

CONSTELLATION

ASSET MANAGEMENT

# CONSTELLATION TALKS #3

Previendo Séries Temporais

Novembro | 2022

```
32 self.file = None
33 self.fingerprints = set()
logdups = True
debug = debug
logger = logging.getLogger(__name__)
th:
self.file = open(os.path.join(path, 'temp
self.file.seek(0)
self.fingerprints.update(re.sub(r'[\s\S]
method
n_settings(cls, settings):
ug = settings.getbool('SUPERFILTER_JUNK')
urn cls(job_dir(settings), debug)
quest_seen(self, request):
= self.request_fingerprint(request)
fp in self.fingerprints:
return True
self.fingerprints.add(fp)
f self.file:
self.file.write(fp + os.linesep)
request_fingerprint(self, request):
return request_fingerprint(request)
```

# Quem é a **Constellation?**

SOMOS UMA GESTORA DE  
INVESTIMENTOS **QUE  
BUSCA GERAR VALOR  
NO LONGO PRAZO**

# Quem é a **pessoa** **que vos fala?**

**LEONARDO PAZ**

Head of Software Development  
and Technology

Gamer nas horas que sobram

Machine Learning Enthusiast

# Requisitos

Precisamos garantir que quem está assistindo já possua alguns conhecimentos básicos

| Noções de Aprendizado de Máquina e estatística





Encontre o conteúdo em  
**<https://github.com/Constellation-Dev-Team>**

# Agenda



- | O que é aprendizado de máquina?
- | Como uma máquina pode aprender?
- | O que são séries temporais
- | Observando o comportamento no tempo
- | Diferenças ao prever séries temporais
- | Estacionariedade
- | Decomposição de séries temporais
- | Exemplos

# O que é aprendizado de máquina?

Campo de estudo que dá aos computadores a habilidade de aprender sem serem explicitamente programados

*Arthur Samuel (1959)*



# O que é aprendizado de máquina?

- Toda função matemática, estatística ou computacional que consegue extrair parâmetros de de uma série de dados para representar aquele conjunto de dados.
- Pode-se, posteriormente, usar essas funções e parâmetros para prever novas amostras desses conjunto de dados (Aprendizado Supervisionado).
- Pode-se, também, utilizar essa nova representação para identificar padrões nos dados (Aprendizado não-supervisionado).



# Tipos de aprendizado de máquina

| Supervisionado   | Não-Supervisionado   | Por Reforço  |
|--|--|--|
| Dados $X \in \mathbb{R}^n, y \in R$<br>Encontre $f$ , t.q.<br>$f(X) \mapsto y$   | Dado $X \in \mathbb{R}^n$ , encontre<br>$f$ , t.q. $f(X)$ extraia<br>alguma informação<br>relevante de $X$ | Dados um conjunto de ações<br>$A$ , um conjunto de<br>estados $S$ e uma função de<br>recompensa $R$ , encontre<br>$\pi(s, a) = \max_{a \in A, s \in S} R(s)$ |
| <i>Você conhece <math>X</math> e <math>y</math> e<br/>quer encontrar algo<br/>que preveja <math>y</math> baseado<br/>em <math>X</math></i> | <i>Você quer encontrar<br/>algum padrão relevante<br/>em <math>X</math>.</i>                               | <i>Você quer encontrar uma<br/>solução onde você não<br/>consegue definir certos e<br/>errados, apenas<br/>recompensar o for<br/>correto</i>                 |
| <i>Baseado no preço anterior,<br/>quero definir o próximo<br/>preço de uma ação</i>  | <i>Quero encontrar grupos<br/>semelhantes na minha<br/>amostra</i>   | <i>Quero que um robô jogue<br/>xadrez</i>  |

