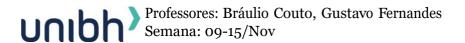
Análise de Dados e Big Data Prática 13: Regressão Logística e Linear Aplicadas



Objetivo: Analisar dados de amostra de vinhos brancos e tintos, construindo e interpretando dois modelos, uma para predizer a chance de um vinho vir a ser classificado como "bom" e outro modelo para estimar o seu teor de álcool com base nas suas características de composição físico-química.

Os dados usados nessa prática podem ser obtidos <u>aqui</u>. No *data-set* é possível observar 13 variáveis:

- 1. Tipo de vinho (branco ou tinto)
- 2. Acidez fixa
- 3. Acidez volatil
- 4. Acido citrico
- 5. Acucar residual
- 6. Cloretos
- 7. Dioxido de enxofre livre
- 8. Dioxido de enxofre total
- 9. Densidade
- 10. pH
- 11. Sulfatos
- 12. Concetracao final de alcool
- 13. Vinho de boa qualidade (sim ou não)

Parte 1: "Entendendo" os dados

- 1) Quais das 13 variáveis são quantitativas e quais delas são categóricas?
 - a. Categóricas:
 - b. Quantitativas:
- 2) Considere a variável "Vinho de boa qualidade" como desfecho ou variável resposta ou variável dependente ou variável de interesse. Neste cenário, quais são as variáveis explicativas ou independentes?
 - a. Resposta:
- 3) Considere agora "Concetração final de alcool" como desfecho, quais são as variáveis explicativas?
 - a. Resposta:

Parte 2: Modelo de Regressão Linear Múltipla

Resumo estatístico dos dados summary(Dados_Vinho)

2) Construa diagramas de dispersão e avalie a possível relação entre as variáveis explicativas em relação ao desfecho *Concetração final de alcool*. Use o commando (R):

```
ggplot(data=data_set, mapping = aes(x = X, y = Y)) + geom_point() +
geom_smooth(method = lm, se = FALSE)

Opção:
scatter.smooth(X ~ Y)
```

3) Faça a regressão linear múltipla, considerando o desfecho *Concetracao final de álcool*. Use os comandos (R):

```
modelo.linear <- lm(Y ~ X1 + X2 + X3 + ... + Xn, data = data_set)
summary(modelo.linear)

#############
contrasts(Tipo.de.vinho)
modelo.alcool = lm(Concetracao.final.de.alcool~ Tipo.de.vinho +
Acidez.fixa + Acidez.volatil + Acido.citrico + Acucar.residual +
Cloretos + Dioxido.de.enxofre.livre + Dioxido.de.enxofre.total +
Densidade + pH + Sulfatos)

summary(modelo.alcool)</pre>
```

- a. Responda às perguntas:
 - O tipo de vinho (tinto versus branco) afeta a sua concentração alcóolica final?
 - Quais variáveis têm correlação positiva significativa com a concentração alcóolica final do vinho de tal forma que, quando esta variável aumenta, o teor alcóolico do vinho também aumenta?
 - Quais variáveis têm correlação negativa significativa com a concentração alcóolica final do vinho de tal forma que, quando esta variável aumenta, o teor alcóolico do vinho também aumenta?

Parte 3: Modelo de Regressão Logística

a) Construa um modelo de regressão para a qualidade do vinho. **Interprete os resultados!**

(gerar modelo de regressão logística)

```
modelo.logistico <- glm(Y ~ X1 + X2 + X3 + ... + Xn, data = data_set, family = binomial)
contrasts(Tipo.de.vinho)

modelo.logistico <- glm(Vinho.de.boa.qualidade~ Tipo.de.vinho + Acidez.fixa +
Acidez.volatil + Acido.citrico + Acucar.residual + Cloretos +
Dioxido.de.enxofre.livre + Dioxido.de.enxofre.total + Densidade + pH +
Sulfatos + Concetracao.final.de.alcool, family = binomial)

summary(modelo.logistico)

# Gerar as porbabilidades do vinho ser "bom"
modelo.probs <- predict(modelo.logistico,type = "response")</pre>
```

Gerar a "matriz de confusão" e taxa de classificação

```
modelo.pred <- ifelse(modelo.probs > seu_limiar, "limite superior", "limite inferior")
modelo.pred <- ifelse(modelo.probs > 0.5, "Sim", "Nao")

table(modelo.pred, data_set$Y)
table(modelo.pred, Vinho.de.boa.qualidade)

mean(modelo.pred == data_set$Y)
mean(modelo.pred == Vinho.de.boa.qualidade)
```

Regras e Organização:

- a. O valor da prática é de 2 pontos;
- b. O grupo deve ser composto de até <u>5 alunos</u>;
- c. A prática deverá ser submetida até o fim da aula corrente na tarefa indicada no uLife.
- d. O professor irá validar os resultados antes da submissão do arquivo.
- e. Alguém do grupo deverá criar um arquivo.pdf contendo as respostas. Não esquecer os nomes dos componentes do grupo.
- f. No fim, o professor irá discutir o desempenho esperado com os grupos.
 - Metas de compreensão: Analisar, produzir e interpretar informações.