsilon_i; \mu_i = X_i\beta_p$$

$$\mu_i = \mathbb{E}(Y_i)$$

Generalized Linear Models (GLMs)

$$\boldsymbol{g(\mu) = \eta = X\beta},$$

em que $\mathbb{E}[Y_i] = \mu_i = g^{-1}(\eta)$.

- Distribuição Poisson, a função de ligação $g(.)$, é o `log` natural.
  - $\log \mu = \eta$
- Distribuição Binomial é o `logit`.
  - $p_i = \dfrac{1}{1 + \exp(X_i\beta_p)}$
  - $g(\mu) = \eta = \texttt{logit}(p_i) = \log(\dfrac{p_i}{1 - p_i}) = X_i\beta_p$
- Distribuição Gaussiana é a identidade.
  - $g(\mu) = \mu = \eta$

Generalized Additive Models (GAMs)

$$g(\mu) = \eta = \beta_0 + X\beta + f(x_1) + f(x_2) + ... + f(x_p) + f(x_1, x_2) + ... + f(x_{p-1}, x_p)$$

- Regressão via *Splines*
- Método para aproximar funções(Interpolação)

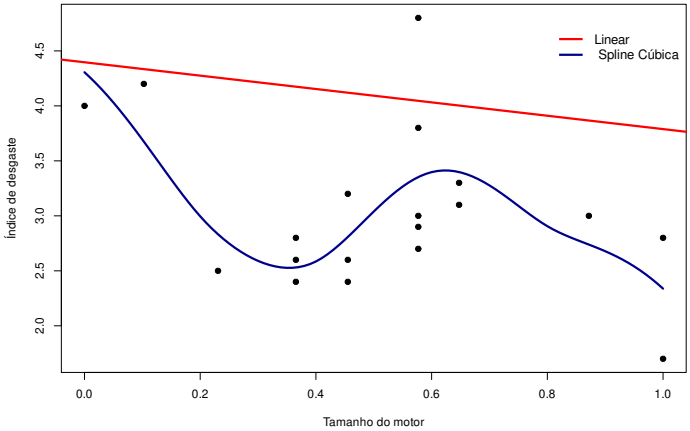
$$y_i = S(x_i) + \epsilon_i$$

- Função por partes (partição do intervalo$[a, b]$)

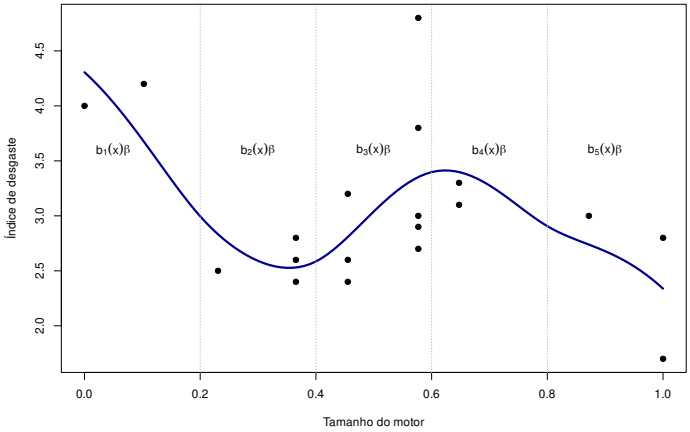
$$S(x) = \sum b_j(x)\beta_j$$

em que $b_j(x)$ é a j-ésima base da função.

Exemplo


Linear
Spline Cúbica
Índice de desgaste
Tamanho do motor
2.0
2.5
3.0
3.5
4.0
4.5
0.0
0.2
0.4
0.6
0.8
1.0

Exemplo


Índice de desgaste
Tamanho do motor
b1(x)β
b2(x)β
b3(x)β
b4(x)β
b5(x)β
2.0
2.5
3.0
3.5
4.0
4.5
0.0
0.2
0.4
0.6
0.8
1.0

Weight of evidence ou peso de evidência - WoE

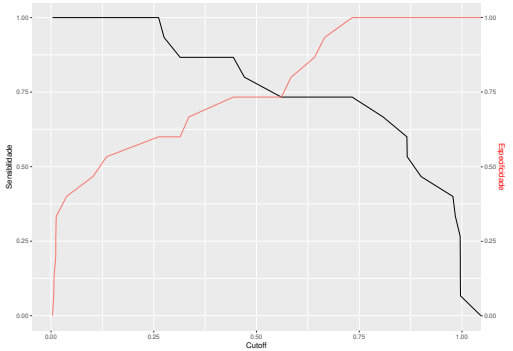
- São pontuações que descrevem a probabilidade da classe de interesse Y que observa-se em cada faixa de valores.
- Pode ser aplicada se tivermos uma variável Y binária e qualquer tipo de variável preditora.

$$WoE = \log\left(\frac{Dist_{Rel}.Positiva_i}{Dist_{Rel}.Negativa_i}\right)$$

- Valor WoE positivo, significa que a distribuição positiva é maior que a distribuição negativa.
- Valor WoE negativo, significa que a distribuição positiva é menor que a distribuição negativa.

