Sistemas Inteligentes Aplicados

Carlos Hall

Programa do Curso

- Limpeza/Integração de Dados
- Transformação de Dados
  - Discretização de Variáveis Contínuas
  - Transformação de Variáveis Discretas em Contínuas
  - Transformação de Variáveis Contínuas
- Análise e Seleção de Variáveis (Redução de Dados)

Transformação de Dados

- Discretização de Variáveis Contínuas/Transformação de Variáveis Discretas em Contínuas
  - Adequação aos métodos inteligentes a serem utilizados posteriormente
  - Melhoria de desempenho
- Transformação de Variáveis Contínuas
  - Melhoria na distribuição dos dados
  - Melhoria de desempenho dos métodos inteligentes

Discretização de Variáveis Contínuas

- Transforma atributos contínuos em atributos categóricos
  - Absolutamente essencial se o método inteligente só manuseia atributos categóricos
  - Em alguns casos, mesmo métodos que manuseiam atributos contínuos têm melhor desempenho com atributos categóricos

Discretização de Variáveis Contínuas

- Diversos métodos de discretização
  - Discretização pelo Método 1R (1-rule)
  - Discretização Não-supervisionada

Discretização de Variáveis Contínuas

- Discretização pelo Método 1R (1-rule)
  - Sub-produto de uma técnica de extração automática de regras
  - Utiliza as classes de saída para discretizar cada atributo de entrada separadamente
    - Ex: Base de dados hipotética de meteorologia x decisão de realizar ou não um certo jogo

Discretização pelo Método 1R (1-rule)

- Base de Dados Meteorológicos

Tempo	Temperatura	Umidade	Vento	Jogar? (CLASSE)
Sol	85	85	Não	Não
Sol	80	90	Sim	Não
Nublado	83	86	Não	Sim
Chuva	70	96	Não	Sim
Chuva	68	80	Não	Sim
Chuva	65	70	Sim	Não
Nublado	64	65	Sim	Sim
Sol	72	95	Não	Não
Sol	69	70	Não	Sim
Chuva	75	80	Não	Sim
Sol	75	70	Sim	Sim
Nublado	72	90	Sim	Sim
Nublado	81	75	Não	Sim
Chuva	71	91	Sim	Não

Discretização pelo Método 1R (1-rule)

- Primeiro passo: ordenar pela coluna Temperatura

Tempo	Temperatura	Umidade	Vento	Jogar? (CLASSE)
Nublado	64	65	Sim	Sim
Chuva	65	70	Sim	Não
Chuva	68	80	Não	Sim
Sol	69	70	Não	Sim
Chuva	70	96	Não	Sim
Chuva	71	91	Sim	Não
Sol	72	95	Não	Não
Nublado	72	90	Sim	Sim
Chuva	75	80	Não	Sim
Sol	75	70	Sim	Sim
Sol	80	90	Sim	Não
Nublado	81	75	Não	Sim
Nublado	83	86	Não	Sim
Sol	85	85	Não	Não

Discretização pelo Método 1R (1-rule)

- Segundo passo: discretizar pela Classe de saída

Tempo	Temperatura	Umidade	Vento	Jogar? (CLASSE)
Nublado	64	65	Sim	Sim
Chuva	65	70	Sim	Não
Chuva	68	80	Não	Sim
Sol	69	70	Não	Sim
Chuva	70	96	Não	Sim
Chuva	71	91	Sim	Não
Sol	72	95	Não	Não
Nublado	72	90	Sim	Sim
Chuva	75	80	Não	Sim
Sol	75	70	Sim	Sim
Sol	80	90	Sim	Não
Nublado	81	75	Não	Sim
Nublado	83	86	Não	Sim
Sol	85	85	Não	Não

Discretização pelo Método 1R (1-rule)

- Segundo passo: discretizar pela Classe de saída

Tempo	Temperatura	Umidade	Vento	Jogar? (CLASSE)
Nublado	64	65	Sim	Sim
Chuva	65	70	Sim	Não
Chuva	68	80	Não	Sim
Sol	69	70	Não	Sim
Chuva	70	96	Não	Sim
Chuva	71	91	Sim	Não
Sol	72	95	Não	Não
Nublado	72	90	Sim	Sim
Chuva	75	80	Não	Sim
Sol	75	70	Sim	Sim
Sol	80	90	Sim	Não
Nublado	81	75	Não	Sim
Nublado	83	86	Não	Sim
Sol	85	85	Não	Não

