Aprendizagem de Máquina

Advanced Institute for Artificial Intelligence

https://advancedinstitute.ai

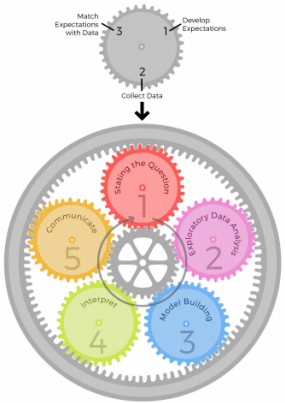
Estatística

Agenda

- Processo de Ciência de Dados
- Análise exploratória
- Distribuição Normal
- Correlação

Introdução a Estatística

Epíciclos de ciência dos dados [1]



[1] "The Art of Data Science A Guide for Anyone Who Works with Data". Roger D. Peng and Elizabeth Matsui. 2016

Introdução a Estatística

- Os epíciclos de ciência dos dados são etapas genéricas que podem ter complexidade variável
- Um projeto pode ser feito seguindo cada etapa em apenas um dia
- Projetos maiores podem demandar a execução de cada etapa por dias ou meses

Introdução a Estatística

Etapas do epiciclo:

- Definir a questão
- Análise de Dados Exploratória
- Construção de um modelo
- Interpretação
- Comunicação

Cada etapa segue um processo de alinhamento de expectativa

- Estabelecendo Expectativas
- Coletando informações (dados), comparando disponibilidade de dados e expectativas
- Revisão de expectativas e/ou ajuste de dados para que seu dados e suas expectativas correspondam.

Introdução a Estatística

A definição da questão

- Descritivas: sumarização ou consolidação de um conjunto de dados
- Exploratória: busca por padrões, tendências, a partir dos resultados é possível criar hipóteses
- Inferência: buscar relações entre os dados que confirmem ou refutem uma hipótese, por exemplo mostrar que uma amostra é singifinicativa de uma população, comparando variáveis que caracterizam a população

Introdução a Estatística

A definição da questão

- Preditiva: buscar quais padrões ou características nos dados levam aos aspectos que se pretende predizer
- Causal: buscar parâmetros que ao serem modificados em um conjunto de variáveis, provocam um detemrinado comportamento nos dados
- Mecanicista : uma busca direta por uma determinada conclusão, tendo como evidência os dados apresentados

Estatística

Análise Exploratória

- Exploração de dados para aprender mais sobre um fenômeno ou como pré-processamento para aprendizagem de máquina
- Aprofundar o entendimento quanto aos dados disponíveis
- Preparação adequada dos dados
- Avaliar se há Dados faltantes
- Presença de Outliers

Descrição de Dados

- Dados organizados em tabelas não são facilmente visualizáveis
- Gráficos permitem análises de modo mais prático e visual
- Medidas de tendência central e variabilidade complementam tais análises e facilitam comparações

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

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.

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

Distribuição

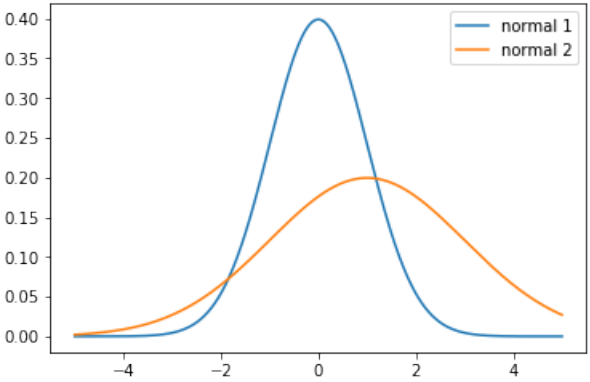
- Uma distribuição define uma curva (gráfico), e a área sob a curva define a probabilidade de acontecer um evento relacionado à distribuição.
- Matemáticos definiram muitas distribuições "famosas", que representam fenômenos importantes do mundo real.

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

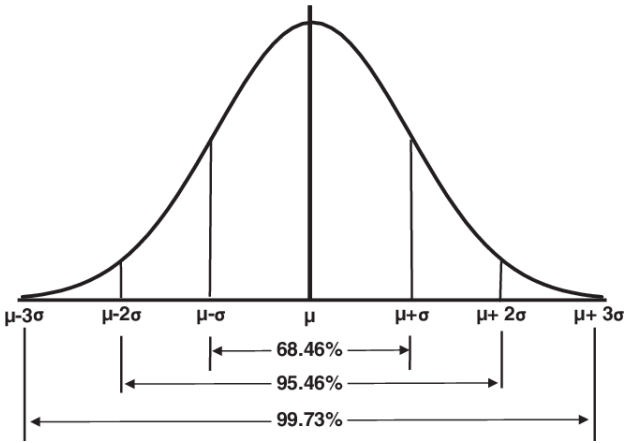
Distribuições Normais com diferentes valores de média e desvio padrão



Distribuições paramétricas

- Supondo que os dados seguem uma distribuição, os parâmetros da distribuição permitem calcular a probabilidade de um fenômeno (no caso da distribuição norma, os parâmetros são média e desvio padrão)
  - $\mu + \sigma$ e μ - σ: 68%
  - μ + 2*σ e μ - 2*σ: 95%
  - μ + 3*σ e μ - 3*σ: 99%
- z-score: representa a distância entre uma dada medida e a média em termos de desvio padrão

Estatística



Estatística

- Covariância é uma medida usada 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

Correlação

- Utilizada para medir o relacionamento entre 2 variáveis - útil para remover variáveis ambíguas.
- **Pearson**: Identifica correlação *linear* entre variáveis (se uma variável varia, a outra varia proporcionalmente)
- **Spearman**: Variáveis variam "juntas", mas não necessariamente linearmente.
- **Kendall**: Mais preciso para medir correlação em conjuntos de treinamento de tamanho reduzido.