Fundamentos da Aprendizagem de Máquina

Lista de Atividades

Aula - Modelos de Regressão Logística

Ricardo Augusto (ricardojunior@inatel.br)

Inatel

Índice

Regressão Logística

1. Atividades

Esse arquivo consiste em uma lista de atividades a serem realizadas para o módulo sobre modelos de regressão logística, do curso introdução à ciência de dados e decisões. A lista é composta pelas seguintes atividades:

- **Exercícios de Múltipla Escolha**
  - São dez (7) questões de múltipla escolha sobre os fundamentos de regressão linear discutidos em aula.
- **Exercícios Computacionais**
  - São três (3) exercícios computacionais relacionados com os modelos de regressão linear simples e múltipla

A composição da nota avaliativa desse módulo, denotada como N2 é dada pela combinação linear das atividades citadas, considerando pesos equilibrados, de acordo com

$$\text{N2} = 0.50 \times \text{Exercícios de Múltipla Escolha} + 0.50 \times \text{Exercícios Computacionais} \tag{1.1}$$

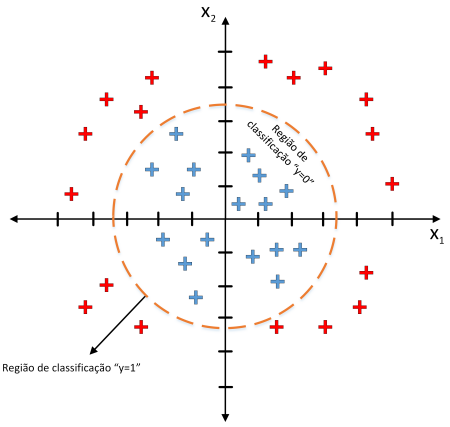
1.1 Exercícios de Múltipla Escolha

1. **Exercício 1 (Aprendizagem Supervisionada - Classificação)**

   Qual das alternativas abaixo retrata, de forma correta, um dos principais fundamentos da aprendizagem supervisionada que visa a tarefa de classificação:

   - a) O princípio fundamental da aprendizagem supervisionada que objetiva a classificação está nos valores discretos dos parâmetros dos modelos usados na construção dos classificadores.
   - b) Um classificador é o processo principal na fase de análise exploratória de dados.
   - c) O princípio fundamental da aprendizagem supervisionada que objetiva a classificação está na construção de um classificador que realiza predições contínuas sobre as variáveis explicativas.
   - d) O princípio fundamental da aprendizagem supervisionada que objetiva a classificação está nos valores discretos da variável de saída a ser predita.

2. **Exercício 2 (Fronteiras de Decisão)**
   Marque a alternativa correta a respeito do conceito das **fronteiras de decisão**, ilustrado na figura abaixo:



   - a) As fronteiras de decisão de um classificador dependem, exclusivamente dos dados de treinamento.
   - b) As fronteiras de decisão de um classificador baseado em regressão logística dependem da variância da variável de saída contínua y.
   - c) As fronteiras de decisão de um classificador baseado em regressão logística dependem dos parâmetros da função hipótese logística, que são obtidos a partir do processo de treinamento.
   - d) As fronteiras de decisão de um classificador baseado em regressão logística dependem dos parâmetros da função hipótese logística, que são obtidos a partir do processo de teste do classificador.

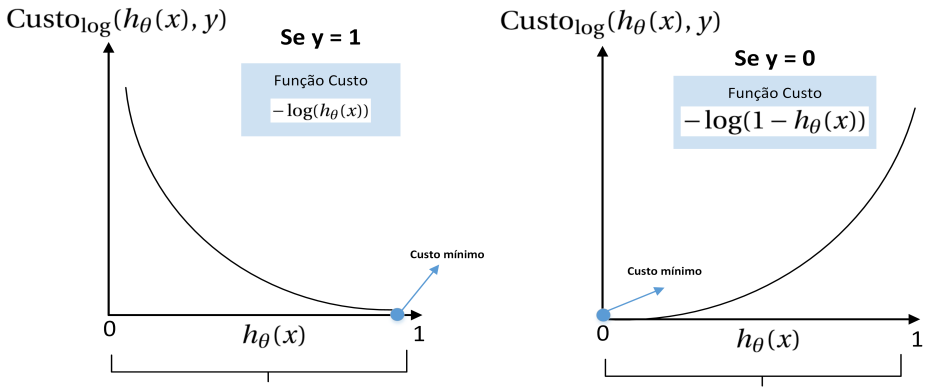
3. **Exercício 3 (Interpretação do custo logístico)**

   As figuras abaixo ilustram o conceito de logaritmo aplicado ao custo da regressão logística, descrito analiticamente a seguir:

$$\text{custo}_{\log}(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)), & \text{se } y = 1 \\ -\log(1 - h_\theta(x)), & \text{se } y = 0 \end{cases} \tag{1.2}$$

   Marque a alternativa correta sobre o comportamento do custo por dado de treinamento.

