Introdução a machine learning com aplicações em R

Aula 3 - Tipos de aprendizados

Felipe Barletta

Departamento de Estatística

05 novembro, 2020

Sumário

Como as máquinas aprendem?

1. **Armazenamento dos dados:** utiliza a observação para fornecer uma base para o raciocínio adicional;
2. **Abrstração:** envolve a tradução dos dados armazenados em representações e conceitos;
3. **Generalização:** cria conhecimento e inferência que direcionam ações em novos contextos;
4. **Avaliação:** fornece um mecanismo de feedback para medir a utilidade do conhecimento adquirido e informar potenciais melhorias.

Como as máquinas aprendem?

1. **Armazenamento dos dados:** utiliza a observação para fornecer uma base para o raciocínio adicional;
2. **Abrstração:** envolve a tradução dos dados armazenados em representações e conceitos;
3. **Generalização:** cria conhecimento e inferência que direcionam ações em novos contextos;
4. **Avaliação:** fornece um mecanismo de feedback para medir a utilidade do conhecimento adquirido e informar potenciais melhorias.

Como as máquinas aprendem?

1. **Armazenamento dos dados:** utiliza a observação para fornecer uma base para o raciocínio adicional;
2. **Abrstração:** envolve a tradução dos dados armazenados em representações e conceitos;
3. **Generalização:** cria conhecimento e inferência que direcionam ações em novos contextos;
4. **Avaliação:** fornece um mecanismo de feedback para medir a utilidade do conhecimento adquirido e informar potenciais melhorias.

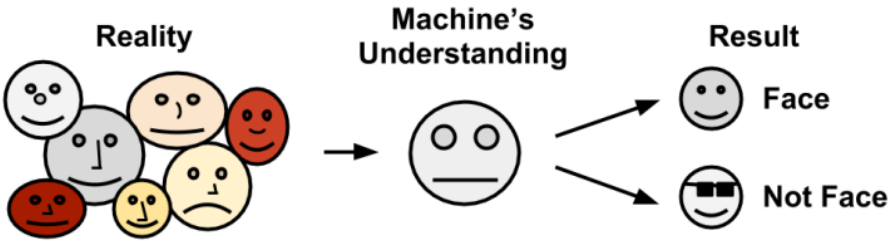
Como as máquinas aprendem?

1. **Armazenamento dos dados:** utiliza a observação para fornecer uma base para o raciocínio adicional;
2. **Abrstração:** envolve a tradução dos dados armazenados em representações e conceitos;
3. **Generalização:** cria conhecimento e inferência que direcionam ações em novos contextos;
4. **Avaliação:** fornece um mecanismo de feedback para medir a utilidade do conhecimento adquirido e informar potenciais melhorias.

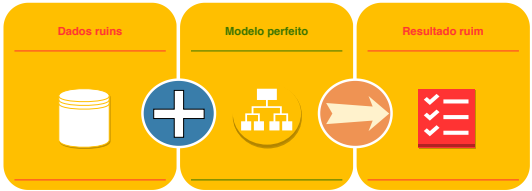
Exemplo: Abstração

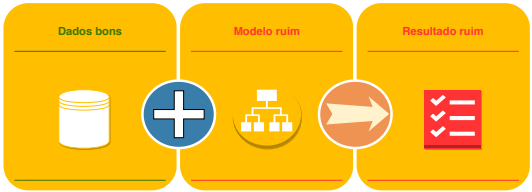
- http://collections.lacma.org/node/239578