Discretização pelo Método 1R (1-rule)

- Terceiro passo: ajustar divisões

Tempo	Temperatura	Umidade	Vento	Jogar? (CLASSE)
Nublado	64	65	Sim	Sim
Chuva	65	70	Sim	Não
Chuva	68	80	Não	Sim
Sol	69	70	Não	Sim
Chuva	70	96	Não	Sim
Chuva	71	91	Sim	Não
Sol	72	95	Não	Não
Nublado	72	90	Sim	Sim
Chuva	75	80	Não	Sim
Sol	75	70	Sim	Sim
Sol	80	90	Sim	Não
Nublado	81	75	Não	Sim
Nublado	83	86	Não	Sim
Sol	85	85	Não	Não

Discretização pelo Método 1R (1-rule)

- Terceiro passo: ajustar divisões

Tempo	Temperatura	Umidade	Vento	Jogar? (CLASSE)
Nublado	(1) 64	65	Sim	Sim
Chuva	65 (2)	70	Sim	Não
Chuva	68	80	Não	Sim
Sol	(3) 69	70	Não	Sim
Chuva	70	96	Não	Sim
Chuva	71 (4)	91	Sim	Não
Sol	72	MUITAS DIVISÕES!		Não
Nublado	72			Sim
Chuva	(5) 75	80	Não	Sim
Sol	75	70	Sim	Sim
Sol	80 (6)	90	Sim	Não
Nublado	(7) 81	75	Não	Sim
Nublado	83	86	Não	Sim
Sol	85 (8)	85	Não	Não

Discretização pelo Método 1R (1-rule)

- Quarto passo: mínimo de valores da maior classe (ex: 3)

Tempo	Temperatura	Umidade	Vento	Jogar? (CLASSE)
Nublado	64	65	Sim	Sim
Chuva	65	70	Sim	Não
Chuva	68	80	Não	Sim
Sol	69	70	Não	Sim
Chuva	70	96	Não	Sim
Chuva	71	91	Sim	Não
Sol	72	95	Não	Não
Nublado	72	90	Sim	Sim
Chuva	75	80	Não	Sim
Sol	75	70	Sim	Sim
Sol	80	90	Sim	Não
Nublado	81	75	Não	Sim
Nublado	83	86	Não	Sim
Sol	85	85	Não	Não

Discretização pelo Método 1R (1-rule)

- Quarto passo: mínimo de valores da maior classe (ex: 3)

Tempo	Temperatura	Umidade	Vento	Jogar? (CLASSE)
Nublado	64	65	Sim	Sim
Chuva	65	70	Sim	Não
Chuva	68	80	Não	Sim
Sol	69	70	Não	Sim
Chuva	70	96	Não	Sim
Chuva	71	91	Sim	Não
Sol	72	95	Não	Não
Nublado	72	90	Sim	Sim
Chuva	75	80	Não	Sim
Sol	75	70	Sim	Sim
Sol	80	90	Sim	Não
Nublado	81	75	Não	Sim
Nublado	83	86	Não	Sim
Sol	85	85	Não	Não

Discretização pelo Método 1R (1-rule)

- Quarto passo: mínimo de valores (ex: 3)

Tempo	Temperatura	Umidade	Vento	Jogar? (CLASSE)
Nublado	64	65	Sim	Sim
Chuva	65	70	Sim	Não
Chuva	68 (1)	80	Não	Sim
Sol	69	70	Não	Sim
Chuva	70	96	Não	Sim
Chuva	71	91	Sim	Não
Sol	72	95	Não	Não
Nublado	72 (2)	90	Sim	Sim
Chuva	75	80	Não	Sim
Sol	75	70	Sim	Sim
Sol	80	90	Sim	Não
Nublado	81 (3)	75	Não	Sim
Nublado	83	86	Não	Sim
Sol	85	85	Não	Não

Discretização pelo Método 1R (1-rule)

- Quarto passo: mínimo de valores (ex: 3)

Tempo	Temperatura	Umidade	Vento	Jogar? (CLASSE)
Nublado	64	65	Sim	Sim
Chuva	65	70	Sim	Não
Chuva	68 (1)	80	Não	Sim
Sol	69	70	Não	Sim
Chuva	70	96	Não	Sim
Chuva	71	91	Sim	Não
Sol	72 (2)	95	Não	Não
Nublado	72	90	Sim	Sim
Chuva	75	80	Não	Sim
Sol	75	70	Sim	Sim
Sol	80	90	Sim	Não
Nublado	81 (3)	75	Não	Sim
Nublado	83	86	Não	Sim
Sol	85	85	Não	Não

