Ciência e Dados

Data Science For Professionals

8 Conceitos Estatísticos Fundamentais Para Data Science

Posted on 16 de março de 2021

Estatística é “um ramo da matemática que lida com a coleta, análise, interpretação e apresentação de massas de dados numéricos”. Junte Programação, Ciência da Computação e Machine Learning e você terá uma boa descrição das principais habilidades em Data Science.

A Estatística é usada em quase todos os aspectos da Ciência de Dados. É usada para analisar, transformar e limpar dados,

avaliar e otimizar algoritmos de aprendizado de máquina e também é usada na apresentação de percepções e descobertas.

O campo da Estatística é extremamente amplo e pode ser difícil determinar exatamente o que você precisa aprender e em que ordem. Mas o fato é que nem tudo é necessário em Ciência de Dados e não é necessário graduação em Estatística para trabalhar como Cientista de Dados.

No artigo a seguir, veremos 8 Conceitos Estatísticos Fundamentais Para Data Science que você precisa entender ao estudar ou trabalhar com Ciência de Dados. Estas não são técnicas particularmente avançadas, mas são uma seleção dos requisitos básicos que você precisa saber antes de passar para o aprendizado de métodos mais complexos.

1- Amostragem

Em Estatística, todos os dados brutos que você pode ter disponíveis para um teste ou experimento é conhecido como população. Por uma série de razões, não é viável medir os padrões e tendências em toda a população. As estatísticas nos permitem tomar uma amostra, realizar alguns cálculos sobre o conjunto de dados e, usando a probabilidade e algumas suposições, podemos com um certo grau de certeza compreender as tendências para toda a população ou prever eventos futuros.

Digamos, por exemplo, que queremos entender a prevalência de uma doença como o câncer de mama em toda a população do Brasil. Por razões práticas, não é possível rastrear toda a população. Em vez disso, podemos pegar uma amostra aleatória e medir a prevalência entre esses dados. Supondo que nossa amostra seja suficientemente aleatória e representativa de toda a população, podemos obter uma medida de prevalência e fazer inferências sobre toda a população.

2- Estatística Descritiva

A estatística descritiva, como o nome sugere, nos ajuda a descrever os dados. Em outras palavras, permite-nos compreender as características. Aqui o objetivo não é prever algo, fazer suposições ou inferência, mas simplesmente fornecer uma descrição da aparência da amostra de dados que temos.

As estatísticas descritivas são normalmente calculadas a partir dos dados. Isso inclui as medidas de tendência central, tal como:

Média – o valor médio dos dados.
Mediana – o valor central se ordenarmos os dados em orem crescente e dividirmos exatamente pela metade.
Moda – o valor que ocorre com mais frequência.

3- Distribuições

As estatísticas descritivas são úteis, mas muitas vezes podem ocultar informações importantes sobre o conjunto de dados. Por exemplo, se um conjunto de dados contém vários números que são muito maiores do que os outros, a média pode ser distorcida e não nos dará uma representação verdadeira dos dados.

Uma distribuição pode ser representada por um gráfico, geralmente um histograma, que exibe a frequência com que cada valor aparece em um conjunto de dados. Este tipo de gráfico nos fornece informações sobre a dispersão e a assimetria dos dados.

Uma distribuição geralmente formará um gráfico semelhante a uma curva, que pode ser inclinada mais para a esquerda ou direita.

Uma das distribuições mais importantes é a distribuição normal, comumente chamada de curva em sino devido ao seu formato. É de forma simétrica com a maioria dos valores agrupados em torno do pico central e os valores mais distantes distribuídos igualmente em cada lado da curva. Muitas variáveis na natureza formarão uma distribuição normal, como a altura das pessoas e as pontuações de QI. A distribuição normal de uma variável é a suposição de vários algoritmos de Machine Learning.

4- Probabilidade

Probabilidade, em termos simples, é a probabilidade de um evento ocorrer. Em Estatística, um evento é o resultado de um experimento que pode ser algo como o lançamento de um dado ou os resultados de um teste AB.

A probabilidade de um único evento é calculada dividindo o número de eventos pelo número total de resultados possíveis. Considere, por exemplo, conseguir um seis ao lançar um dado. Como existem 6 resultados possíveis, a chance de rolar um seis é 1/6 = 0,167, e às vezes isso também é expresso como uma porcentagem, então 16,7%.

Os eventos podem ser independentes ou dependentes.

Com eventos dependentes, um evento anterior influencia o evento subsequente. Digamos que temos um pacote de M&M e queremos determinar a probabilidade de escolher aleatoriamente um M&M vermelho. Todas as vezes que removêssemos um M&M do pacote, a probabilidade de escolher o vermelho mudaria devido ao efeito de eventos anteriores.