Exemplo: Generalização

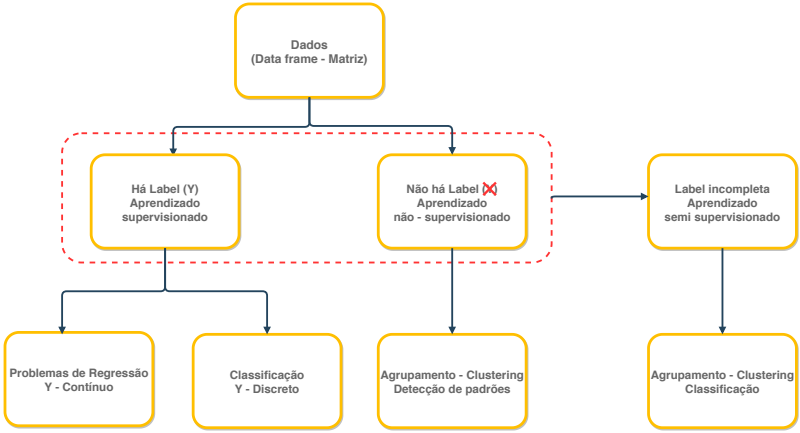


Aspectos dos dados





Tipos de aprendizado



Métodos de reamostragem

- **Treinamento** - Aprende com os exemplos apresentados ao algoritmo;
- **Validação** - Valida-se ou simula-se o aprendizado com exemplos em que se conhece a respota - Como um simulado do vestibular
- **Teste** - Testa o conhecimento obtido no treinamento com exemplos novos que não foram expostos ao algoritmo no treinamento - Como a prova de um vestibular

Métricas de avaliação - Classificação

Matriz de confusão

- É um layout de tabela que permite a visualização do desempenho do algoritmo.

	Negativo Observado	Positivo Observado
Negativo Predito	**VN**	**FN**
Positivo Predito	**FP**	**VP**

Métricas de avaliação - Classificação

Medidas da matriz de confusão

```
    obs
pred     0    1
   0 18057 1828
   1   351  556
attr(,"class")
[1] "confusion.matrix"
```

Métricas de avaliação - Classificação

Acurácia

- Mede o quanto o modelo está classificando corretamente, tanto os casos negativos quanto os casos positivos.

$$Acuracia = \frac{(VP + VN)}{(VP + FP + VN + FN)}$$

Métricas de avaliação - Classificação

Recall ou sensibilidade

Dado que o estado verdadeiro é positivo, qual a proporção de verdadeiro positivo.

$$Recall = \frac{VP}{VP + FN}$$

$$Recall = \frac{556}{(556 + 1828)} = 0.2332$$

Métricas de avaliação - Classificação

Especificidade

Medida de quanto o modelo está classificando corretamente os verdadeiros negativos.

$$Especificidade = \frac{VN}{VN + FP}$$

$$Especificidade = \frac{18057}{(18057 + 351)} = 0.9809$$

Métricas de avaliação - Classificação

Precision ou valor preditivo positivo(VPP)

É o número de verdadeiros positivos dividido pelo número de positivos estimados pelo modelo.

$$Precision = \frac{VP}{VP + FP}$$

$$Precision = \frac{556}{(556 + 351)} = 0.613$$

Métricas de avaliação - Classificação

Valor preditivo negativo(VPN)

É o número de verdadeiros negativos dividido pelo número de negativos estimados pelo modelo.

$$VPN = \frac{VN}{VN + FN}$$

$$VPN = \frac{18057}{(18057 + 1828)} = 0.9081$$

Métricas de avaliação - Classificação

F1 score

Esta métrica combina a *Recall* (sensibilidade) e a *Precisão*(VPP) do modelo.

$$F1 = 2\left(\frac{precison.recall}{precision + recall}\right)$$

Métricas de avaliação - Classificação

Matthews correlation coefficient (mcc)

Medida de qualidade de duas classificações dicotômicas (ou binárias).
Usada para verificar o quanto um modelo está classificando corretamente os valores preditos com relação ao valores observados.

$$mcc = \frac{(VP.VN) - (FP.FN)}{\sqrt{(VP + FP)(VP + FN)(VN + FP)(VN + FN)}}$$

$$mcc = 0.3340531$$

Métricas de avaliação - Classificação

Teste Kolmogorov-Smirnov (K-S)

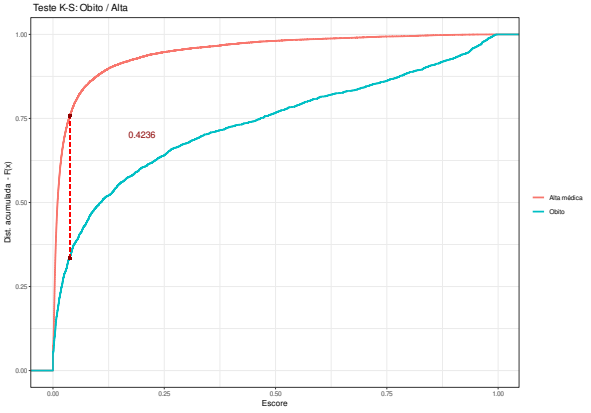
- Testa a aderência entre duas distribuições de densidade.

$$KS = \max \mid F(X_1)_m - F(X_2)_n \mid$$

- m é o número de exemplos da classe positiva.
- n é o número de exemplos da classe negativa.
- $F(X_1)_m$ é o vetor de valores da distribuição empírica da classe positiva.
- $F(X_2)_n$, é o vetor de valores da distribuição empírica da classe negativa.

