

Caderno_Inferencia_Estatistica

2024-03-12

Sumario

Instrumentação Matemática para Estatística – Prof Wagner Hugo Bonat – 09/03/2024

- Matemática (5 partes)
- Probabilidade (1 parte)
- Inferência (3 partes)

Tópicos em matemática customizados para DS:

- Fornecer base matemática para entender e criar técnicas de análise de dados
- Visão geral e intuitiva
- Focar nos resultados e suas aplicações
- Não ser exaustivo em cada tópico ou matematicamente (muito) rigoroso
- Suporte computacional para compreender conceitos matemáticos abstratos
- Formar uma base sólida para entender técnicas avançadas:
- Modelagem estatística
- Machine learnig

O curso não é de receitas, é de fundamentos Os objetivos desta abordagem são:

- Desmistificar o processo peos quais os algoritmos resolvem problemas
- Mostrar que apesar de existir um conjunto enrme de técnicas, muitas delas são pequenas melhorias em técnicas já existentes
- Promover um uso qualificado das ferramentas já disponíveis

Referencias:

- Deep Learning – Ian Goodfellow et al, 2016
- Mathematics for Machine Learning – Marc Peter Deisenroth, 2019

O que precisamos saber?

- Cálculo Diferencial e Integral
- Funções, limites e continuidade
- Derivadas
- Integrais
- Álgebra Matricial
- Vetores e escalares

- Matrizes
- Sistemas de equações lineares
- Decomposições matriciais
- Métodos Numéricos
- Sistemas de equações não-lineares
- Diferenciação e integração numérica
- Otimização

Material:

- Livro: Matemática para Ciências de Dados

Exemplos motivacionais:

Classificador binário Ferramenta popular em modelagem estatística e aprendizagem de máquina

Objetivo: classificar um indivíduo ou observação em uma entre duas categorias

Exemplos:

- Classificar um paciente como saudável ou doente
- Classificar um cliente como bom ou mal pagador etc

Diversos algoritmos disponíveis:

- Árvores de classificação
- Máquinas de vetores de suporte
- Redes neurais
- Gradient boost
- Regressão logística é muito popular

Descrição matemática:

- Suponha que temos um conjunto de dados y_i para $i = 1, \dots, n$.
- Cada $y_i \in [0, 1]$ (é zero ou 1) -> sim ou não, saudável ou doente etc

Potenciais objetivos:

- Descrever o relacionamento de y_i com um conjunto de variáveis explanatórias x_{ij} com $j = 1, \dots, p$
- Classificar uma nova observação como 0 ou 1

Exemplo - Conjunto de dados com 3 colunas:

- Renda anual do usuário
- Anos de experiência do usuário
- Se é premium ou não

Objetivos:

- Identificar como as covariáveis renda e anos influenciem a compra premium
- Prever se um novo usuário será ou não premium
- Orientar campanha de marketing

[Gráfico em R]

$y = f(x_{i1} = renda, x_{i2} = anos)$ 1 0 18 10

2 1 18 10

3 1 20 3

4 1 20 8

$y_i = f(x_{ij})$ $y_i = f(x_{ij}) + erro$ $erro = y_i - f(x_{ij})$

Construção do classificador:

- Explicar o modelo que descreve a relação entre y_i e x_{ij} (i linha-usuário, j coluna-covariável)

$y = f(renda, xp)$, ou seja, y é função dependente de renda e xp

- Especificar função perda (medida de erro)

$erro = g(y_i, f(x_{ij}))$ função g

- Características satisfaça duas equações de distância: $d(y, \mu) > 0 | y = \mu$ e $d(y, \mu) = 0 | \mu = f(x_{ij})$

Otimizar a função perda:

- Qual algoritmo escolher?
- Como implementá-la?
- Analisar o modelo ajustando

Kmeans

Clusterização usando kmeans

- Agrupar indivíduos semelhantes
- Indivíduos no mesmo grupo sejam mais parecidos do que indivíduos em grupos diferentes
- Distância da media

Math part

Linha reta

$$y_i = \beta_0 + \beta_1 * renda$$

Sigmoid

$$y_i = \frac{1}{1 + \exp^{-(\beta_0 + \beta_1 * renda + \beta_2 * anos)}}$$

Combinando o modelo logístico com a função perda

$$SQ_{logit}(\beta) = \sum_{i=1}^n \left(y_i - \frac{1}{1 + \exp^{-(\beta_0 + \beta_1 * renda + \beta_2 * anos)}} \right)^2$$

```
# f_logit <- funcao(par, y, renda, anos) {  
#   mu <- 1 / (1 + exp(-(par[1] + par[2] * renda + par[3] * anos)))  
#   SQ_logit <- sum((y - mu) ^ 2)  
#   return(SQ_logit)  
# }
```