Mineração de Dados

Tratamento de Dados pt2.

Prof.

Fernando S. de Aguiar Neto

Sumário

Discretização

Motivação

- Vimos algoritmos que exigem dados categóricos e algoritmos que exigem dados numéricos
- E se tivermos uma base com dados contínuos e precisarmos efetuar contagens?
- Precisaríamos transformar esses números em faixas numéricas

Discretização

- O termo usado para esse tipo de transformação é discretização
- Pois passaremos os dados de um domínio contínuo para um domínio discreto[1]

[1] contável, limitado

Exemplo

```
array([0.09686948, 2.27929692, 2.35933674, 1.43318587, 1.3419051 ,
       1.57901547, 2.67600397, 3.6918435 , 0.0627299 , 3.2984076 ,
       1.4778862 , 2.08394509, 1.73778478, 1.96309427, 0.4918363 ,
       2.06207856, 2.64528463, 1.21344181, 0.        , 1.70406866,
       3.4846894 , 2.75117159, 1.4396345 , 0.59413107, 2.94872034,
       3.30536107, 0.58561338, 1.91985733, 2.96036234, 2.0246309 ])
```

Figura 1: 30 números aleatórios de uma distribuição normal (media≈0, std≈1)

Exemplo

- Esperamos que mais números estejam próximos à média
- Mas se contarmos os números, notaremos que cada número aparece apenas uma vez..

Exemplo

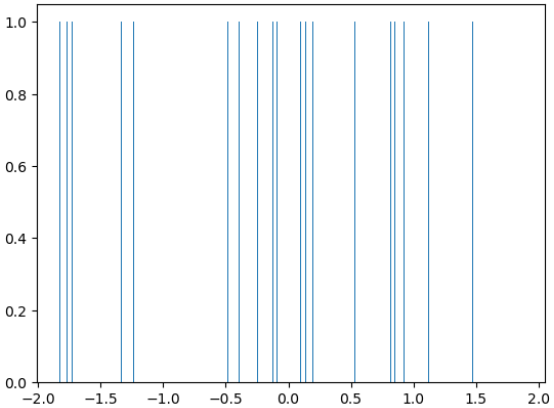


Figura 2: Contagem de valores únicos

Binning

- Para solucionar esse problema iremos utilizar a técnica chamada *binning*
- Iremos dividir esse espaço em 15 intervalos, 'caixas'[2], de tamanho igual
- As amostras que estiverem no intervalo irão contar para todo o intervalo
- Notem que mudamos a ideia, não estamos mais contando os **valores** e sim **intervalos**

