Qualidade do vinho português com base nas suas propriedade físico-químicas

Trabalho de Álgebra Linear

Gabriel Matos dos Santos
Edgard Júnior

Uma breve apresentação da parte teórica e experimental.

Escola de Matemática Aplicada
Fundação Getúlio Vargas
Rio de Janeiro - Brasil
Outubro de 2023

Sumário

Resumo

A busca por modelos que se ajustem aos dados é fundamental em diversas áreas do conhecimento e pesquisa. Modelos eficazes permitem prever resultados, avaliar o impacto de variáveis específicas e obter dados estatísticos valiosos. Nesse contexto, a análise da qualidade do vinho português, com base em suas propriedades físico-químicas, torna-se uma tarefa relevante. Esta análise é conduzida com o intuito de identificar quais dessas propriedades possuem maior ou menor influência na qualidade do vinho.

1 Parte teórica

Para avaliar a importância das variáveis que afetam a qualidade do vinho, utilizamos o conceito de regressão linear múltipla. Essa abordagem assume que a qualidade do vinho pode ser expressa como uma combinação linear de suas propriedades. Por meio dessa suposição, estimamos os coeficientes de cada atributo do vinho, o que nos permite entender seu impacto no produto final. Antes de explorar os detalhes de nossa análise, é importante compreender os fundamentos da regressão linear múltipla.

1.1 Medição de uma constante

Vamos começar com um exemplo simples que demonstra como podemos estimar uma constante a partir de várias medições.

Ao tentar medir um valor constante, geralmente obtemos resultados ligeiramente diferentes devido a imprecisões no processo de medição. Suponhamos ser k a constante a ser determinada e $b = [x_1 \ \cdots \ x_n]^T \in \mathbb{R}^n$ o vetor onde a i-éssima coordenada representa o valor obtido na i-éssima medição.

Caso não houvesse imprecisões, obteríamos o vetor $a = [k \ \cdots \ k]^T \in \mathbb{R}^n$ composto apenas da constante k. Entretanto, devido às imprecisões, procuramos um vetor da forma $c = [\alpha \ \cdots \ \alpha]^T \in \mathbb{R}^n$ que é "mais próximo"de b. Note que $\alpha \neq k$.

Para determinar "o mais próximo"utilizamos a norma euclidiana. Desta forma, desejamos minimizar a distância $\|c - b\|$ entre os vetores b e c. Analisando o caso $n = 2$, a minimização buscada ocorre quando $\langle c - b, \ a \rangle = 0$, i.e. $c - b$ é ortogonal a a.

Estendendo a ideia para o caso geral, teríamos que $\|c - b\|$ é mínimo quando

$$\langle c - b, \ a \rangle = \langle [\alpha \ \cdots \ \alpha] - [x_1 \ \cdots \ x_n] \, , \ [k \ \cdots \ k] \rangle = \sum_{k=1}^{n} (\alpha - x_k) \, k = 0$$

e isso ocorre justamente quando $\alpha = (x_1 + \cdots + x_n) \, / n$. Portanto, a melhor estimativa que conseguimos para k é a média aritmética dos resultados.

No próximo exemplo exemplo buscaremos estender o conceito de ortogonalidade nesse contexto.

1.2 Mínimos quadrados e regressão linear múltipla

Vamos começar com um cenário geral que fornecerá uma compreensão sólida do método.

Suponha que uma variável b dependa linearmente das variáveis x_1, ..., x_n. Nosso objetivo é estimar os parâmetros que relacionam b com as demais incógnitas. Se β_i é o parâmetro associado a x_i, teremos as m medições, representadas por:

$$\begin{cases} \beta_1 a_{1,1} + \cdots + \beta_n a_{1,n} = b_1 \\ \beta_1 a_{2,1} + \cdots + \beta_n a_{2,n} = b_2 \\ \quad\vdots \qquad\qquad\quad \vdots \\ \beta_1 a_{m,1} + \cdots + \beta_n a_{m,n} = b_m \end{cases}$$

Aqui, $a_{i,j}$ representa o valor da variável x_j na i-ésima medição. Essas equações podem ser representadas matricialmente como:

$$\begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m,1} & a_{m,2} & \cdots & a_{m,n} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}$$

Chamemos essas matrizes de A, y e b, respectivamente Nosso objetivo é um vetor $y \in \mathbb{R}^n$ de forma que a norma $\|Ay - b\|$ seja a menor possível. Uma forma equivalente e mais confortável de determinar y é minimizar a o quadrado de $\|Ay - b\|$, daí o nome de *método de mínimos quadrados*.

Seja p a projeção ortogonal de b no espaço coluna de A[1]. Usando propriedades de ortogonalidade, podemos decompor $Ay - b$ unicamente em $Ay - p \in C(A)$ e $b - p \in C(A)^{\perp}$. Assim, temos:

$$\|Ay - b\|^2 = \|Ay - p\|^2 + \|b - p\|^2$$

Note que

$$\|Ay - b\|^2 \geq \|b - p\|^2$$

é minimizado quando

$$\|Ay - p\|^2 = 0, \text{ o que implica } Ay - p = 0$$

Assumindo A invertível, podemos determinar os parâmetros desejados resolvendo

$$A^T A y = A^T b$$

Efetuando as multiplicações matriciais, basta resolver o seguinte sistema linear

$$\begin{bmatrix} \sum_{k=1}^{n} a_{1,k}^2 & \cdots & \sum_{k=1}^{n} a_{1,k} a_{m,k} \\ \vdots & \ddots & \vdots \\ \sum_{k=1}^{n} a_{m,k} a_{1,k} & \cdots & \sum_{k=1}^{n} a_{m,k}^2 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_n \end{bmatrix} = \begin{bmatrix} \sum_{k=1}^{m} b_k a_{1,k} \\ \vdots \\ \sum_{k=1}^{m} b_k a_{n,k} \end{bmatrix}$$

Este procedimento é conhecido como *regressão linear múltipla*. Quando temos apenas uma variável x_i (ou seja, $b = \beta x$), chamamos de apenas de *regressão linear*.

[1]isto é, $p \in C(A)$ é tal que $p - b \in C(A)^{\perp}$.

A montagem e resolução desse sistema é implementado na biblioteca `statsmodels` e com ele que conseguimos realizar a modelagem de base de dados.