Discretização pelo Método 1R (1-rule)

- Quarto passo: mínimo de valores (ex: 3)

Tempo	Temperatura	Umidade	Vento	Jogar? (CLASSE)
Nublado	64	65	Sim	Sim
Chuva	65	70	Sim	Não
Chuva	68 (1)	80	Não	Sim
Sol	69	70	Não	Sim
Chuva	70	96	Não	Sim
Chuva	71	91	Sim	Não
Sol	72	95	Não	Não
Nublado	72	90	Sim	Sim
Chuva	75	80	Não	Sim
Sol	75	70	Sim	Sim
Sol	80	90	Sim	Não
Nublado	81 (2)	75	Não	Sim
Nublado	83	86	Não	Sim
Sol	85	85	Não	Não

Discretização pelo Método 1R (1-rule)

- Quarto passo: mínimo de valores (ex: 3)

Tempo	Temperatura	Umidade	Vento	Jogar? (CLASSE)
Nublado	64	65	Sim	Sim
Chuva	65	70	Sim	Não
Chuva	68 (1)	80	Não	Sim
Sol	69	70	Não	Sim
Chuva	70	96	Não	Sim
Chuva	71			
Sol	72			
Nublado	72			
Chuva	75			
Sol	75	70	Sim	Sim
Sol	80	90	Sim	Não
Nublado	81 (2)	75	Não	Sim
Nublado	83	86	Não	Sim
Sol	85	85	Não	Não

Categoria 1: Temperatura ≤ 77.5
Categoria 2: Temperatura > 77.5

Discretização pelo Método 1R (1-rule)

- Base de Dados Meteorológicos - Umidade

Tempo	Temperatura	Umidade	Vento	Jogar? (CLASSE)
Sol	85	85	Não	Não
Sol	80	90	Sim	Não
Nublado	83	86	Não	Sim
Chuva	70	96	Não	Sim
Chuva	68	80	Não	Sim
Chuva	65	70	Sim	Não
Nublado	64	65	Sim	Sim
Sol	72	95	Não	Não
Sol	69	70	Não	Sim
Chuva	75	80	Não	Sim
Sol	75	70	Sim	Sim
Nublado	72	90	Sim	Sim
Nublado	81	75	Não	Sim
Chuva	71	91	Sim	Não

Discretização pelo Método 1R (1-rule)

- Base de Dados Meteorológicos - Umidade

Tempo	Temperatura	Umidade	Vento	Jogar? (CLASSE)
Nublado	64	65	Sim	Sim
Chuva	65	70	Sim	Não
Sol	69	70	Não	Sim
Sol	75	70	Sim	Sim
Nublado	81	75	Não	Sim
Chuva	68	80	Não	Sim
Chuva	75	80	Não	Sim
Sol	85	85	Não	Não
Nublado	83	86	Não	Sim
Sol	80	90	Sim	Não
Nublado	72	90	Sim	Sim
Chuva	71	91	Sim	Não
Sol	72	95	Não	Não
Chuva	70	96	Não	Sim

Discretização pelo Método 1R (1-rule)

- Base de Dados Meteorológicos - Umidade

Tempo	Temperatura	Umidade	Vento	Jogar? (CLASSE)
Nublado	64	65	Sim	Sim
Chuva	65	70	Sim	Não
Sol	69	70 (1)	Não	Sim
Sol	75	70	Sim	Sim
Nublado	81	75	Não	Sim
Chuva	68	80	Não	Sim
Chuva	75	80	Não	Sim
Sol	85	85	Não	Não
Nublado	83	86	Não	Sim
Sol	80	90 (2)	Sim	Não
Nublado	72	90	Sim	Sim
Chuva	71	91	Sim	Não
Sol	72	95	Não	Não
Chuva	70	96 (3)	Não	Sim

Discretização pelo Método 1R (1-rule)

- Base de Dados Meteorológicos - Umidade

Tempo	Temperatura	Umidade	Vento	Jogar? (CLASSE)
Nublado	64	65	Sim	Sim
Chuva	65	70	Sim	Não
Sol	69	70 (1)	Não	Sim
Sol	75	70	Sim	Sim
Nublado	81	75	Não	Sim
Chuva	68			
Chuva	75			
Sol	85			
Nublado	83			
Sol	80			
Nublado	72	90	Sim	Sim
Chuva	71	91	Sim	Não
Sol	72	95	Não	Não
Chuva	70	96 (3)	Não	Sim