Os eventos independentes não são afetados por eventos anteriores. No caso do pacote de M&M se cada vez que selecionamos um o colocamos de volta no pacote, a probabilidade de selecionar vermelho permaneceria a mesma todas as vezes.

Se um evento é independente ou não, é importante, pois a maneira como calculamos a probabilidade de vários eventos muda dependendo do tipo.

A probabilidade de vários eventos independentes é calculada simplesmente multiplicando a probabilidade de cada evento. No exemplo do lançamento de dados, digamos que quiséssemos calcular a chance de lançar um 6 três vezes. Isso seria parecido com o seguinte:

1/6 = 0,167

1/6 = 0,167

1/6 = 0,167

0,167 * 0,167 * 0,167 = 0,005

O cálculo é diferente para eventos dependentes, também conhecido como probabilidade condicional. Se tomarmos o exemplo do M&M, imagine que temos um pacote com apenas duas cores vermelho e amarelo, e sabemos que o pacote contém 3 vermelhos e 2 amarelos e queremos calcular a probabilidade de escolher dois vermelhos em uma fileira. Na primeira escolha, a probabilidade de escolher um vermelho é 3/5 = 0,6. Na segunda escolha, removemos um M&M, que por acaso era vermelho, então nosso segundo cálculo de probabilidade é 2/4 = 0,5. A probabilidade de escolher dois vermelhos em uma fileira é, portanto, 0,6 * 0,5 = 0,3.

5- Viés

Como discutimos anteriormente, usamos amostras de dados para fazer estimativas sobre todo o conjunto de dados. Da mesma forma, para modelagem preditiva, usaremos alguns dados de treinamento e tentaremos construir um modelo que possa fazer previsões sobre novos dados.

Viés é a tendência de um modelo estatístico ou preditivo de super ou subestimar um parâmetro. Isso geralmente se deve ao método usado para obter uma amostra ou à forma como os erros são medidos. Existem vários tipos de vieses comumente encontrados nas estatísticas. Aqui está uma breve descrição de dois deles.

Viés de seleção – ocorre quando a amostra é selecionada de forma não aleatória. Em Data Science, um exemplo pode ser interromper um teste AB mais cedo quando o teste está em execução ou selecionar dados para treinar um modelo de aprendizado de máquina de um período de tempo que pode mascarar os efeitos sazonais.

Viés de confirmação – ocorre quando a pessoa que realiza alguma análise tem uma suposição predeterminada sobre os dados. Nessa situação, pode haver uma tendência de gastar mais tempo examinando variáveis que provavelmente apoiarão essa suposição.

6- Variância

Como discutimos anteriormente neste artigo, a média é uma medida de tendência central. A variância mede a distância de cada valor no conjunto de dados da média. Essencialmente, é uma medida da dispersão dos números em um conjunto de dados.

O desvio padrão é uma medida comum de variação para dados que têm uma distribuição normal. É um cálculo que fornece um valor para representar a extensão da distribuição dos valores. Um desvio padrão baixo indica que os valores tendem a ficar muito próximos da média, enquanto um desvio padrão alto indica que os valores estão mais dispersos.

Se os dados não seguem uma distribuição normal, outras medidas de variância são usadas. Normalmente, o intervalo interquartil é usado. Essa medida é derivada primeiro ordenando os valores por classificação e, em seguida, dividindo os pontos de dados em quatro partes iguais, chamadas quartis. Cada quartil descreve onde 25% dos pontos de dados se encontram de acordo com a mediana. O intervalo interquartil é calculado subtraindo a mediana dos dois quartos centrais, também conhecidos como Q1 e Q3.

7- Tradeoff Viés/Variância

Os conceitos de viés e variância são muito importantes em Machine Learning. Quando construímos um modelo de aprendizado de máquina, usamos uma amostra de dados conhecida como conjunto de dados de treinamento. O modelo aprende padrões nesses dados e gera uma função matemática que é capaz de mapear o rótulo de destino correto ou valor (y) para um conjunto de entradas (X).

Ao gerar esta função de mapeamento, o modelo usará um conjunto de suposições para melhor aproximar o alvo. Por exemplo, o algoritmo de regressão linear assume uma relação linear (linha reta) entre a entrada e o destino. Essas suposições geram viés no modelo.

Como cálculo, o viés é a diferença entre a previsão média gerada pelo modelo e o valor verdadeiro.

Se tivéssemos de treinar um modelo usando diferentes amostras de dados de treinamento, obteríamos uma variância nas previsões que são retornadas. A variância no aprendizado de máquina é uma medida de quão grande é essa diferença.

Em aprendizado de máquina, o viés e a variância constituem o erro geral esperado para nossas previsões. Em um mundo ideal, teríamos baixo viés e baixa variância. No entanto, na prática, minimizar o viés geralmente resultará em um aumento na variância e vice-versa. A compensação de viés / variância descreve o processo de equilibrar esses dois erros para minimizar o erro geral de um modelo.

