# Aprendizado de Máquina

# Aula 7: Algoritmos baseados em probabilidade (parte 4)

André C. P. L. F de Carvalho
ICMC/USP
andre@icmc.usp.br

# Tópicos

- Introdução
- Modelos preditivos
- Tarefas de regressão
- Aproximação de funções
- Regressão linear simples
- Regressão polinomial
- Algoritmo de ajuste de parâmetros
- Regressão linear múltipla
- Regularização

# Regressão múltipla

- De acordo com o número de variáveis independentes consideradas, uma tarefa de regressão pode ser
    - Simples: uma única variável independente (atributo preditivo)
        - $f(x) = \mathrm{w}_0 x_0 + \mathrm{w}_1 x_1$
    - Múltipla: mais de uma variável independente (atributo preditivo)
        - Linear
            - $f(x) = \mathrm{w}_0 x_0 + \mathrm{w}_1 x_1 + \mathrm{w}_2 x_2 + \ldots + \mathrm{w}_d x_d$
        - Polinomial
            - $f(x) = \mathrm{w}_0 x_0 + \mathrm{w}_1 x_1^{3} - \mathrm{w}_2 x_1^{2} + \mathrm{w}_3 x_2^{4} + \mathrm{w}_4 x_1^{7}$

# Dados estruturados para tarefa de regressão

Atributos de entrada (preditivos)

| Estudou | Assistiu | Jogou | Leu | Nota |
|---|---|---|---|---|
| 20 | 6 | 5 | 2 | 7.0 |
| 10 | 2 | 4 | 1 | 3.0 |
| 30 | 8 | 0 | 3 | 10.0 |
| 10 | 7 | 1 | 2 | 8.0 |
| 6 | 6 | 3 | 1 | 6.0 |
| 4 | 0 | 0 | 0 | 1.0 |

Atributo alvo

# Dados estruturados para tarefa de regressão

Atributos de entrada (preditivos)

| Estudou | Assistiu | Jogou | Leu | Nota |
|---|---|---|---|---|
| 20 | 6 | 5 | 2 | 7.0 |
| 10 | 2 | 4 | 1 | 3.0 |
| 30 | 8 | 0 | 3 | 10.0 |
| 10 | 7 | 1 | 2 | 8.0 |
| 6 | 6 | 3 | 1 | 6.0 |
| 4 | 0 | 0 | 0 | 1.0 |

Atributo alvo

Obs.:
$x_2 = 8$
$x_0$ é uma constante

# Regressão múltipla

- Busca função capaz de associar
  - Valores de mais de um atributo preditivo ao
  - Valor de um atributo alvo

$$f(x) = w_0 + w_1 x_1 + w_2 x_2 + \ldots + w_d x_d$$

onde:

$w_j$ é o parâmetro $j$
$x_j$ é o valor da variável $x_j$
$d$ é o número de parâmetros

# Regressão múltipla

- Busca função capaz de associar
  - Valores de mais de um atributo preditivo a
  - Valor de um atributo alvo

$$f(x) = w_0 x_0 + w_1 x_1 + w_2 x_2 + \ldots + w_d x_d$$

onde:

$w_j$ é o parâmetro $j$
$x_j$ é o valor da variável $x_j$
$x_0$ é o valor constante 1
$d + 1$ é o número de parâmetros

  - Ex.: $f(x) = 3x_0 - 2{x_1}^2 + {x_1}^3 + 5{x_2}^5 + 2x_3$

# Regressão múltipla

- Aplicação de um modelo de regressão linear pode ser visto como uma multiplicação de matrizes

$$f(x) = w_0 x_0 + w_1 x_1 + w_2 x_2 + \ldots + w_d x_d$$
$$= w^T . x$$

- Exemplo para d = 4: w = $\begin{pmatrix} 1 \\ 1 \\ -1 \\ 100 \\ 1 \end{pmatrix}$ e x = $\begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}$

# Regressão múltipla

- Aplicação de um modelo de regressão linear pode ser visto como uma multiplicação de matrizes

$$f(x) = w_0 x_0 + w_1 x_1 + w_2 x_2 + \ldots + w_d x_d$$
$$= w^T . x$$

- Exemplo para d = 4: $w = \begin{pmatrix} 1 \\ 1 \\ -1 \\ 100 \\ 1 \end{pmatrix}$ e $x = \begin{pmatrix} 1 \\ 0 \\ 1 \\ 4 \\ -2 \end{pmatrix}$

$$w^T = [1, 1, -1, 100, 1]$$

# Regressão múltipla

- Encontrar valor para cada parâmetro $\mathrm{w}_i$ que reduza o erro cometido pela função
  - Erro: diferença entre valor predito e valor verdadeiro para todos os exemplos
    - Idealmente erro = 0
  - Minimizar função de custo $\mathrm{J}(\mathrm{w}_0, \mathrm{w}_1, \mathrm{w}_2 \dots \mathrm{w}_d)$

$$\mathrm{J}(\mathrm{w}_0,\ \mathrm{w}_1,\ \mathrm{w}_2 \dots \mathrm{w}_d) = \frac{1}{2n}\sum_{i=1}^{n}(y^i - f(x^i))^2$$

$$\mathrm{J}(\mathrm{w}) = \frac{1}{2n}\sum_{i=1}^{n}(yi - f(x^i))^2$$

*onde:*

$n$ é o número de objetos no conjunto de dados
d é o número de atributos preditivos
$x^i$ é o vetor de entrada para o i−ésimo exemplo
$y^i$ é o valor de saída para o i−ésimo exemplo