Categoria 1: Umidade ≤ 82.5
Categoria 2: $82.5 <$ Umidade ≤ 95.5
Categoria 3: Umidade > 95.5

Discretização de Variáveis Contínuas

- Discretização Não-Supervisionada
  - O método 1R é supervisionado
    - Considera a variável de saída (classe) na discretização
  - Métodos Não Supervisionados consideram somente o atributo a ser discretizado
  - São a única opção no caso de problemas de agrupamento (clustering), onde não se conhecem as classes de saída

Métodos de Discretização Não Supervisionada

- Três abordagens básicas:
  - Número pré-determinado de intervalos uniformes (equal-interval binning)
  - Número uniforme de amostras por intervalo (equal-frequency binning)
  - Agrupamento (clustering): intervalos arbitrários

Métodos de Discretização Não Supervisionada

- Número pré-determinado de intervalos uniformes (equal-interval binning)
  - No exemplo (temperatura):

64 65 68 69 70 71 72 72 75 75 80 81 83 85

  - Bins com largura 6:

x ≤ 60	
60 < x ≤ 66	
66 < x ≤ 72	
72 < x ≤ 78	
78 < x ≤ 84	
84 < x ≤ 90	

Métodos de Discretização Não Supervisionada

- Número pré-determinado de intervalos uniformes (equal-interval binning)
  - No exemplo (temperatura):

64 65 | 68 69 70 71 72 72 | 75 75 | 80 81 83 | 85

  - Bins com largura 6:

x ≤ 60:	n.a.
60 < x ≤ 66:	64, 65
66 < x ≤ 72:	68, 69, 70, 71, 72, 72
72 < x ≤ 78:	75, 75
78 < x ≤ 84:	80, 81, 83
84 < x ≤ 90:	85

Métodos de Discretização Não Supervisionada

- Equal-interval binning: Problemas
  - Como qualquer método não supervisionado, arrisca destruir distinções úteis, devido a divisões muito grandes ou fronteiras inadequadas
  - Distribuição das amostras muito irregular, com algumas bins com muitas amostras e outras com poucas amostras

Métodos de Discretização Não Supervisionada

- Número uniforme de amostras por intervalo (equal-frequency binning)
  - Também chamado de equalização do histograma
  - Cada bin tem o mesmo número aproximado de amostras
  - Histograma é plano
  - Heurística para o número de bins: $\sqrt{N}$
    - N = número de amostras

Métodos de Discretização Não Supervisionada

- Número uniforme de amostras por intervalo (equal-frequency binning)
  - No exemplo (temperatura):

64 65 68 69 | 70 71 72 72 | 75 75 80 | 81 83 85

  - 14 amostras: 4 Bins

x ≤ 69,5:	64, 65, 68, 69
69,5 < x ≤ 73,5:	70, 71, 72, 72
73,5 < x ≤ 80,5:	75, 75, 80
x > 80,5:	81, 83, 85

Métodos de Discretização Não Supervisionada

- Agrupamento (Clustering)
  - Pode-se aplicar um algoritmo de agrupamento no caso unidimensional
  - Para cada grupo (cluster), atribuir um valor discreto

Programa do Curso

- Limpeza/Integração de Dados
- Transformação de Dados
  - Discretização de Variáveis Contínuas
  - Transformação de Variáveis Discretas em Contínuas
  - Transformação de Variáveis Contínuas
- Análise e Seleção de Variáveis (Redução de Dados)

Transformação de Variáveis Discretas em Contínuas

- Transforma atributos categóricos em atributos contínuos
  - Problema inverso à discretização
  - Alguns algoritmos de aprendizado (KNN, FCM, regressão, etc) conseguem manusear somente atributos numéricos
  - Como estender a aplicação para atributos categóricos?

