Aprendizagem de Máquina

Advanced Institute for Artificial Intelligence

https://advancedinstitute.ai

Estatística

Agenda

- Bibliotecas
- Análise Estatística Descritiva
  - Tendência central
  - Dispersão
  - Interpretação
- Correlação
- Análise de probabilidades

Estatística

Bibliotecas Python para Estatística

- Scikit-Learn
  - Provê suporte a utilização de algoritmos de aprendizagem de máquina
- Scipy Stats
  - Função estatísticas elementares
- MatplotLib
  - Biblioteca para montar gráficos
- Pandas e Numpy
  - Algumas funções estatísticiticas e de gráficos estão embutidas nessas bibliotecas

Estatística

Análise Estatística

- Passo fundamental para utilização de técnicas de aprendizagem de máquina
- Aprofundar o entendimento quanto aos dados disponíveis
- Preparação adequada dos dados
- Ex: Dados faltantes
- Ex: Outliers
- Ex: Segmentação adequada

Estatística

Descrição de Dados

- Dados organizados em tabelas dificultam análises
- Gráficos são essenciais para realização de análises de modo mais prático
- Medidas de tendência central e variabilidade complementam tais análises e facilitam comparações

Estatística

Tipos de Dados e Gráficos

- Dados Qualitativos
  - Gráfico de barra
  - Gráfico de pizza
- Dados Quantitativos
  - Histograma
  - Box
  - Área
  - Linha
  - Scatter

Estatística

Estimativa de densidade por Kernel (KDE - Kernel Density Estimate)

- forma não-paramétrica para estimar a Função densidade de probabilidade de uma variável aleatória.
- Possui a propriedade de estimar de forma continua de acordo com um kernel adequado (curva normal por exemplo)
- Opção de visualização a histogramas

Estatística

Tendência central

- Média (μ): soma dos valores dividido pela quantidade de dados
- Mediana: valor situado no meio da amostra, quando a amostra está ordenada
- Moda: valor mais frequente na amostra

Estatística

Dispersão

- Desvio médio: módulo da média aritmética dos desvios de cada elemento da série para a média da série
- Variância: mesmo conceito do desvio médio trocando módulo por elevar a diferença ao quadrado
- Desvio Padrão (σ): é representado como a raiz quadrada da variância

Interpretação da Dispersão

- Quanto mais uniforme forem os valores, mais próximo de zero estará o desvio padrão.
- Quando todos valores são iguais o desvio padrão é zero. Assim a amostra é perfeitamente uniforme.
- Quando estamos interessados em saber qual conjunto de valores possui uma maior regularidade podemos usar tanto a variância, como o desvio padrão.
- O desvio padrão é expresso na mesma unidade de medida das variáveis do conjunto.

Estatística

Variância:

- Variância simples: soma do desvio quadrado de cada valor em relação a média dividido por n-1 (população -1)
- Desvio padrão simples: raiz quadrada da variância simples

- Quando utilizamos a população completa é comum utilizar a população -1 para ajuste estatístico

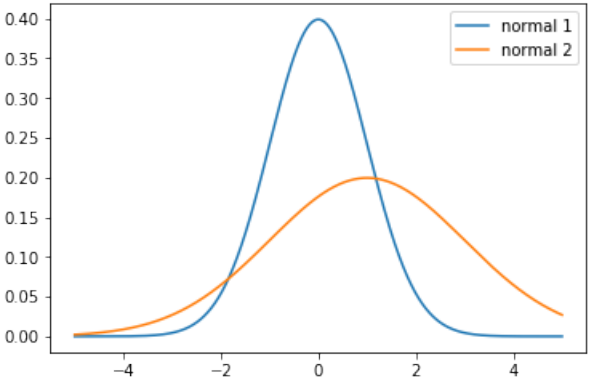
Estatística

Distribuição Normal

- É uma mais importantes distribuições de probabilidade que caracteriza muitos fenômenos aleatórios
  - Fenômenos naturais
  - Altura
  - pressão sangüínea
- Desempenham papel importante nos métodos de inferência estatística
- A distribuição é normal é uma variável aleatória contínua tem uma distribuição em forma de sino.

Estatística

Distribuições Normais com diferentes valores de média e desvio padrão

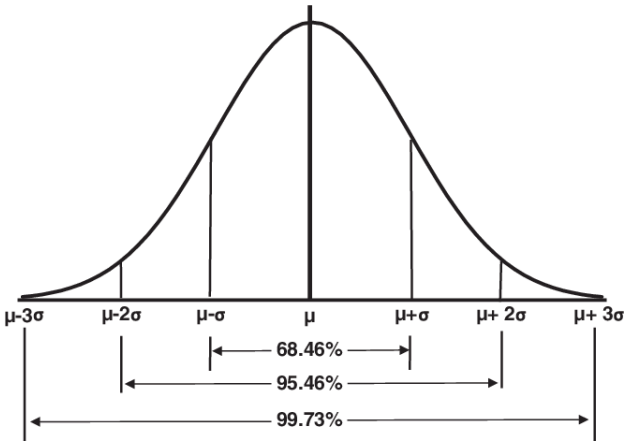


Estatística

Distribuições paramétricas

- regra empírica : define a probabilidade dos valores estarem em intervalos definidos pela média e desvio padrão
  - $\mu + \sigma$ e μ - σ: 68%
  - $\mu + 2*\sigma$ e μ - $2*\sigma$: 95%
  - $\mu + 3*\sigma$ e μ - $3*\sigma$: 99%
- z-score: representa a distância entre uma dada medida e a média em termos de desvio padrão

Estatística



Estatística

- Covariância é uma medida sada para comparar o comportamento de duas ou mais variáveis
- Mede como duas ou mais variáveis variam em conjunto de suas médias
- É possível identificar se diferentes variáveis possuem algum padrão comum entre si.

Estatística

- Por exemplo, uma variável que mede acidentes por dia em uma região e outra variável que mede velocidade média nessa mesma região.
- Tais padrões comuns permitem tomar conclusões a respeito da base em estudo
- Importante destacar que correlação não implica causalidade obrigatoriamente

Correlação

- Utiliza-se a covariância e o desvio padrão como base para definir métricas de correlação
- A correlação perto de -1 é uma anti-correlação perfeita
- A correlação perto de 0 indica que não há correlação
- A correlação perto de 1 é uma correlação perfeita