   - a) Se a classe de saída de um dado de treinamento $y = 1$ e o modelo de classificação forneceu a predição $h_\theta(x) \to 0$, então o custo do treinamento é mínimo, i.e., **custo** $\to 0$.
   - b) Se a classe de saída de um dado de treinamento $y = 0$ e o modelo de classificação forneceu a predição $h_\theta(x) \to 0$, então o custo do treinamento é máximo, i.e., **custo** $\to \infty$.
   - c) Se o resultado da função hipótese se aproxima de $h_\theta(x) \to 0$ e a classe de saída $y = 0$, então o modelo converge para a predição correta e o custo de treinamento logístico segue a expressão $-\log(1 - h_\theta(x))$.
   - d) Se o resultado da função hipótese se aproxima de $h_\theta(x) \to 1$ e a classe de saída $y = 1$, então o modelo não converge para a predição correta e o custo de treinamento logístico segue a expressão $-\log(h_\theta(x))$.



4. **Exercício 4 (Função hipótese logística)**

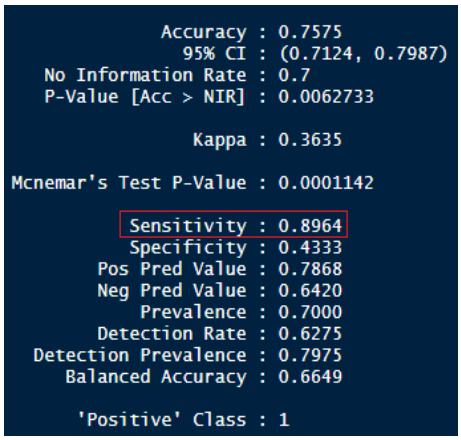
   Marque a alternativa correta a respeito da influência da função hipótese logística, mostrada abaixo, no processo de aprendizagem baseado na minimização da função custo.

$$\text{Função hipótese da regressão logística} \implies h_\theta(\mathbf{x}) = \frac{1}{1 + e^{-\boldsymbol{\theta}^{\mathrm{T}}\mathbf{x}}}, \tag{1.3}$$

   - a) O uso da função logística $h_\theta(\mathbf{x})$ no problema da minimização da função custo baseada no erro quadrático médio não tem influência sobre os algoritmos de aprendizagem como o gradiente descendente.
   - b) O uso da função sigmoid na regressão logística torna a função custo linear.
   - c) A função custo da regressão logística é do tipo sigmoid e, com isso, é convexa, possibilitando a convergência dos algoritmos de aprendizagem como o gradiente descendente.
   - d) O uso da função sigmoid torna a função hipótese logística não linear impactando o cálculo de $J(\theta)$ e fazendo com que esta tenha uma característica não convexa.

5. **Exercício 5 (Interpretação de Modelos de Regressão Logística)**

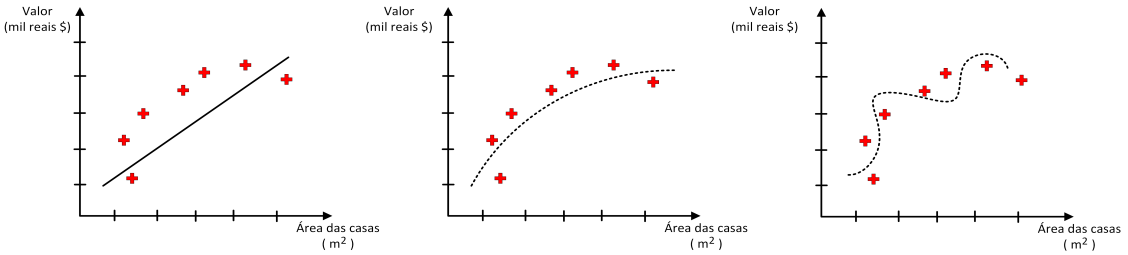
A Figura abaixo consiste na matriz de confusão associada aos resultados de um classificador construído na linguagem R. Baseado nos conceitos da matriz de confusão, marque a alternativa abaixo que retrata, corretamente, a interpretação do resultado de saída destacado em vermelho:



- a) O resultado de saída *sensitivity* é conhecido como *recall* e expressa pela razão entre os verdadeiros positivos e a soma entre os verdadeiros positivos e os falsos negativos gerados pelo classificador.
- b) O resultado de saída *sensitivity* é conhecido como *precision* e expressa pela razão entre os verdadeiros positivos e a soma entre os verdadeiros positivos e os falsos positivos gerados pelo classificador.
- c) O resultado de saída *sensitivity* é conhecido como *accuracy* e expressa a acurácia do classificador construído.
- d) O resultado de saída *sensitivity* é conhecido como *recall* e expressa pela razão entre os verdadeiros positivos e a soma entre os verdadeiros positivos e os verdadeiros negativos gerados pelo classificador.

6. **Exercício 6 (Overfitting)**

A figura abaixo ilustra os conceitos e características relacionadas com o problema de overfitting por meio de três gráficos.



$\theta_0 + \theta_1 x$ **(a)** $\theta_0 + \theta_1 x + \theta_2 x^2$ **(b)** $\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$ **(c)**

Marque a alternativa correta sobre as ideias transmitidas pelos gráficos em relação ao problema de overfitting.

- a) O problema de overfitting ocorre sempre que usamos os modelos de regressão linear no processo de treinamento, como no gráfico em (a).
- b) O gráfico em (c) transmite a ideia de ajuste aderente aos dados de treinamento e, portanto, a melhor generalização.
- c) O gráfico em (a) é um na subestimação ou subajuste do modelo, por não capturar a variabilidade dos dados em sua totalidade.
- d) O gráfio em (b) é um caso híbrido entre sobreajuste e, por isso, é apresenta a pior generalização.

7. **Exercício 7 (Regularização)**

Sobre a expressão da função custo mostrada abaixo, marque a alternativa correta quanto ao procedimento de regularização.

$$J(\theta) = \frac{1}{2m}\Big[\sum_{i=1}^{m}\Big(h_\theta(x^{(i)}) - y^{(i)}\Big)^2 + \lambda \sum_{j=1}^{n} \theta_j^2\Big], \tag{1.4}$$

- a) A expressão da função custo apresentada não apresenta a modificação relacionada com a regularização.
- b) O parâmetro λ controla a intensidade do fator de regularização e, com isso, se $\lambda \Rightarrow \infty$ temos a situação de sobreajuste e minimização completa da função custo.
- c) O efeito da regularização na função custo se dá pelo termo do somatório de parâmetros. Isso significa que a regularização maximiza os valores dos parâmetros a fim de manter as variáveis explanatórias presentes no modelo de ML.
- d) A ideia da regularização consiste em manter as variáveis explanatórias do modelo, mas reduzir os valores dos seus respectivos parâmetros, modificando $J(\theta)$.

1.2 Exercícios Computacionais

1. **Exercício 1 (Regressão Logística)**

   Considere o conjunto de dados **iris**, amplamente conhecido e usado como exemplo em diversos livros de aprendizagem de máquina. Esse conjunto já vem incorporado em diversos pacotes das linguagens R e Python, de acordo com

   Linguagem R: library(datasets)
   Linguagem Python: from sklearn import datasets

   A partir do conjunto de dados carregado capture somente os dados relacionados com a classe **virginica**, que consiste em uma **espécie** de uma planta (flor). O objetivo desse exercício é construir um classificador binário a partir de um modelo de regressão logística, permitindo verificar se uma espécie a ser testada é ou não do tipo **virginica**. Isso significa que um pré-processamento deve ser realizado sobre o dataset iris a fim de obtermos apenas duas classes (e.g., 0 -> não é virginica e 1 -> é virginica).
   As variáveis explanatórias desse conhecido dataset são:

   - **Petal.Length**: comprimento da pétala da flor
   - **Petal.Width**: largura da pétala da flor
   - **Sepal.Length**: comprimento da sépala da flor
   - **Sepal.Width**: largura da sépala da flor