Transformação de Variáveis Discretas em Contínuas

- Caso 1: valores não ordenados
  - Alternativa 1: Redefinir a distância entre dois valores categóricos como:
    - Valores idênticos: distância = 0
    - Valores distintos: distância = 1

Transformação de Variáveis Discretas em Contínuas

- Caso 1: valores não ordenados
  - Alternativa 2: transformação de atributos
    - Atributo categórico com k valores distintos → k atributos binários
    - Utilizando-se o cálculo tradicional da distância Euclidiana, é equivalente à alternativa 1
    - Podem ser empregados pesos w_i para os k atributos, que não seriam mais {0,1}, e sim {0, w_i}
      - Reflete a importância relativa dos possíveis valores categóricos do atributo original

Transformação de Variáveis Discretas em Contínuas

- Caso 2: valores ordenados
  - Alternativa 1: atribuir um valor inteiro a cada um dos valores categóricos
    - Possível problema: implica não somente em ordenamento, mas também em uma métrica sobre os valores do atributo

Transformação de Variáveis Discretas em Contínuas

- Caso 2: valores ordenados
  - Alternativa 2: transformação de atributos
    - Atributo categórico com k valores distintos → k–1 atributos binários sintéticos, funcionando como "termômetro"
    - Exemplo: atributo categório Tamanho, com 4 possíveis valores {pequeno, médio, grande, enorme}
      - Cria-se 3 atributos numéricos binários:
        - 000: pequeno
        - 001: médio
        - 011: grande
        - 111: enorme
    - Também implica em uma métrica sobre os valores do atributo, mas não implica em igual distância

Programa do Curso

- Limpeza/Integração de Dados
- Transformação de Dados
  - Discretização de Variáveis Contínuas
  - Transformação de Variáveis Discretas em Contínuas
  - Transformação de Variáveis Contínuas
- Análise e Seleção de Variáveis (Redução de Dados)

Transformação de Variáveis Contínuas

- Se bem empregada, pode solucionar problemas nos dados ainda existentes após a Limpeza, tais como:
  - Não Normalidade
  - Não Linearidade
  - Não Homocedasticidade (atributos com variâncias muito diferentes entre si)
  - Outros problemas específicos

Transformação de Variáveis Contínuas

- Cuidados/problemas específicos:
  - Geralmente deseja-se que os dados transformados tenham distribuição normal ou quase normal
  - Escala dos dados influencia a utilidade das transformações
    - São mais efetivas com escalas arbitrárias
  - Dificuldade de interpretação dos dados transformados
  - Necessidade de pós-processamento para inverter a transformação

Transformação de Variáveis Contínuas

- Construção de atributos
- Normalização da Faixa Dinâmica
- Normalização de Distribuição
- Equalização
- Transformações de coordenadas (PCA)

Construção de atributos

- Utilização dos atributos originais para construir novos atributos derivados
- Também chamado de extração de atributos
  - Atributos sugeridos pelos atributos originais, específicos da área de aplicação, ou sugeridos por propriedades do algoritmo de aprendizado
    - Ex. 1: Dois atributos de data (nascimento, atual) podem ser subtraídos para construir o atributo idade
    - Ex. 2: Uma imagem digital pode ser pré-processada para extrair atributos descritivos, como textura, cor, etc
    - Ex. 3: Diferenciação de um atributo, quando a informação mais importante está na variação temporal

Transformação de Variáveis Contínuas

- Construção de atributos
- Normalização da Faixa Dinâmica
- Normalização de Distribuição
- Equalização
- Transformações de coordenadas (PCA)

Normalização da Faixa Dinâmica

- Freqüentemente chamado simplesmente de Normalização
- 3 Tipos mais comuns:
  - Padrão
  - Pela faixa de valores
  - Softmax

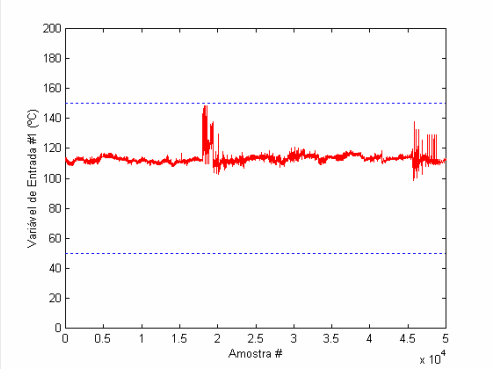
Normalização da Faixa Dinâmica

- Normalização padrão:
  - Sinal normalizado tem
    - Média zero
    - Variância unitária

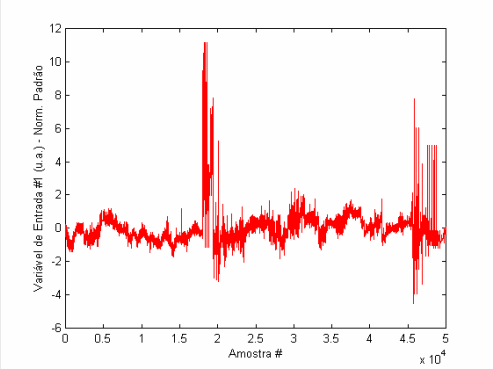
$$x_n = \frac{x - \mu}{\sigma}$$

  - Difícil implementação em tempo real:
    - Sinal normalizado não é limitado em faixa de valores
    - Necessário estimar a média e a variância dos sinais a cada instante.

