Introdução à Modelagem Estatística

Mini-curso de Introdução a ML e AI

Mário O. de Menezes

Maio/2020

Modelagem Estatística

Modelos

Em 1976, um estatístico britânico chamado George Box escreveu:
"Todos os modelos são errados, alguns são úteis."

O que é um modelo

- Um modelo é uma abstração da realidade;
- Necessariamente deixa de lados alguns aspectos *menos relevantes*;
- É uma simplificação proposital para um propósito específico.

Modelagem Estatística

- Uma das coisas mais sensíveis e importantes quando se começa é a escolha do tipo correto de análise estatística. A escolha depende:
  - da natureza dos dados
  - da questão que se quer responder, entre outras coisas.
- A chave é entender que tipo de variável *resposta* você tem e saber a natureza de suas variáveis *explicativas*.
  - A variável *resposta* é coisa com a qual você está trabalhando:
    - é a variável cuja variação você está tentando entender!
    - é a variável que você está tentando prever;
    - é a que vai no eixo y do gráfico.
  - A variável *explicativa* vai no eixo x do gráfico.
    - você está interessado em entender como a variação da variável *resposta* está associada com a variação da variável *explicativa*.

Modelagem Estatística

- Você também precisa considerar o *modo* que as variáveis na sua análise medem o que elas se propõem a medir.
- Uma medida contínua é uma variável do tipo altura ou peso que pode assumir valores com números reais.
- Uma variável categórica é um fator com dois ou mais níveis:
  - sexo é um fator com dois níveis (masculino e feminino)
  - cor pode ser um fator com sete níveis (vermelho, laranja, amarelo, verde, azul, índigo e violeta)
- Portanto, é essencial responder às seguintes questões:
  - Qual das variáveis é **a variável resposta**?
  - Quais são as variáveis explicativas?
  - As variáveis explicativas são contínuas ou categóricas, ou uma mistura de ambas?
  - Que tipo de variável resposta temos:
    - é uma medida contínua? uma contagem? uma proporção? um tempo (ocasião) de morte? ou uma categoria?

Variáveis

Car_Name	Year	Selling_Price	Present_Price	Kms_Driven	Fuel_Type
ritz	2014	3.35	5.59	27000	Petrol
sx4	2013	4.75	9.54	43000	Diesel
ciaz	2017	7.25	9.85	6900	Petrol

age	sex	bmi	children	smoker	region	charges
19	female	27.90	0	yes	southwest	16884.924
18	male	33.77	1	no	southeast	1725.552
28	male	33.00	3	no	southeast	4449.462

LOW	LWT	RACE
0	182	2
0	155	3
0	105	1

Método Estatístico Apropriado

Algumas *chaves* simples para a escolha do método estatístico apropriado

As variáveis explicativas

1. Todas as variáveis explicativas são contínuas ⇒ **Regressão**
   - É possível realizar uma Regressão com variáveis explicativas contínuas e categóricas (transformando)
2. Todas as variáveis explicativas são categóricas ⇒ **Análise de Variância (ANOVA)**
3. Variáveis explicativas são tanto contínuas como categóricas ⇒ **Análise de Covariância (ANCOVA)**

A variável resposta

1. Contínua ⇒ **Regressão Normal, ANOVA ou ANCOVA**
2. Proporção ⇒ **Regressão Logística**
3. Contagem ⇒ **Modelos log-linear**
4. Binária ⇒ **Análise logística binária**
5. Tempo na morte ⇒ **Análise de sobrevivência**

Objetivo da Modelagem Estatística

- Determinar os valores dos parâmetros em um modelo específico que *levam ao melhor ajuste do modelo aos dados*
- Os dados são *sacrosantos*; eles nos dizem o que realmente aconteceu sob determinadas circunstâncias.
  - É um erro comum dizer "os dados foram ajustados ao modelo" como se os dados fossem flexíveis, e nós tivéssemos uma estrutura clara do modelo.
  - É o contrário: o que se procura é o modelo **mínimo adequado** que descreva os dados.
  - O modelo é ajustado aos dados; não o contrário!
- O melhor modelo é o que produz o mínimo de variação não explicada (o *mínimo desvio dos resíduos*), sujeito à restrição de que todos os parâmetros no modelo devem ser estatisticamente significantes

