Aprendizagem Estatística em Altas Dimensões

[MAE0501/MAE5904/IBI5904]

Departamento de Estatística (MAE) - IME/USP

Profª Associada Florencia Graciela Leonardi

Entrega #1: Proposta de Trabalho

Ícaro Maia Santos de Castro (Nº USP 11866921)

✉ icaromsc@usp.br

Rayssa de Carvalho Roberto (Nº USP 10940828)

✉ rayscarvalho@usp.br

Rodrigo Aoyama Nakahara (Nº USP 3510922)

✉ nakahara@usp.br

Rodrigo Marcel Araujo Oliveira (Nº USP 9299208)

✉ rodrigo.marcel.oliveira@usp.br

Vitor Hugo Vieira de Lima (Nº USP 10263886)

✉ vitorhugo@usp.br

30 de Setembro de 2020

Membros do Grupo

O Grupo 11 é formado por:

- Ícaro Maia Santos de Castro (11866921) — Doutorado Direto - Bioinformática - IME-USP
- Rayssa de Carvalho Roberto (10940828) — Mestrado - Biologia Genética - IB-USP
- Rodrigo Aoyama Nakahara (3510922) — Bacharelado - Estatística - IME-USP
- Rodrigo Marcel Araujo Oliveira (9299208) — Bacharelado - Estatística - IME-USP
- Vitor Hugo Vieira de Lima (10263886) — Bacharelado - Estatística - IME-USP

Descrição do Problema

O objetivo geral do problema é prever se a pessoa possui ou não diabetes *mellitus*, com base em uma série de variáveis preditoras. O desafio foi posto declaradamente como um problema de aprendizagem.

O banco de dados possui todas suas observações constituídas de pessoas do gênero feminino, com idade superior a 21 anos, de ascendência do povo *Pima* (grupo de nativos norte-americanos). Os dados, que são atualmente abertos e gratuitos para *download* (`CC0 1.0`), são provenientes do Instituto Nacional de Diabetes e Distúrbios Digestivos e do Rim (NIDDK) dos EUA e mantidos pela *UC Irvine Machine Learning Repository* (University of California-Irvine).

Descrição do Conjunto de Dados

O banco de dados é composto pelas seguintes variáveis:

Tabela 1: Descrição das Variáveis

Variável	Descrição
Outcome	Variável resposta categórica (1 se diabético, 0 se não diabético)
Pregnancies	Quantidade de gestações
Glucose	Concentração de glicose no plasma após 2 horas em um teste oral de tolerância a glicose
BloodPressure	Pressão arterial diastólica (mm Hg)
SkinThickness	Espessura da dobra da pele do tríceps (mm)
Insulin	Insulina sérica de 2-horas (mu U/ml)
BMI	Índice de massa corporal (peso em kg/(altura em m)^2)
DiabetesPedigreeFunction	Função "pedigree" de diabetes
Age	Idade (anos)

Para um conhecimento preliminar das variáveis, são mostradas algumas medidas-resumo.

Tabela 2: Estatísticas Descritivas: Medidas-Resumo Gerais

	N.Valid	**Mean**	**Std.Dev**	**CV**	**Skewness**	**Kurtosis**
Age	768	33.240	11.760	0.353	1.125	0.621
BloodPressure	768	69.105	19.355	0.280	-1.836	5.117
BMI	768	31.992	7.884	0.246	-0.427	3.244
DiabetesPedigreeFunction	768	0.471	0.331	0.702	1.912	5.528
Glucose	768	120.894	31.972	0.264	0.173	0.619
Insulin	768	79.799	115.244	1.444	2.263	7.133
Outcome	768	0.3489	0.476	1.366	0.632	-1.601
Pregnancies	768	3.845	3.369	0.876	0.898	0.142
SkinThickness	768	20.536	15.952	0.776	0.108	-0.530

Tabela 3: Estatísticas Descritivas: 5 Números e Medidas Robustas

	Min	**Q1**	**Median**	**Q3**	**Max**	**MAD**	**IQR**
Age	21	24	29	41	81	10.378	17
BloodPressure	0	62	72	80	122	11.861	18
BMI	0	27.3	32	36.6	67.1	6.82	9.3
DiabetesPedigreeFunction	0.078	0.244	0.373	0.627	2.42	0.249	0.383
Glucose	0	99	117	140.5	199	29.652	41.25
Insulin	0	0	30.5	127.5	846	45.219	127.25
Outcome	0	0	0	1	1	0	1
Pregnancies	0	1	3	6	17	2.965	5
SkinThickness	0	0	23	32	99	17.791	32

O estudo pioneiro de SMITH *et alii* (1988)[1] foi um dos primeiros a selecionar as variáveis desse banco de dados. Por esse motivo, esse estudo descreve em maiores detalhes a constituição de cada uma dessas variáveis.