[2] *Bins*

Binning - Exemplo 15

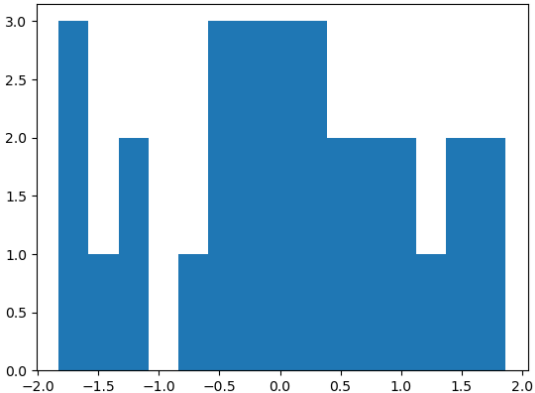


Figura 3: Contagem, dividindo o intervalo em 15 partes

Binning - Exemplo 10

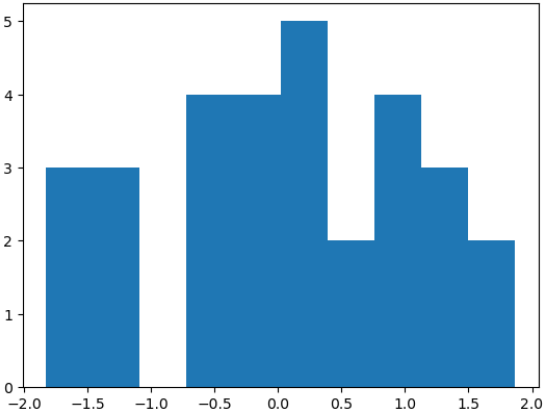


Figura 4: Contagem, dividindo o intervalo em 10 partes

Binning - Exemplo 5

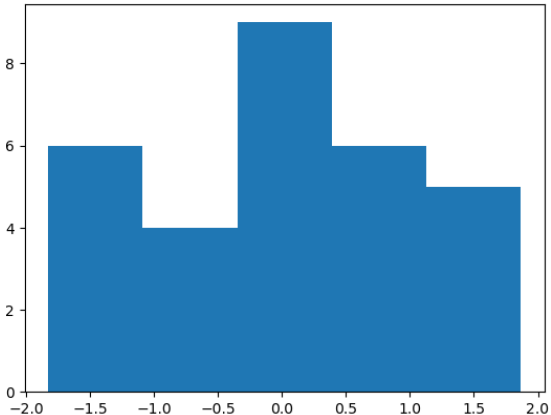


Figura 5: Contagem, dividindo o intervalo em 5 partes

Binning - Exemplo 3

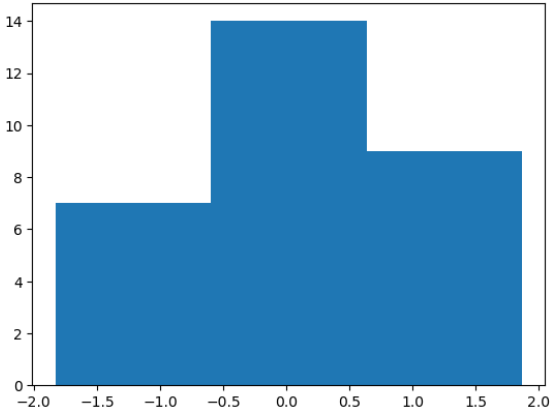


Figura 6: Contagem, dividindo o intervalo em 3 partes

Binning - Exemplo 2

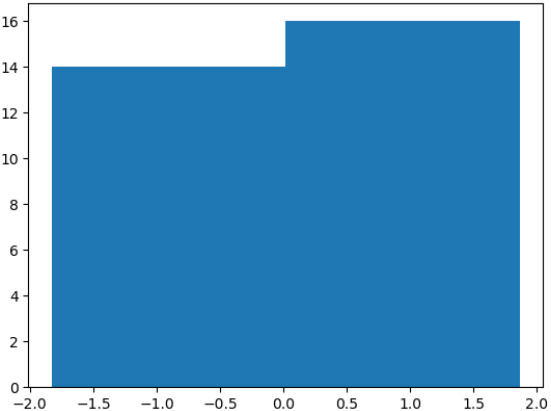


Figura 7: Contagem, dividindo o intervalo em 2 partes!

Como escolher a quantidade de bins?

- Notem que conforme alteramos a quantidade de a distribuição começa a se aproximar do que esperamos em uma distribuição normal
- Temos apenas 30 amostras e é por isso que precisamos deixar tão poucos intervalos

Como escolher a quantidade de bins?

- Podemos ir variando a quantidade de intervalos até notarmos que se aproxima da distribuição esperada
- Podemos também definir um quantidade esperada de intervalos, e.g. 3 (pouco, médio, muito)
- Vamos ver o efeito da escolha da quantidade de bins, em uma amostra de 3000