Normalização - Nenhuma



Normalização Padrão



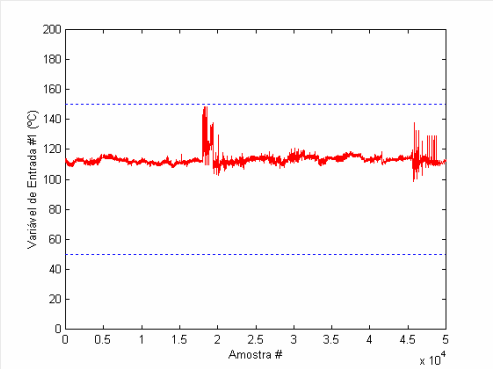
Normalização da Faixa Dinâmica

- Normalização pela Faixa de Medição:
  - O atributo tem uma faixa de valores rigidamente definida: $[m_i, M_i]$.
  - Utiliza-se essa faixa para normalização.

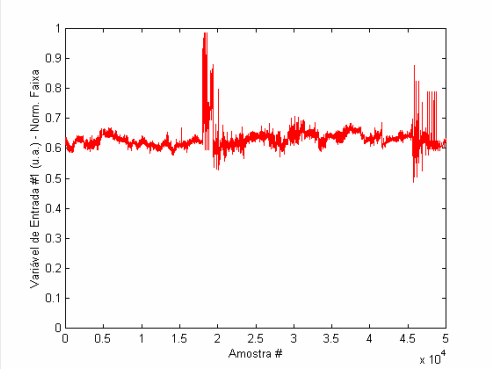
$$x_n = \frac{x - m}{M - m}$$

  - Fácil implementação, conhecendo-se a faixa
  - Pode desperdiçar faixa dinâmica
  - Solução: equalização

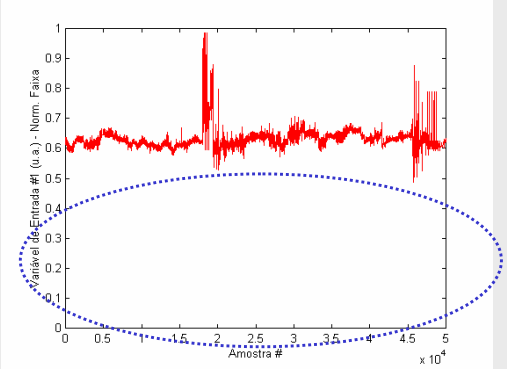
Normalização - Nenhuma



Normalização pela Faixa



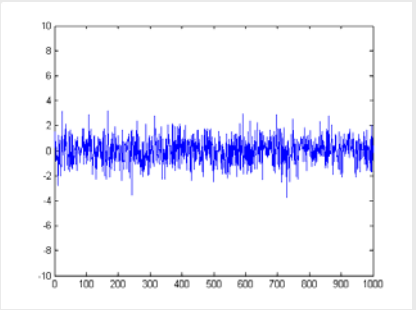
Normalização pela Faixa



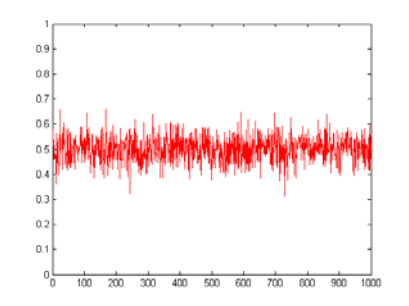
Normalização da Faixa Dinâmica

- Normalização Softmax:
  - Baseada na função logística: $x_n = \frac{1}{1+e^{-x}}$
  - Fácil implementação
  - Transforma valores de [-∞,+∞] para [0,1]
  - Lida naturalmente com valores fora da faixa esperada
  - Função tem um trecho linear
  - Também pode desperdiçar faixa dinâmica