8- Correlação

Correlação é uma técnica estatística que mede as relações entre duas variáveis. A correlação é considerada linear (formando uma linha quando exibida em um gráfico) e é expressa como um número entre +1 e -1, conhecido como coeficiente de correlação.

Um coeficiente de correlação de +1 indica uma correlação perfeitamente positiva (quando o valor de uma variável aumenta o valor da segunda variável também aumenta), um coeficiente de 0 indica nenhuma correlação e um coeficiente de -1 indica uma correlação negativa perfeita.

Importante ainda ressaltar que correlação não implica causalidade. O fato de haver correlação entre duas variáveis não significa que uma causa a ocorrência da outra. Para afirmar isso teríamos que realizar estudos adicionais e uma análise de causalidade.

A Estatística é um campo amplo e complexo. Este artigo pretende ser uma breve introdução a algumas das técnicas estatísticas mais comumente usadas em Data Science.

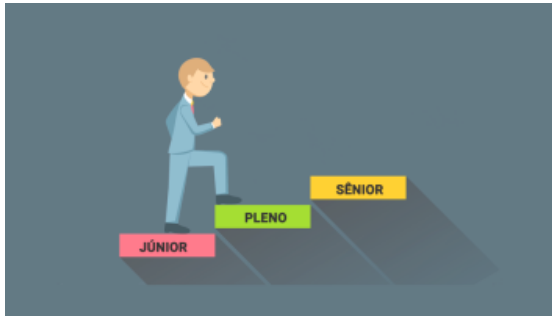
David Matos

Referências:

8 Fundamental Statistical Concepts for Data Science

Análise Estatística Para Data Science

O Papel da Estatística na Ciência de Dados
2 de novembro de 2015
Em "Estatística"



Cientista de Dados – Júnior, Pleno e Sênior
14 de novembro de 2021
Em "Carreira"

As 10 Habilidades de um Cientista de Dados
25 de setembro de 2017
Em "Cientista de Dados"

2 thoughts on “8 Conceitos Estatísticos Fundamentais Para Data Science”

Jorge Olimpio de Souza disse:
30 de março de 2021 às 4:53 PM

É seu trabalho de Tese?
Atividade construtiva, fonte de conhecimento.

Responder

David Matos disse:
31 de março de 2021 às 12:15 AM

Olá Jorge. Somente colaboração com outros mesmo.

Responder

Deixe um comentário

O seu endereço de e-mail não será publicado. Campos obrigatórios são marcados com *

COMENTÁRIO

NOME *

E-MAIL *

☐ NOTIFIQUE-ME SOBRE NOVOS COMENTÁRIOS POR E-MAIL.

☐ NOTIFIQUE-ME SOBRE NOVAS PUBLICAÇÕES POR E-MAIL.

Publicar comentário

ASSINAR BLOG POR E-MAIL

Digite seu endereço de e-mail para assinar este blog e receber notificações de novas publicações por e-mail.