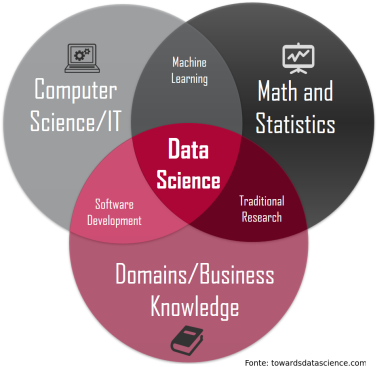
Ponto de corte *(cutoff ou threshold)*

- Modelos de classificação binário - *default* $= 0,5$
- Encontrar um novo ponto de corte que estabilize as estimativas de erro do modelo
- Análise útil para dados desbalanceados



Comunicação

- Capacidade de comunicação visual, oral e escrita.
- Narrativa está atrelada aos dados.
- Por meio de relatórios com gráficos e tabelas.
- Forma um enredo que atrai e cativa a todos.
- *Data Storytelling*.



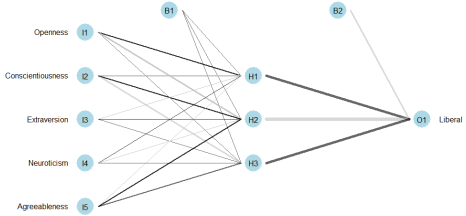
Fonte: towardsdatascience.com

Relatórios dinâmicos

1. `.Rnw`
   - Entrelaçamento de código *LaTeX* com `R`.
2. `.Rmd`
   - Entrelaçamento de código *Markdown* com `R` ou `python`.
3. `jupyter-notebook`
   - Entrelaçamento de código *Markdown* com `R` ou `python`.
4. `Shiny`
   - Interface interativa do `R`.

Redes Neurais Artificiais - RNA

1. Aprendizado não-supervisionado
   - *Clustering*
2. Aprendizado supervisionado
   - Classificação
   - Regressão
3. Séries Temporais
   - Previsão de valores futuros. Por ex. *Long short-term memory - LSTM*

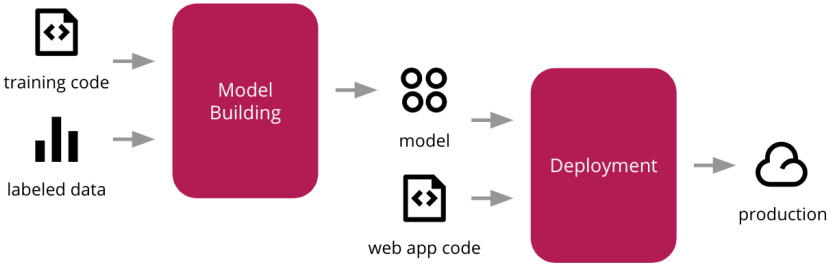


Extensões do Gradient Boosting

- https://lightgbm.readthedocs.io/en/latest/

- https://catboost.ai/docs/concepts/about.html

Deploy - Colocar o modelo em produção



Desafio 5

a. Com os gerados no desafio 4, tente melhorar suas predições com as técnicas apresentadas na aula de hoje

b. Faça uma tabela com as novas métricas de desempenho comparando o modelo antigo e novo.

c. Faça um breve relatório contando a história dos seus dados e os resultados do modelo que escolheu.

Referências

1. Hastie, T., Tibshirani, R. e Friedman, J., **The Elements of Statistical Learning**, 2009.
2. James, G., Witten, D., Hastie, T. e Tibshirani, **An Introduction to Statistical Learning**, 2013.
3. L. Breiman. Statistical modeling: **The two cultures. Statistical Science**, 16(3):199-231, 2001.
4. Lantz, B., **Machine Learning with R**, Packt Publishing, 2013.
5. Tan, Steinbach, and Kumar, **Introduction to Data Mining**, Addison-Wesley, 2005.
6. Mitchell, T. M. (1997). **Machine Learning**. McGraw-Hill.
7. Wickham, H.; Grolemund, G. **R for data science: import, tidy, transform, visualize, and model data**. " O'Reilly Media, Inc.", 2016.


multumese
camonban
mochchakkeram
gracias
merci
kiitos
dankedir
chokrane
obrigado
thankyou
arigato
asante
welalin
grazie
maturnuwun
termakasih