   - **Questões Avaliativas**
     - 1) Realize o pré-processamento necessário para extração dos dados relacionados à classe **virginica**.
     - 2) Faça a análise de dados das variáveis explanatórias para o conjunto de dados.
     - 3) Realize a divisão do conjunto de treino e teste em 90/10.
     - 4) Para reprodução dos resultados use o set.seed(12).
     - 5) Forneça visualizações de dispersão e densidades das variáveis explanatórias de treino e as classes.
     - 6) Construa e treine o modelo preditivo de ML baseado em regressão logística.
     - 7) Faça a síntese do modelo e interprete os seus resultados.
     - 8) Encontre as variáveis explanatórias mais relevantes para o modelo.
     - 9) Faça as predições para os dados de teste e avalie os resultados
     - 10) A partir dos novos dados de entrada colocados abaixo, realize as classificações com o modelo

```r
# ------------------------------------------------------------------------------
# Realize predições para essas duas espécies

flor1 <- data.frame(Sepal.Length=6.4, Sepal.Width=2.8, Petal.Length=4.6, Petal.Width=1.8)
flor2 <- data.frame(Sepal.Length=6.3, Sepal.Width=2.5, Petal.Length=4.1, Petal.Width=1.7)
```

   - **Dicas para o Exercício**
     - Escolha o ambiente de desenvolvimento e a linguagem que for mais confortável para você (e.g., R/RStudio, Python/Jupyter, MATLAB, Java, entre outras), mas não deixe de visitar soluções diferentes, conversando com outros alunos, por exemplo).
     - Independente da linguagem, entenda o algoritmo e interprete-o como ferramenta, colocando o enfoque sobre a solução do problema.

2. **Exercício 2 (Regressão Logística)**

   Considere o desenvolvimento do modelo de classificação do Exercício Computacional 1 - obtenha a **matriz de confusão** do classificador. O ponto chave aqui é realizar a **interpretação** dos resultados obtidos.

3. **Exercício 3 (Regressão Logística)**

   Uma instituição financeira nos forneceu um conjunto de dados relacionados à créditos financeiros presentes no banco de dados da instituição. A instituição está trabalhando em um projeto de ciência de dados para previsão de risco de crédito. Nós iremos participar de uma fase específica desse projeto com o objetivo de construir um **classificador**, que possa auxiliar na análise de risco de crédito de diversos clientes da instituição.

   O modelo de ML (i.e., classificador) deve prever se um determinado cliente deve ou não receber créditos de produtos financeiros ofertados pela instituição. Isso significa que teremos acesso a um conjunto de dados com informações diversas sobre inúmeros clientes da instituição.

   O conjunto de dados consiste em vinte (20) variáveis explanatórias que consistem em informações diversas sobre os clientes incluindo: duração e tamanho do crédito, indicadores de saldo e operações financeiras, além de dados dos clientes como idade, dependentes, emprego e até contatos como telefone. Tais informações são apresentadas com codificação que são processadas pela instituição para posteriores interpretações. Com isso, nossa tarefa consiste em lidar com os dados codificados e a variável de saída **credit.rating**, que indica o estado de **aprovação** (1) ou **desaprovação** (0) de crédito para cada dados de treinamento (registro) do dataset.

   Os pacotes listados abaixo serão fundamentais para as questões avaliativas:
   - library(caret)
   - library(ROCR)
   - library(e1071)

   - **Questões Avaliativas**
     - 1) Realize a importação do arquivo .csv fornecido para o RStudio
     - 2) Faça a análise exploratória do dataset e verifique:
       - i) a necessidade de normalização dos dados
       - ii) conversão para fatores
     - 3) Considerando o item 2) identifique quais são as variáveis numéricas e os fatores.
       - Crie duas funções: para normalização e conversão em fatores.
     - 4) Realize a divisão do conjunto de treino e teste em 60/40.
     - 5) Para reprodução dos resultados use o set.seed(100).
     - 6) Construa e treine o modelo preditivo de ML baseado em regressão logística.
     - 7) Faça a síntese do modelo e interprete os seus resultados.
     - 8) Faça as predições para os dados de teste e avalie os resultados com a matriz de confusão.
     - 9) Use a função varImp do pacote **caret** para descobrir as variáveis explanatórias mais relevantes para o modelo.
     - 10) Obtenha a curva **ROC** do classificador antes da análise varImp.
     - 11) Obtenha a curva **ROC** do classificador após da análise com varImp.

   - **Dicas para o Exercício**
     - Escolha o ambiente de desenvolvimento e a linguagem que for mais confortável para você (e.g., R/RStudio, Python/Jupyter, MATLAB, Java, entre outras), mas não deixe de visitar soluções diferentes, conversando com outros alunos, por exemplo).
     - Independente da linguagem, entenda o algoritmo e interprete-o como ferramenta, colocando o enfoque sobre a solução do problema.