Em linhas gerais, a variável resposta `Outcome` é categórica (dicotômica, 0 ou 1, indicando a presença ou ausência da diabetes) e apresentou 268 como diabéticos no total da amostra de tamanho 768.

Além disso, a função "pedigree" de diabetes refere-se a uma função de linhagem que avalia a probabilidade de diabetes com base no histórico familiar. Sua metodologia de cálculo é devidamente detalhada no mencionado *paper*.

[1] Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C., & Johannes, R.S. (1988). *Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus*. In: Proceedings of the Symposium on Computer Applications and Medical Care (pp. 261–265). IEEE Computer Society Press.

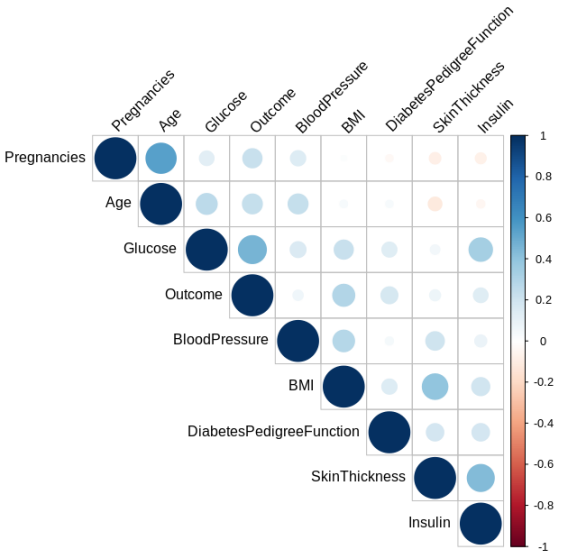

Figura 1: Correlação entre as variáveis.

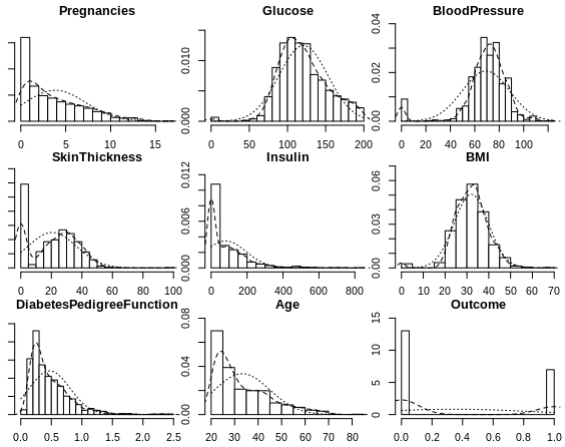

Figura 2: Histograma da distribuição das variáveis.

Em relação a correlação entre as variáveis, pode-se verificar conforme a Figura 1 que a variável `glucose` possui uma correlação moderada com a variável resposta `Outcome`. Essas associações entre as variáveis serão detalhadamente analisadas e discutidas no decorrer do trabalho. Adicionalmente, na Figura 2, podemos verificar as variáveis que parecem seguir uma distribuição normal, o que pode ser um ponto importante a ser considerado na definição dos métodos de aprendizagem estatística a serem aplicados.

Avaliação do Problema

Em essência, o conjunto de dados será analisado como um problema de inferência (dado que se tem uma variável resposta disponível), mas pode eventualmente ser tratado com um problema de predição, caso se queira utilizar o modelo para prever a condição de novos indivíduos, conforme entendimento de JAMES, HASTIE & TIBSHIRANI (2013).

Considerando a natureza categórica da variável resposta, o problema será, em princípio, um problema de classificação (em contraste com os problemas de regressão). No entanto, como reconhecem JAMES, HASTIE & TIBSHIRANI (2013), essa distinção pode não ser muito clara quando se trabalha com variáveis resposta dicotômicas e se deseja estimar, por exemplo, probabilidades esperadas em um modelo de regressão logístico, ao invés de se desejar uma classificação/discriminação. Nesse sentido, como a utilização de modelos de regressão com respostas categóricas é uma das possibilidades a serem consideradas para a análise desses dados, o problema poderia ser tanto de classificação quanto de regressão.

No que se refere ao tipo de aprendizagem, o conjunto das possíveis técnicas a serem utilizadas direciona o problema para que seja do tipo supervisionado. Dentre as muitas possibilidades, pode-se citar a própria regressão logística e a máquina de suporte vetorial (*support vector machine*). Mas, como se tem disponível uma variável resposta, será também possível utilizar técnicas tipicamente não-supervisionadas (tais como a análise de agrupamentos ou *clusters*) com apoio de uma validação cruzada.

Endereço da Página do Problema

O conjunto dos dados, bem como outras informações, podem ser obtidos em:

https://www.kaggle.com/uciml/pima-indians-diabetes-database