Métricas de avaliação - Classificação

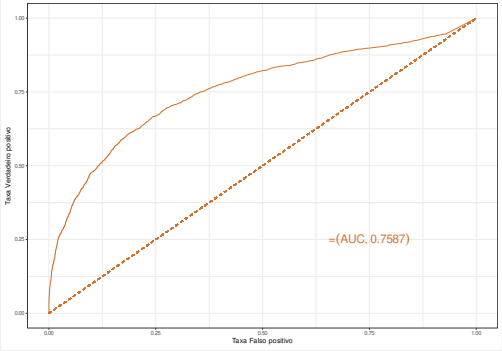
Teste Kolmogorov-Smirnov (K-S)



Métricas de avaliação - Classificação

Curva ROC(Receiver Operating Characteristic) e AUC(Area Under Curve)

- Varia-se o ponto de corte ao longo da amplitude dos escores fornecido pelo modelo. Para cada ponto de corte obtêm-se os respectivos valores para as medidas de sensibilidade e especificidade.



Métricas de avaliação - Regressão

Erro quadrático médio (EQM)

Essa métrica é mais útil quando erros grandes são particularmente indesejáveis (advindos de outliers, por exemplo), uma vez que os penaliza com mais intensidade.

$$EQM[y_i, h(\mathbf{x})] = \frac{1}{n}\sum_{i=1}^{n}[y_i - h(x_i)]^2$$

Métricas de avaliação - Regressão

Erro absoluto médio (EAM)

Mede a diferença média entre o valor observado e o estimado.
Mais robusto (menos influenciado) a estas questões de valores atípicos.

$$EAM[y_i, h(\mathbf{x})] = \frac{1}{n}\sum_{i=1}^{n}|y_i - h(x_i)|$$

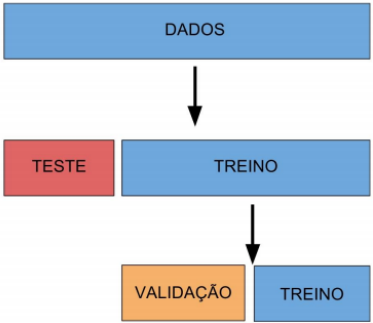
Métricas de avaliação - Regressão

Observações

1. $[EAM] \leq [REQM]$,o resultado da raiz quadrada do erro quadrático médio será sempre maior ou igual ao erro absoluto médio. Se todos os erros tiverem a mesma magnitude, então $[EAM] = [REQM]$.
2. $[REQM] \leq [EAM \times \sqrt{n}]$, em que n é o número de observações de validação.
3. Uma grande vantagem do EQM é evitar a função módulo, que é bastante indesejável em muitos cálculos matemáticos.
4. Coeficiente de correlação linear - valor observado e predito.
5. Gráfico de dispersão - valor observado vs predito.

Validação cruzada ***holdout***

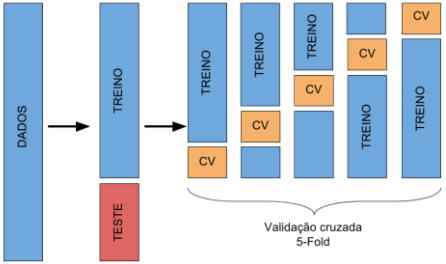
- Dividimos os dados em apenas duas partes:
  1. Treino
  2. Validação



Validação cruzada por *k-fold*

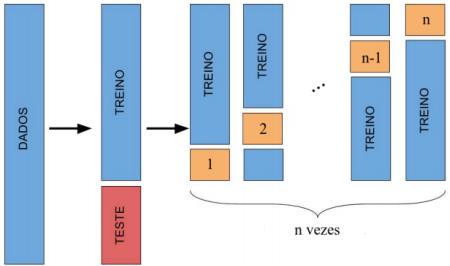
- Dividimos os dados em K partes iguais.