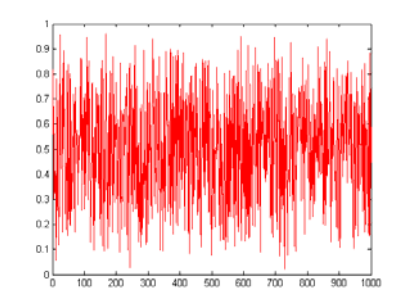
Sem Normalização



Normalização pela Faixa [-10,10]

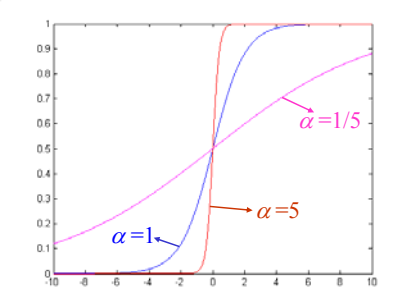


Normalização Softmax



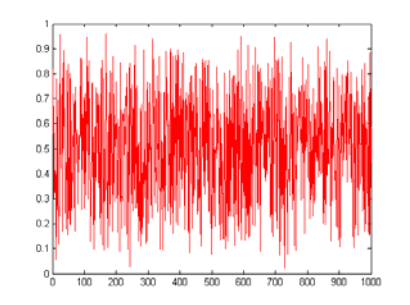
Normalização da Faixa Dinâmica

- Normalização Softmax:
  - Pode-se ajustar o tamanho do trecho linear:

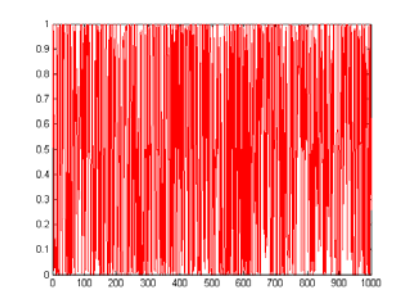


$$x_n = \frac{1}{1+e^{-\alpha x}}$$

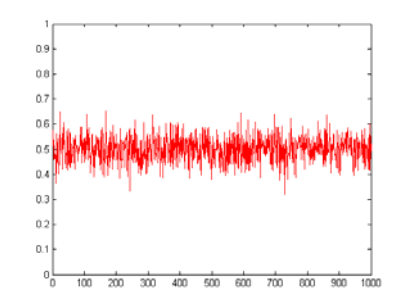
Normalização Softmax (α=1)



Normalização Softmax (α=5)



Normalização Softmax (α=1/5)



Transformação de Variáveis Contínuas

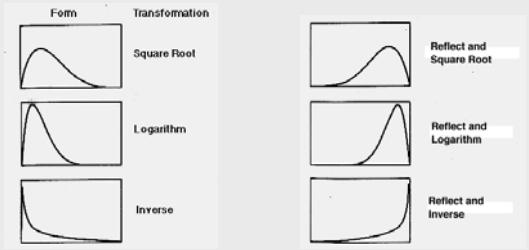
- Construção de atributos
- Normalização da Faixa Dinâmica
- Normalização de Distribuição
- Equalização
- Transformações de coordenadas (PCA)

Normalização de Distribuição

- Muitas ferramentas e algoritmos assumem que os atributos têm distribuição Normal (Gaussiana)
- Contudo, muitos atributos reais têm distribuições não Normais
- Uma transformação algébrica pode aproximar a distribuição de uma Gaussiana

Normalização de Distribuição

- Exemplos:



Normalização de Distribuição

- Transformação de Box-Cox:
  - Família de funções do tipo $x_n = \frac{x^{\lambda} - 1}{\lambda}$
  - Aplicável somente a atributos com valores estritamente positivos
  - Parâmetro λ pode ser estimado a partir dos valores do atributo original
  - Caso $\lambda = 0$, tem-se $x_n = \ln(x)$

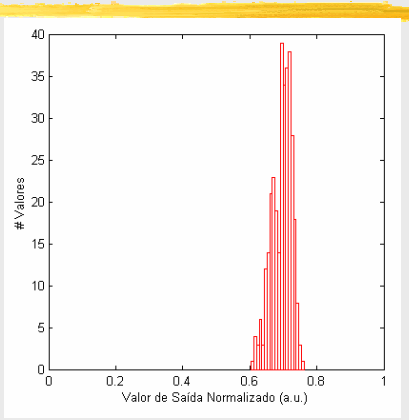
Transformação de Variáveis Contínuas

- Construção de atributos
- Normalização da Faixa Dinâmica
- Normalização de Distribuição
- Equalização
- Transformações de coordenadas (PCA)