Binning - 3000 amostras

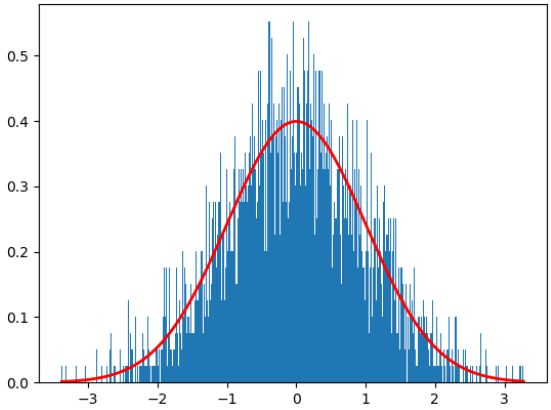


Figura 8: Função Densidade de Probabilidade, dividindo o intervalo em 500 partes

Binning - 3000 amostras

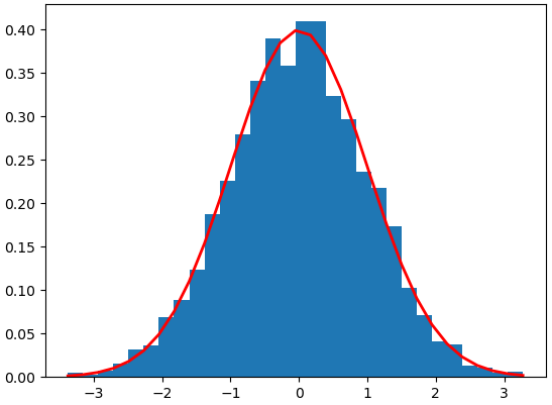


Figura 9: Função Densidade de Probabilidade, dividindo o intervalo em 30 partes

Binning - 3000 amostras

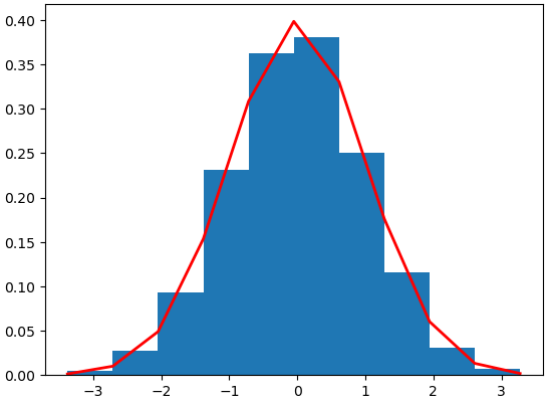


Figura 10: Função Densidade de Probabilidade, dividindo o intervalo em 10 partes

Discussões

- Precisamos deixar sempre intervalos de tamanho igual?
- E se tivéssemos situações como as seguintes?

Binning - 3000 amostras

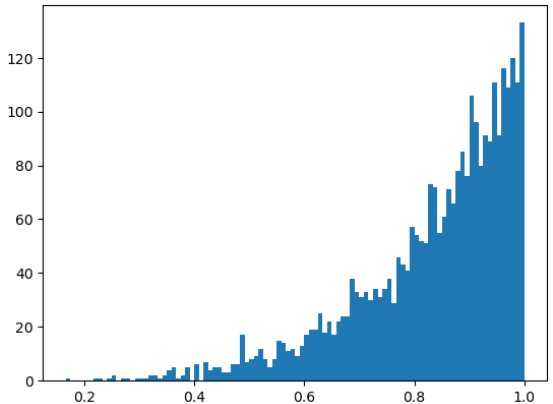


Figura 11: Contagem, distribuição de potência (exp=5), dividindo o intervalo em 100 partes

Binning - 3000 amostras

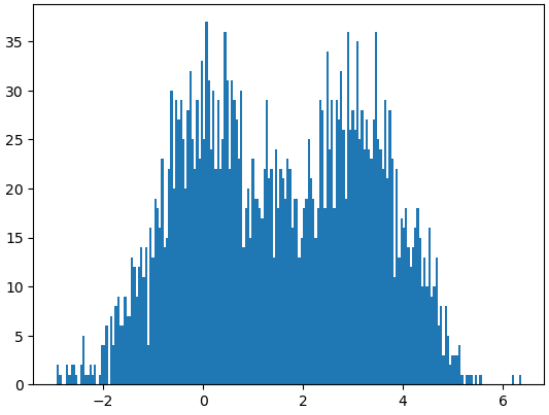


Figura 12: Contagem, duas normais próximas, dividindo o intervalo em 100 partes

Outras aplicações

- Essa técnica pode ser útil para permitir métodos que exijam contagem
- Mas também é útil se precisarmos alterar o sentido do que estamos estudando
- e.g. número de tênis para infantil, adulto
- e.g. dividir salário em baixo e alto
- e.g. Ou como vimos na base Supermarket, o gasto foi dividido em pouco ou muito

Categorico para numérico

- E se nosso algoritmo esperar dados numéricos, e.g. k-means
- Podemos transformar nossas categorias em números?

Categorico para numérico

- CUIDADO!
- Ao discretizar, perdemos um pouco de informação ao juntar vários valores diferentes nas mesmas 'caixinhas'
- Ao pegar uma categoria e transformá-la em número, estamos adicionando informação que **não existe**, o que pode criar situações indesejadas, esse tipo de transformação exige cautela

Imputando Valores

1 Discretização

2 Imputando Valores

- Estatísticas Descritivas
- Vizinhança
- Vizualização e Agrupamento

Motivação

Muitas vezes valores estarão faltando em bases de dados, algumas causas:

- Falha em sensor/medidor
- Erro de digitação
- Questões legais, e.g. privacidade

Motivação

- Não queremos descartar **toda** a informação da amostra por causa de **um** atributo faltando
- Como preencher esses valores que faltam?