# Aprendizado Supervisionado

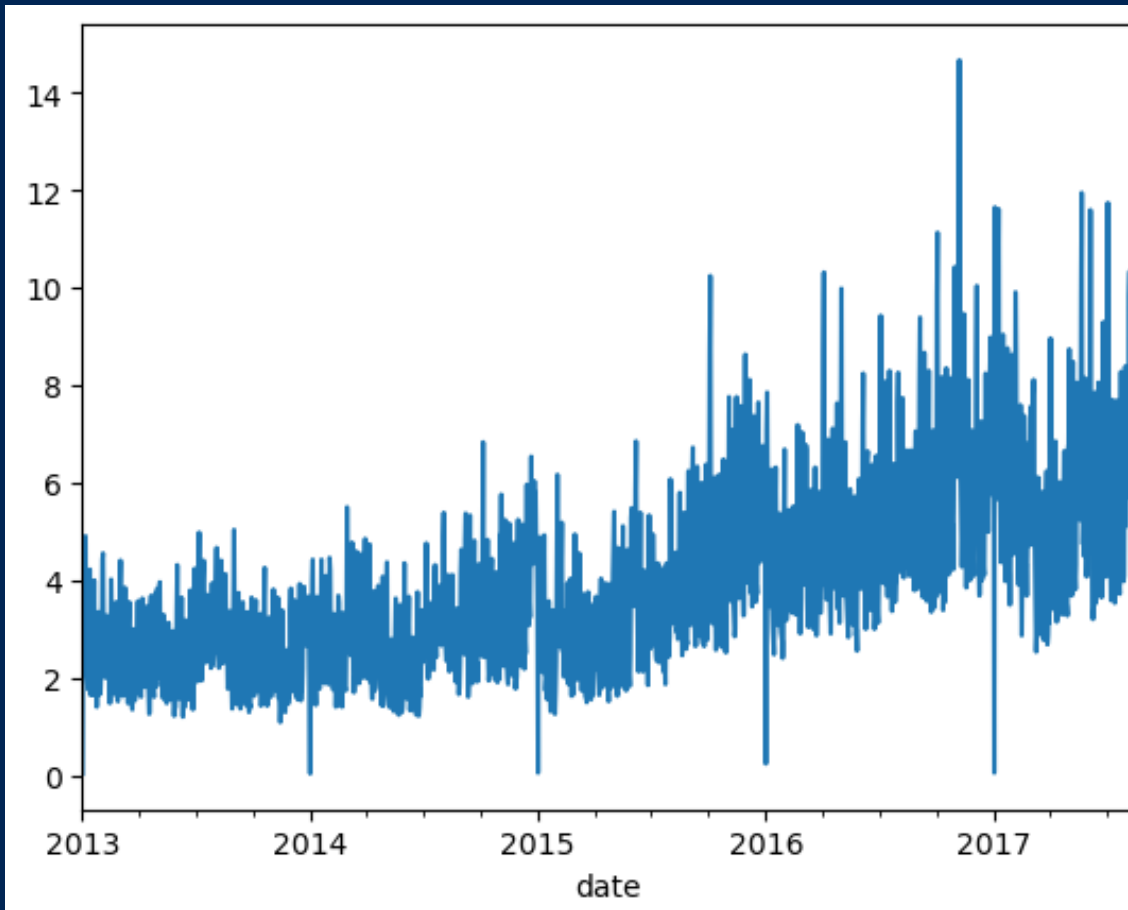
| Regressão  | Classificação   |
|--|---|
| Dados $X \in \mathbb{R}^n, y \in R$<br>Encontre $f$ , t.q.<br>$f(X) \mapsto y$ | Dados $X \in \mathbb{R}^n, y \in \{1,0\}^m$<br>Encontre $f$ , t.q.<br>$f(X) \mapsto y$              |
| <i><math>y</math> é sempre um número real</i>                                  | <i><math>y</math> é um conjunto de classes finitas definidos por inteiros ou um vetor "One Hot"</i> |
| <i>Baseado no preço anterior, quero definir o próximo preço de uma ação</i>    | <i>Baseado no preço anterior, quero se no próximo período a ação vai subir ou cair</i>              |

# O que é uma série temporal?

- | Uma série temporal é um conjunto de dados INDEXADO pelo tempo.
- | Em geral, tentamos usar um modelo de regressão. Também pode ser utilizado com classificação, mas possui um uso menos frequente.
- | Existe uma noção natural e intrínseca de sequência e ordenação nos dados

# Exemplo

Dado uma amostra de dados, podemos encontrar uma função linear



# Prevendo uma série temporal

Mas o que poderia ser modelado dessa forma?

