Universidade Federal de Minas Gerais

Exercício 2

A.P. Braga

Agosto de 2020

1 Problema Não-Linearmente Separável

Considerando-se o conjunto de dados representado na Figura 1, pede-se que seja implementada, em R ou Python, uma projeção não linear arbitrária que torne o problema linearmente separável.

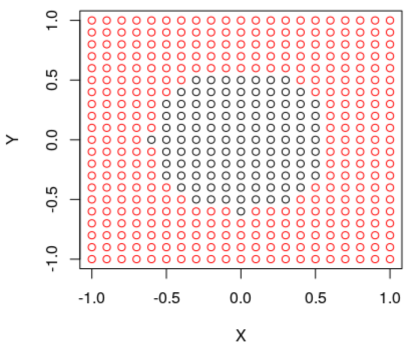

Figure 1: Pontos em vermelho pertencem à classe positiva e pontos em preto pertencem à classe negativa

Os dados foram gerados em R a partir da seguinte rotina:

```
x = seq(-1,1,by = 0.1)
y = seq(-1,1,by = 0.1)
create_grid <- expand.grid(x,y)

circle <- function(x,y) {
  return(sqrt(x^2+y^2))
}

raio = 0.6

classe = 1*(circle(create_grid$Var1,create_grid$Var2)>raio)
```

2 *Overfitting* e *Underfitting*

Considerando-se a Figura 2, que apresenta os dados de treinamento para um problema de regressão:

- Qual dos 3 modelos construídos (preto, vermelho e azul) parece ser o mais adequado aos dados, sabendo-se que existe um ruído na amostragem?
- Qual dos modelos apresenta menor erro de treinamento?
- Quais os problemas com cada um dos modelos inadequados?

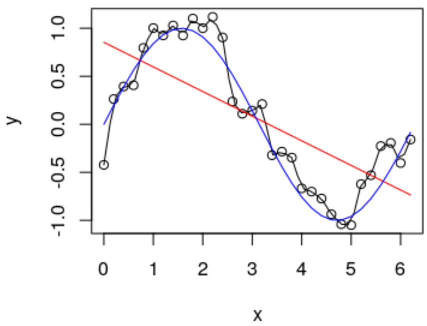

Figure 2: Ajuste de três modelos para um problema de regressão

3 Aproximação Polinomial

Introdução Teórica

Considere um polinômio de grau p conforme representado na sua forma geral na Equação 1.

$$p(x) = w_p x^p + w_{p-1} x^{p-1} + \cdots + w_1 x + w_0 \tag{1}$$

em que x é o argumento e w_i é o coeficiente do termo de grau i.

Dadas as observações (x_i, y_i) representadas na forma do conjunto de dados $D = \{x_i, y_i\}_{i=1}^N$, deseja-se encontrar o polinômio de grau p que melhor aproxima a função geradora $f_g(x)$ do conjunto D. O objetivo é, a partir das amostras de dados, encontrar o grau p e os coeficientes w_i de forma tal que $p(x) \approx f_g(x)\ \forall x$. A aproximação de $f_g(x)$ é usualmente feita com base na minimização do erro dos termos quadráticos $(y_i - p(x_i))^2$ $(i = 1 \cdots N)$. Espera-se que o conjunto D contenha informação suficiente para que seja possível aproximar $f_g(x)$ por $p(x)$ com base somente nas suas N amostras. Os parâmetros de $p(x)$ são ajustados de forma tal que $y_i = w_p x_i^p + w_{p-1} x_i^{p-1} + \cdots + w_1 x_i + w_0\ \forall x_i \in D$, conforme representado no sistema de equações 2.

$$\begin{array}{lllllll}
y_1 = w_p x_1^p & +w_{p-1} x_1^{p-1} & + & \cdots & +w_1 x_1 & +w_0 \\
y_2 = w_p x_2^p & +w_{p-1} x_2^{p-1} & + & \cdots & +w_1 x_2 & +w_0 \\
\quad \vdots & \quad \vdots & & \cdots & \vdots & \\
y_N = w_p x_N^p & +w_{p-1} x_N^{p-1} & + & \cdots & +w_1 x_N & +w_0
\end{array} \tag{2}$$

O sistema representado em 2 possui N equações e p incógnitas, podendo também ser representado na forma matricial 3.

$$\mathbf{H}\mathbf{w} = \mathbf{y} \tag{3}$$

em que $\mathbf{H}$, $\mathbf{w}$ e $\mathbf{y}$ são representados em 4, 5 e 6.

$$\mathbf{H} = \begin{bmatrix} x_1^p & x_1^{p-1} & \cdots & x_1 & 1 \\ x_2^p & x_2^{p-1} & \cdots & x_2 & 1 \\ \vdots & \vdots & \cdots & \vdots & \\ x_N^p & x_N^{p-1} & \cdots & x_N & 1 \end{bmatrix} \tag{4}$$

$$\mathbf{w} = \begin{bmatrix} w_p \\ w_{p-1} \\ \vdots \\ w_1 \\ w_0 \end{bmatrix} \tag{5}$$

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} \tag{6}$$

A matriz $\mathbf{H}$ possui um papel importante na resolução do problema de aproximação, pois ela contém os termos não-lineares que compõem o polinômio $p(x)$, os quais serão responsáveis pela projeção dos elementos x_i no espaço composto pelo sistema de coordenadas caracterizado pelas colunas de $\mathbf{H}$. Como $\mathbf{H}$ e $\mathbf{y}$ são dados pelo problema, a solução da Equação 3 pode ser obtida por meio da pseudoinversa, conforme Equação 7.

$$\mathbf{w} = \mathbf{H}^{+}\mathbf{y} \tag{7}$$

em que $\mathbf{H}^{+}$ é a pseudoinversa de $\mathbf{H}$.

Exercícios

Considerando a aproximação polinomial das seções anteriores, faça:

- Obtenha aproximações polinomiais a partir de 10 amostras da função geradora $f_g(x) = \frac{1}{2}x^2 + 3x + 10$ somadas com um ruído gaussiano $N(mean = 0, sd = 4)$ amostradas entre $x = -15$ e $x = 10$, com um número de amostras $N = 20$ e grau do polinônimo variando entre $p = 1$ a $p = 8$. Para cada aproximação, mostre um gráfico com a função geradora, as amostras e o polinômio obtido.
- Responda: Ocorreu Overfitting? Ocorreu Underfitting? Em quais casos ocorreu estes fenômenos?
- Repita o procedimento para 100 amostras ao invés de 10. Qual o impacto do número de amostras na aproximação polinomial?

Para o cálculo da pseudo-inversa o aluno deverá usar o pacote *library("corpcor")*.

Instruções: O exercício deve ser entregue em formato *pdf* via moodle. Recomenda-se a utilização de ferramentas como R markdown, Sweave ou Jupyter Notebook (ou similares), para elaboração dos relatórios.