Preocupações

- Estamos adicionando informação ao imputar um valor
- Isso irá alterar os dados de entrada e consequentemente as conclusões obtidas
- Queremos imputar um valor que traga ao mínimo de mudanças/viés

Estatísticas Descritivas

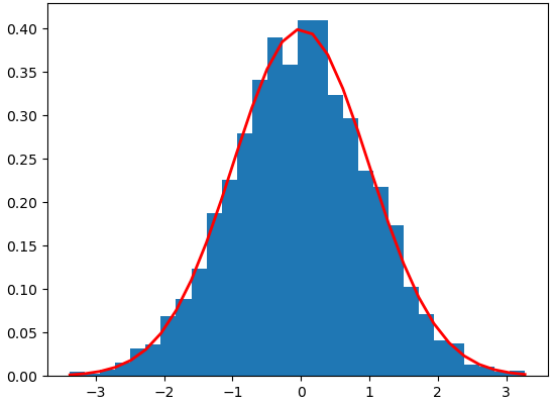
Estatística Descritiva

- Usando e entendendo as estatísticas descritivas da base podemos imputar um valor que seja próximo ao valor esperado
- I.e. possuem o menor erro, serão uma boa aproximação do valor real

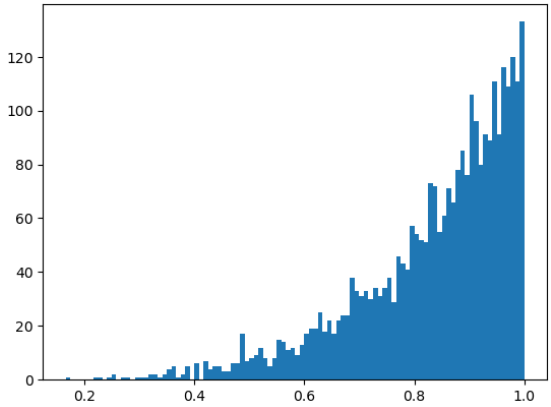
Estatística Descritiva

- Uma estratégia é imputar um valor que seja central
- Média?
- Mediana?
- Moda?

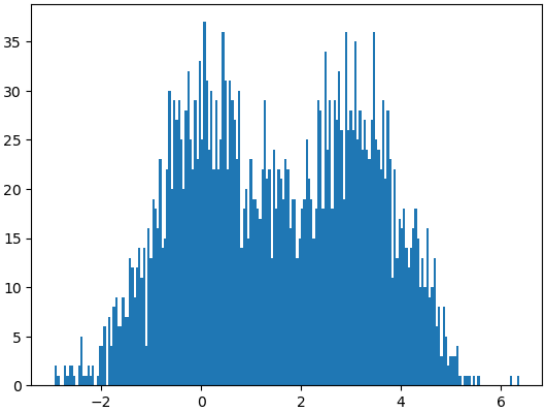
Estatística Descritiva - Centralidade



Estatística Descritiva - Centralidade



Estatística Descritiva - Centralidade



Multiplas Imputações

- Nada impede que imputemos um valor diferente em cada amostra
- e.g. na imagem anterior, temos duas ‘modas’, podemos decidir qual é a adequada amostra a amostra
- Podemos também aplicar uma ‘média local’, por exemplo se temos amostras rotuladas, aplicamos a média dentro do rótulo

Vizinhança

Motivação

- Com a ideia de imputar valores diferentes para cada amostra podemos pensar em quais as amostras mais parecidas
- Essa ideia permite corrigir por grandes variações, desde que valores sejam localmente parecidos

Medidas de Distância

A escolha da medida de distância depende dos dados disponíveis, seguem alguns exemplos:

- Manhattan ou Euclidiana para dados numérico/espaciais
- Correlação de Pearson para vetores como as transações do mercado
- Hamming para dados puramente categóricos

K-Vizinhos Mais Próximos

- Um algoritmo que pode ser aplicado para imputar baseado nos k vizinhos mais próximos
- A proximidade é medida usando os outros atributos (não-faltantes)
- e o valor é obtido em função[3] do valor dos vizinhos mais próximos

[3] média, consenso, escolha aleatória, etc.

Discussão

- Precisamos medir distâncias entre amostras, convém normalizar?
- Quais problemas podem aparecer?
- Quais vantagens e desvantagens de escolher muitos vizinhos ou apenas 1?

Vizualização e Agrupamento

Motivação

- Uma forma 'natural' de encontrar comunidades de amostras localmente parecidas é o uso de agrupamento

Usando K-Means

- Aplicamos o algoritmo normalmente, omitindo amostras com valores faltantes
- O centróide será um representante da amostra .. central de cada grupo, portanto podemos usar os valores dele como valor a ser imputado

Usando K-Means

- Devemos calcular a pertinência das amostras com valores faltantes
- Isso pode ser feito por distância ao centróide, ou inspeção visual
- Então usamos os dados do centróide para imputar valores na amostra

Discussão

- Precisamos medir distâncias entre amostras, convém normalizar?
- Quais problemas podem aparecer?
- Quais vantagens e desvantagens de escolher muitos grupos?