- Garantir algum grau de estacionariedade
- Garantir que a variável prevista é uma variável aleatória independente e identicamente distribuída



# Observando um comportamento no tempo

A realidade é diferente da teoria:

- Nem todo dado de uma série pode ser explicado pela série
- Diferente de outros conjuntos de dados, impactos aleatórios aqui são mais graves e tendem a fazer a predição muito mais volátil.
- Boa parte dos usos reais de predição de séries temporais contem variáveis externas que não estão incluídas no conjunto dado

# Por quê precisamos tratá-las de forma diferente?

É muito fácil cometer erros de Look-ahead

Muito fácil criar modelos que não performam bem fora do conjunto de dados (overfit)

Métricas pouco assertivas para predição de casos reais


# Como podemos prever séries temporais?

- A chave da predição de séries temporais está na criação de um modelo que consegue identificar os padrões cíclicos e não-cíclicos dos dados
- Estacionariedade e Variável IID
- Padrões não-cíclicos:
  - Tendências globais
  - Tendências locais
- Padrões cíclicos:
  - Sazonalidade
  - Outros padrões cíclicos





# Como endereçar os problemas anteriores?

- Separação Treino/Desenvolvimento/Teste sequencial
  - Walking-Forward Validation
  - **Não utilizar validação cruzada normal**
  - Combinatorial Purged Cross Validation em alguns casos  
(Marco Lopes de Prado, 2016)
- 

# Estacionariedade e Var. IID

Antes de seguirmos, vamos escrever essas equações em formato matricial para facilitar nosso futuro

- Podemos utilizar o Augmented Dickey-Fuller Test para verificar estacionariedade da série
- Separar os dados em sets de treino, desenvolvimento e teste podem ser uma forma de garantir que o processo observado é IID
- Caso a variável não seja completamente IID, ela pode ser IID por um período de tempo, ou pelo menos podemos assumir isso.
- Walking-Forward Validation é uma forma de testar períodos onde o processo seja IID

# Decomposição da série temporal

## Modelo Aditivo

$$P(t) = trend + seasonality + ciclic + noise$$

## Modelo Multiplicativo

$$P(t) = trend * seasonality * ciclic * noise$$

# Componentes

- Tendência: componente permanente que permeia toda a série
  - Linear
  - Não linear
- Seasonalidade: componente cíclico que depende do calendário
  - Minutos, Horas, Dias, Semanas, Meses, Semestres e Anos
- Ciclos: demais componentes cíclicos que não depende do calendário
  - Podemos extrair tendências cíclicas utilizando regressões sobre séries de Fourier e as séries
- Auto-regressivos: comportamentos que dependem dos valores anteriores da série

$$P(t) := f(P(t - 1))$$

# Ruído em Séries temporais

**Vamos considerar a equação:**

$$P(t) = \text{trend} + \text{seasonality} + \text{cyclic} + \text{noise}$$

- Uma regressão pode capturar parte do comportamento do ruído como valor significativo da série temporal.
- Séries em que a razão ruído-sinal seja alto, se tornam mais difíceis de prever para períodos fora da amostra (out of sample)

Algumas técnicas para reduzir o ruído:

- Médias móveis
- Filtro de Kalman
- AR models: ARMA, ARIMA, ARIMAX...

# Métricas de Erro em séries temporais

Erro quadrático médio (MSE)

Raiz do Erro quadrático médio (RMSE)

Erro Médio Absoluto (MAE)

Erro Médio Percentual Absoluto (MAPE)



Agora vamos para o código



# Referências

<https://www.kaggle.com/learn/time-series>

<https://machinelearningmastery.com/decompose-time-series-data-trend-seasonality/>

<https://machinelearningmastery.com/time-series-seasonality-with-python/>

<https://medium.com/@khairulomar/deconstructing-time-series-using-fourier-transform-e52dd535a44e>