- Isto é feito para $k = 1 : K$, em seguida os resultados são combinados;



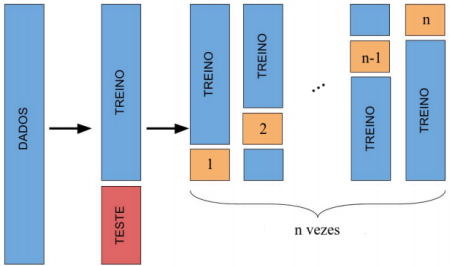
Validação cruzada *leave-one-out*

- Treino feito com $n - 1$ observações e valido com apenas uma observação e repete-se n vezes.



Validação cruzada *leave-one-sample*

- Treino feito com $n-1$ amostras e valido com apenas uma amostra e repete-se n vezes.



Validação cruzada: **certo e errado**

- Considere um problema de classificação em que a ***label(Y)*** possui duas classes:

1. Começando com 50 preditores(*inputs*) e amostra de tamanho 5000.
2. Filtramos os 10 preditores com maior correlação entre as classes.
3. Validamos o modelo utilizando somente os 10 preditores.

Como podemos estimar o desempenho do teste para este classificador?
Validação cruzada

Validação cruzada: **abordagem errada!**

1. Podemos aplicar validaçãao cruzada no Passo 3, esquecendo o Passo 1 (não incorporando o fato de termos eliminado 40 preditores)?

Não!

Validação cruzada: **abordagem errada!**

1. Podemos aplicar validaçãao cruzada no Passo 3, esquecendo o Passo 1 (não incorporando o fato de termos eliminado 40 preditores)?

Não!

Validação cruzada: **abordagem correta!**

1. Devemos aplicar validaçãao cruzada no Passo 1.
2. Não devemos ignorar o fato de que no Passo 1 o procedimento já viu os rótulos de treinamento, e aprendeu com isso.

Sim!

Validação cruzada: **abordagem correta!**

1. Devemos aplicar validaçãao cruzada no Passo 1.
2. Não devemos ignorar o fato de que no Passo 1 o procedimento já viu os rótulos de treinamento, e aprendeu com isso.

Sim!

Validação cruzada: **abordagem errada!**

1. Fazemos uma superamostragem na classe minoritária (***SMOTE***)
2. Aplicamos a validação cruzada nesses dados reamostrados
3. E validamos o modelo

Não!

Validação cruzada: **abordagem errada!**

1. Fazemos uma superamostragem na classe minoritária (***SMOTE***)
2. Aplicamos a validação cruzada nesses dados reamostrados
3. E validamos o modelo

Não!

Validação cruzada: **abordagem errada!**

1. Primeiro fazemos a valodação cruzada
2. Depois a reamostragem
3. E validamos o modelo

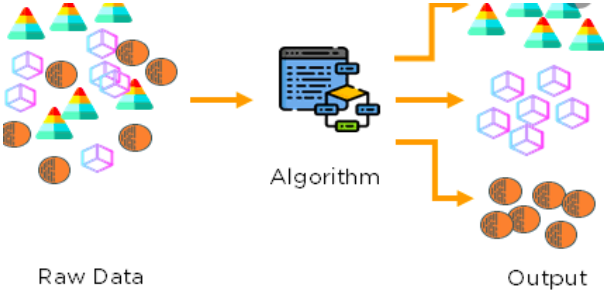
Sim!

Validação cruzada: **abordagem errada!**

1. Primeiro fazemos a valodação cruzada
2. Depois a reamostragem
3. E validamos o modelo

Sim!

Aprendizado não supervisionado



- Redução de dimensionalidade
- Agrupamento (*Clustering*)

Análise de Componentes Principais

- Suponha um vetor de características $X_1, X_2, ..., X_p$

$$Z_k = \phi_{1k}X_1 + \phi_{2k}X_2 + ... + \phi_{pk}X_p$$

- A variância total dos dados é definida como:

$$\sum_{j=1}^{p} Var(Z_j) = \lambda_1 + \lambda_2 + ... + \lambda_p$$

- Assim, a proporção da variância explicada pela j-ésima componente principal é dada por:

$$\frac{\lambda_j}{\lambda_1 + \lambda_2 + ... + \lambda_j}$$

Desafio 3

Com os dados coletados no desafio 1, faça:

- Uma análise de Componentes Principais

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.