# Sumarizando

- Função hipótese: $f(x) = w_0 x_0 + w_1 x_1 + \ldots + w_d x_d$
- Parâmetros: $w_0, w_1, \ldots, w_d$
- Função de custo: $J(w_0, w_1, \ldots w_d) = J(w) = \frac{1}{2n}\sum_{i=1}^{n}(y^i - f(x^i))^2$
- Objetivo: $\min_{w} J(w)$
  - Ajuste simultâneo
    - $aux_0 = w_0 - \alpha \frac{\partial}{\partial w_0} J(w) = w_0 - \alpha \frac{1}{n}\sum_{i=1}^{n}\left(y^i - f(x^i)\right) x_0{}^i$
    - $aux_1 = w_1 - \alpha \frac{\partial}{\partial w_1} J(w) = w_1 - \alpha \frac{1}{n}\sum_{i=1}^{n}\left(y^i - f(x^i)\right) x_1{}^i$
    - ...
    - $aux_d = w_d - \alpha \frac{\partial}{\partial w_d} J(w) = w_d - \alpha \frac{1}{n}\sum_{i=1}^{n}\left(y^i - f(x^i)\right) x_d{}^i$

Importante diferenciar entre:
*$f(x) = x^4 - 4x^2 + 2x + 4$ (PS)*
*$f(x) = x_1^4 - 4x_2^2 + 2x_3 + 4$ (PM)*

$w_0 = aux_0$
$w_1 = aux_1$
...
$w_d = aux_d$

# Regressão com regularização

- Algoritmos de regressão podem induzir modelos sobre-ajustados aos dados (overfitting)
  - Processo de indução do modelo leva em conta detalhes (ruídos) da distribuição dos dados (variância)
  - Possíveis causas de overfitting:
    - Grande número de atributos preditivos ou baixo número de exemplos
    - Baixa proporção do número de exemplos em relação ao número de atributos preditivos
    - A distribuição dos dados é muito complexa
  - Overfitting pode ser reduzido restringindo a complexidade dos modelos induzidos
    - Incorporando um termo de regularização para restringir o valor dos parâmetros
      - Aumenta presença de viés (bias) na função de custo

# Regularização

- Funções conseguem se ajustar melhor aos dados se puderem ter valores positivos e negativos muito grandes
- Adiciona um termo de penalidade à função de custo
  - Reduz a complexidade dos modelos gerados
    - Encolhe (*shrink*) os valores dos coeficientes
    - Penaliza função de custo quando coeficientes têm valores elevados
      - Quanto mais complexo o modelo, maior a penalidade

$$\mathrm{J(w)} = \frac{1}{2n}\sum_{i=1}^{n}(y^i - f(x^i))^2 + \lambda Penalidade$$

- Principais técnicas de regularização:
  - Regressão Lasso
  - Regressão de Ridge
  - Redes elásticas

# Lasso

- Least Absolute Shrinkage and Selection Operator
- Regularização L1 (função de perda, loss function, L1)
    - Penalidade: soma dos valores absolutos dos coeficientes
        - Método de menor desvio absoluto
        - Encolhe modelo gerado

$$J(w) = \frac{1}{2n}\sum_{i=1}^{n}(y^i - f(x^i))^2 + \lambda\|w\|_1$$

$$L1 = \lambda\|w\|_1 = \lambda\sum_{j=1}^{d}|w_j|$$

# Ridge

- Regularização L2 (função de perda, loss function, L2)
  - Penalidade mais usada
  - Quanto maior o valor do parâmetro (positivo ou negativo) , mais ele é penalizado
    - Adiciona à função de custo um termo de penalidade (soma dos quadrados dos coeficientes)

$$J(w) = \frac{1}{2n}\sum_{i=1}^{n}(yi - f(x^i))^2 + \lambda\|w\|_2^2$$

$$L2 = \lambda\|w\|_2^2 = \lambda\sum_{j=1}^{d} w_j^2$$

# Rede elástica

- Elastic net
- Combina a regressão Lasso com a regressão de Ridge
    - Utilizando as penalidades L1 e L2
        - L1 para gerar dados esparsos
        - L2 para remover limite de número de variáveis selecionadas

$$J(w) = \frac{1}{2n}\sum_{i=1}^{n}(yi - f(x^i))^2 + \lambda_1 \sum_{j=1}^{d} |w_j| + \lambda_2 \sum_{j=1}^{d} w_j^2$$

        - Alternativa:

$$J(w) = \frac{1}{2n}\sum_{i=1}^{n}(yi - f(x^i))^2 + \lambda \sum_{j=1}^{d} |w_j| + (1 - \lambda) \sum_{j=1}^{d} w_j^2$$

# Conclusão

- Tarefas de regressão
- Regressão linear
  - Simples
  - Múltipla
- Regressão polinomial
  - Simples
  - Múltipla
- Algoritmo de ajuste de parâmetros
- Regularização

# Fim do apresentação