Primeiros passos

- A melhor coisa a fazer é gastar um tempo substancial, logo de início, para entender os dados e o que eles mostram.
  - Isto vai ajudar a guiar o pensamento para a modelagem estatística mais apropriada.
- **Thinking with Data** – Max Shron
  - Scoping: Why Before How
    - *"Most people start working with data from exactly the wrong end. They begin with a data set, then apply their favorite tools and techniques to it. The result is narrow questions and shallow arguments. Starting with data, without first doing a lot of thinking, without having any structure, is a short road to simple questions and unsurprising results. We don't want unsurprising – we want knowledge."*

Checklist da Modelagem Estatística

Checklist

- Certificar-se de que o `data.frame` está correto em estrutura e conteúdo:
  - Todos os valores de cada variável estão na mesma coluna?
    - **tidy data**
  - Todos os zeros são realmente 0 ou deveriam ser `NA`?
  - Cada linha contém o mesmo número de entradas?
  - Existe algum nome de variável que contém espaço?
- Depois de carregar os dados, a Análise Exploratória de Dados é **essencial**

Checklist da Modelagem Estatística

Sobre o Modelo

- Algumas coisas básicas na escolha do modelo
  - Quais variáveis explicativas deveriam ser incluídas?
  - Transformação da variável resposta é necessária?
  - Interações deveriam ser incluídas?
  - Termos não lineares deveriam ser incluídos? $(X^2, X^3...)$
  - As variáveis explicativas deveriam ser transformadas?
- Tente utilizar o tipo mais simples de análise que seja apropriado para seus dados e para a questão que está tentando responder.
- Ajuste um modelo máximo e vá simplificando-o paulatinamente ao remover parâmetros.
- Faça o *diagnóstico do modelo*
- Por fim, documente tudo o que fizer, e explique cada um dos passos. Desta maneira você entenderá o que fez e porque fez quando retornar à sua análise 6 meses mais tarde!

Especificando o modelo

- Um modelo incorpora nosso entendimento mecanicista das variáveis explicativas envolvidas, e da maneira que elas estão relacionadas com a variável resposta.
- Buscamos um modelo **mínimo** por conta do princípio da *parcimônia*, e também um modelo **adequado**
- É muito importante entender que *não há **um** modelo*.
  - em muitos casos, haverá um grande número de modelos diferentes, uns mais plausíveis do que outros.
- É preciso determinar quais, se algum, dos modelos possíveis, são adequados
  - e depois, dos adequados, qual é o modelo *mínimo adequado*.
  - pode haver um conjunto de modelos que descrevem os dados igualmente bem (ou de modo igualmente probre se a variabilidade for grande)

Objetivo do Modelo – Minimizar os resíduos

- O que, exatamente, queremos dizer quando afirmamos que os valores dos parâmetros devem dar conta do *melhor ajuste do modelo aos dados* ?
- A convenção utilizada é que nossas técnicas devem levar a **estimadores que minimizem a variância e sejam livres de viés**.
- Nós definimos **melhor** em termos da "máxima verosimilhança".
- Uma definição *funcional* para estes termos é:
  - Dados os dados,
  - e dada nossa escolha do modelo,
  - quais valores dos parâmetros deste modelo farão os dados observados mais prováveis?
- Julgamos o modelo com base em *quão prováveis os dados seriam se o modelo estivesse correto*!
  - Ou seja, o modelo que produz os menores resíduos (diferença entre os valores reais e os preditos).

Próximo Bloco

Regressão Linear Simples