Endereço de e-mail

Assinar

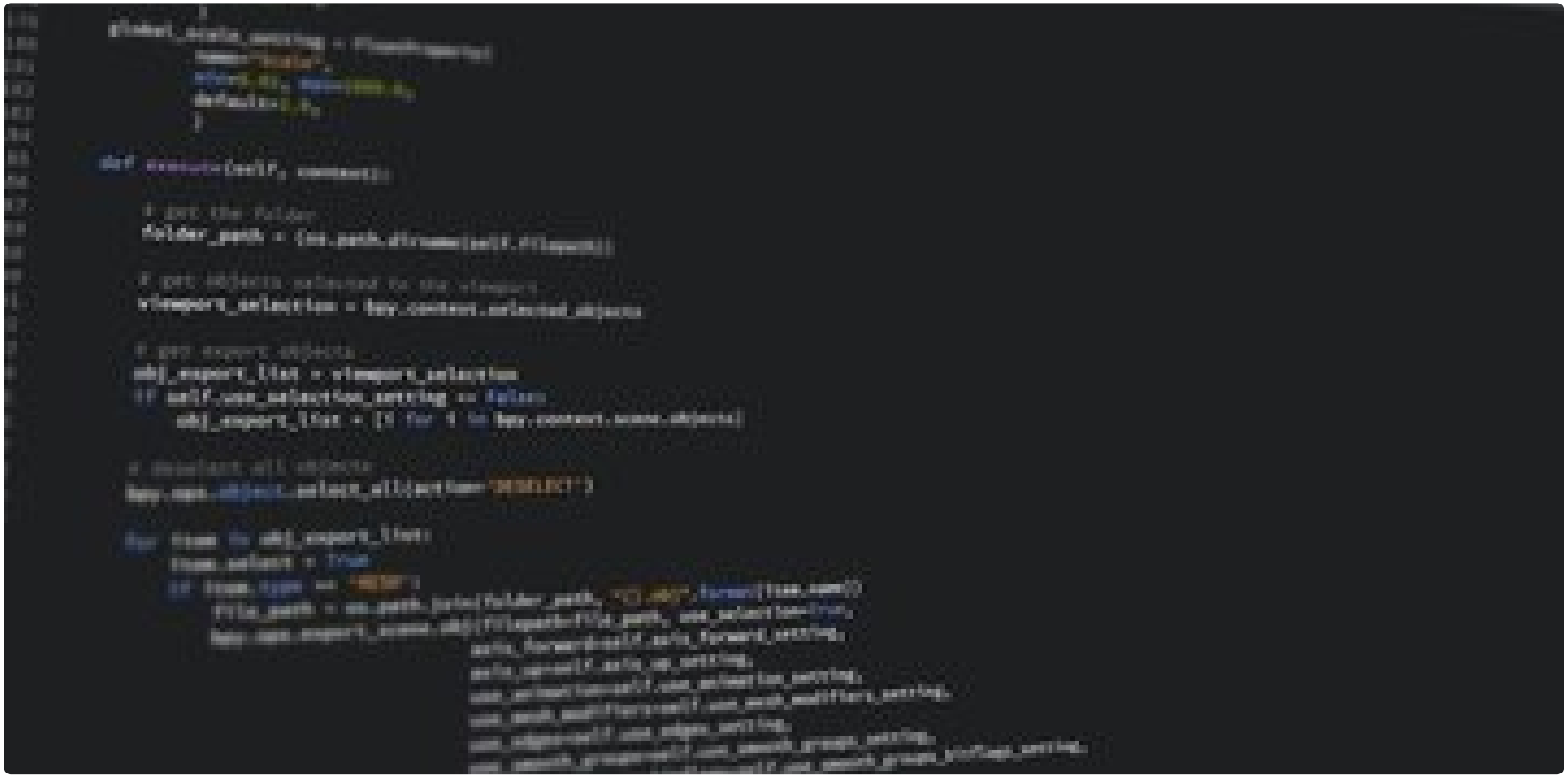
BUSCAR

Pesquisar ...

TWITTER

Ciência e Dados
@cienciaedados

10 linguagens de programação que o mercado vai exigir em 2022canaltech-com-br.cdn.ampproject.org/c/s/canaltech.…



17 de dez. de 2021

Ciência e Dados
@cienciaedados

Na Era da Informação, o aprendizado é data-drivenmittechreview.com.br/na-era-da-info…

Na Era da Informação, o aprendizado é data-driven - MIT Technology Review

A horizontalidade e os fluxos das redes desmonta a verticalidade do sistema clássico de aprendizado, não sendo mais necessário seguir uma

TAGS MAIS COMUNS NOS POSTS

Anaconda Analytics Análise de Negócios Apache Spark AWS Big Data Blockchain Business Intelligence Chief Data Officer Cientista de Dados Cientistas de Dados Ciência de Dados Cloud Computing DaaS Data Lake Data Science Data Scientist Data Warehouse Deep Learning Deploy Descriptive Analytics Diagnostic Analytics Engenheiro de Dados Estatística GPU Hadoop Inteligência Artificial Internet of Things Linguagem Python Linguagem R Machine Learning MapReduce Mercado Financeiro NoSQL NVIDIA Open Data Oracle PaaS Predictive Analytics Prescriptive Analytics Probabilidade Python SaaS Salários Data Science Visualização

HISTÓRICO DE POSTS

abril 2020 (1)

março 2020 (1)

fevereiro 2020 (2)

setembro 2019 (1)

agosto 2019 (2)

julho 2019 (1)

abril 2019 (1)

março 2019 (1)

janeiro 2019 (1)

dezembro 2018 (1)

outubro 2018 (1)

setembro 2018 (2)

julho 2018 (1)

junho 2018 (3)

maio 2018 (1)

abril 2018 (2)

março 2018 (1)

fevereiro 2018 (2)

janeiro 2018 (1)

dezembro 2017 (1)

novembro 2017 (1)

outubro 2017 (1)

setembro 2017 (1)

julho 2017 (1)

junho 2017 (1)

maio 2017 (2)

abril 2017 (1)

janeiro 2017 (1)

novembro 2016 (1)

outubro 2016 (1)

setembro 2016 (1)

julho 2016 (1)

junho 2016 (1)

maio 2016 (1)

abril 2016 (1)

fevereiro 2016 (1)

janeiro 2016 (3)

dezembro 2015 (4)

novembro 2015 (6)

outubro 2015 (9)

setembro 2015 (9)

agosto 2015 (9)