ANÁLISE ESTATÍSTICA E MODELAGEM PREDITIVA DE SÉRIES TEMPORAIS EM PYTHON (DSA)

Aula 1 – O Que São Séries Temporais?

Uma série temporal é um conjunto sequencial de pontos de dados, medido tipicamente em tempos sucessivos.

É matematicamente definido como um conjunto de vetores x (t), t = 0,1,2, ... onde t representa o tempo decorrido. A variável x(t) é tratada como uma variável aleatória.

Espere, o que é uma variável aleatória?

Uma variável aleatória é uma variável quantitativa, cujo resultado (valor) depende de fatores aleatórios. Um exemplo de uma variável aleatória é o resultado do lançamento de um dado que pode ter qualquer número entre 1 e 6 como resultado. Embora possamos conhecer os seus possíveis resultados, o resultado em si depende de fatores de sorte (álea). Uma variável aleatória pode ser uma medição de um parâmetro que pode gerar valores diferentes. O conceito de variável aleatória é essencial em estatística e em outros métodos quantitativos para a representação de fenômenos incertos.

As variáveis aleatórias podem ser classificadas em variáveis aleatórias discretas, contínuas e mistas.

Voltando às séries temporais:

As medições realizadas durante um evento em uma série temporal são organizadas em uma ordem cronológica adequada.

Uma série temporal contendo registros de uma única variável é denominada como univariada e mais de uma variável como multivariada. O tempo não é uma variável dos dados, mas sim o índice dos dados.

As séries temporais são amplamente usadas no mundo dos negócios:

- Previsão de vendas - previsão de vendas de produtos ou serviços.
- Previsão de demanda - usada no gerenciamento de preços, inventário e força de trabalho.
- Previsão de tráfego - otimização de transporte e rotas, projeto de instalações rodoviárias.
- Previsão de receita - orçamento, definição de metas.

A previsão de séries temporais é uma área importante em Machine Learning, que geralmente é negligenciada.

É importante porque existem muitos problemas de previsão que envolvem um componente de tempo. Esses problemas são negligenciados porque é esse componente do tempo que dificulta o manuseio dos problemas de séries temporais.

Um conjunto de dados “comum” de aprendizado de máquina é uma coleção de observações.

Por exemplo:

Registro 1 – idade, altura, peso – 34 anos, 175 cm, 80 Kg
Registro 2 – idade, altura, peso – 42 anos, 182 cm, 93 Kg
Registro 3 – idade, altura, peso – 29 anos, 167 cm, 59 Kg

Podemos usar as variáveis idade e altura para prever o peso de uma pessoa.

São feitas previsões para novos dados quando o resultado real pode não ser conhecido até alguma data futura. O futuro está sendo previsto, mas todas as observações anteriores são quase sempre tratadas de forma igual. Talvez com algumas dinâmicas temporais muito pequenas para superar a ideia de “desvio de conceito”, como usar apenas o último ano de observações em vez de todos os dados disponíveis.

Mas no problema acima, não seria interessante considerar a mudança de peso ao longo do tempo? Afinal, ninguém mantém exatamente o mesmo a vida inteira (ou pelo menos é muito difícil).

Um conjunto de dados de séries temporais é diferente.

As séries temporais adicionam uma dependência explícita da ordem entre as observações: uma dimensão temporal.

Registro 1 – data, idade, altura, peso – 02/01/2020, 34 anos, 175 cm, 80 Kg
Registro 2 – data, idade, altura, peso – 03/01/2020, 42 anos, 182 cm, 93 Kg
Registro 3 – data, idade, altura, peso – 04/01/2020, 29 anos, 167 cm, 59 Kg

Essa dimensão adicional de tempo é uma restrição e uma estrutura que fornece uma fonte de informações adicionais.

Portanto, uma série temporal é uma sequência de observações tomadas sequencialmente no tempo.

Análise de Séries Temporais

Ao usar Estatística Clássica, a principal preocupação é a análise de séries temporais.

A análise de séries temporais envolve o desenvolvimento de modelos que melhor capturam ou descrevem uma série temporal observada para entender as causas subjacentes. Este campo de estudo busca o "porquê" por trás de um conjunto de dados de séries temporais.

Isso geralmente envolve fazer suposições sobre a forma dos dados e decompor as séries temporais em componentes.

A qualidade de um modelo descritivo é determinada por quão bem ele descreve todos os dados disponíveis e a interpretação que fornece para melhor informar o domínio do problema.

O objetivo principal da análise de séries temporais é desenvolver modelos matemáticos que forneçam descrições plausíveis a partir de amostras de dados.

Previsão de Séries Temporais

Fazer previsões sobre o futuro é chamado de extrapolação no tratamento estatístico clássico de dados de séries temporais. Os campos mais modernos se concentram no tópico e se referem a ele como previsão de séries temporais.

A previsão envolve ajustar os modelos aos dados históricos e usá-los para prever observações futuras.

Os modelos descritivos podem emprestar para o futuro (ou seja, para suavizar ou remover o ruído), eles apenas procuram melhor descrever os dados.

Uma distinção importante na previsão é que o futuro está completamente indisponível e só deve ser estimado a partir do que já aconteceu.

Portanto, temos objetivos diferentes, dependendo de estarmos interessados em entender um conjunto de dados ou fazer previsões.

A compreensão de um conjunto de dados, chamado análise de séries temporais, pode ajudar a fazer melhores previsões, mas não é necessária e pode resultar em um grande investimento técnico em tempo e experiência, não diretamente alinhados com o resultado desejado, que está prevendo o futuro.

Na modelagem descritiva ou análise de séries temporais, uma série temporal é modelada para determinar seus componentes em termos de padrões sazonais, tendências, relação a fatores externos e similares. Por outro lado, a previsão de séries temporais usa as informações em uma série temporal (talvez com informações adicionais) para prever valores futuros dessa série, o que também chamamos de Forecasting.

Aí está aliás a diferença entre Modelagem Estatística e Modelagem Preditiva. O primeiro visa descrever e compreender os dados e o segundo visa fazer previsões a partir de dados históricos. Podemos usar um modelo de regressão linear, por exemplo, para as duas tarefas. Tudo depende, como sempre, do objetivo a ser alcançado. E quando adicionamos a isso conhecimento em Programação, Ciência da Computação (Armazenamento e Processamento Distribuído) e expertise em Áreas de Negócio, tudo isso junto, é o que chamamos de Ciência de Dados ou Data Science.

Essa diferenciação entre Modelagem Estatística e Modelagem Preditiva é mostrada em detalhes no curso Data Science Aplicada à Área de Saúde, da FIAMED, quando o mesmo tipo de modelo é usado para tarefas completamente diferentes.

Esta introdução era para ser breve e rápida, mas desculpe. Não conseguimos montar aulas breves e rápidas. Densidade de conteúdo é a marca registrada da DSA. Não existe atalho para o aprendizado.