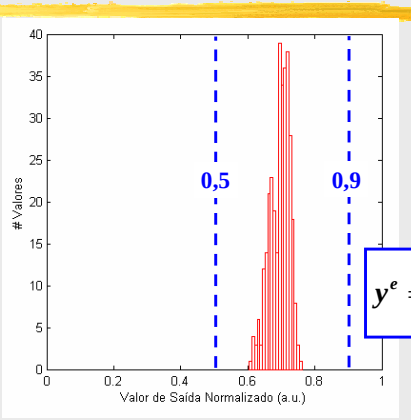
Equalização

- Procurar otimizar a faixa de variação do atributo ou da variável de saída
  - Saída da Rede Neural geralmente tem função de ativação sigmóide, limitada em [0, 1]
- Equalização com base no Histograma
- É uma forma de Normalização da Faixa Dinâmica
- Também chamada de Normalização Linear por Partes

Equalização

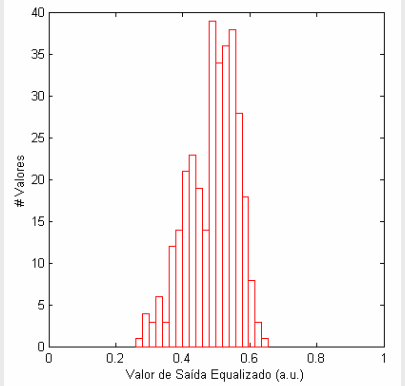


Equalização



$$y^e = \frac{y^n - 0{,}5}{0{,}9 - 0{,}5}$$

Equalização



Transformação de Variáveis Contínuas

- Construção de atributos
- Normalização da Faixa Dinâmica
- Normalização de Distribuição
- Equalização
- Transformações de coordenadas (PCA)

Transformações de coordenadas

- Base de dados com *k* atributos numéricos:
  - Nuvem de pontos no espaço k-dimensional
  - Atributos representam as coordenadas no espaço
- Contudo, os eixos que definem o sistema de coordenadas são arbitrários
- Pode-se redefinir o sistema de coordenadas para outro que seja mais adequado à aplicação pretendida

Transformações de coordenadas

- Geralmente o melhor sistema de coordenadas é definido pelos próprios dados
- Independentemente do sistema de coordenadas usado, a nuvem de pontos tem uma certa variância em cada direção
  - Indica o grau de espalhamento ao redor do valor médio nesta direção

Transformações de coordenadas

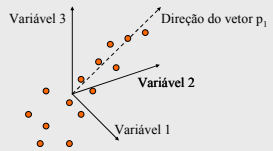
- Sistema de coordenadas Ortogonal:
  - Cada eixo forma ângulos retos com todos os demais eixos
  - Produto interno entre cada par de eixos é nulo
  - Soma das variâncias ao longo de cada eixo se mantém constante após mudança de sistema de coordenadas (desde que o novo sistema também seja ortogonal)

Análise de Componentes Principais

- Método estatístico multivariável para redução de dados.
- PCA projeta dados de um espaço ortonormal multidimensional em outro espaço, cuja base também é ortonormal.
- Analisando os dados nesta nova base, apenas poucos atributos têm importância para a descrição do fenômeno.

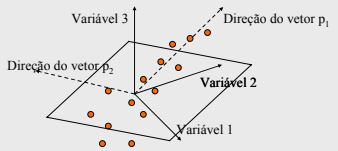
Análise de Componentes Principais

- O primeiro componente principal é determinado ajustando-se uma reta aos *n* pontos no espaço de dimensão *k* : vetor p_1, que é chamado de 1º vetor de loading



Análise de Componentes Principais

- O segundo componente principal (p_2) é determinado de maneira similar. O processo é repetido até que *k* vetores tenham sido ajustados aos dados.



Análise de Componentes Principais

- Os *k* atributos originais, contidos em uma matriz *X* [*n* x *k*], são decompostos em *k* vetores de loading *p* [*k* x 1] e *k* vetores de score *v* [*n* x 1] :

$$\boldsymbol{X} = \boldsymbol{v}_1 \boldsymbol{p}_1^T + \boldsymbol{v}_2 \boldsymbol{p}_2^T + \ldots + \boldsymbol{v}_k \boldsymbol{p}_k^T$$

- Os vetores de score formam a matriz transformada *Y* [*n* x *k*]

- Pode-se fazer uma transformação com resíduo:

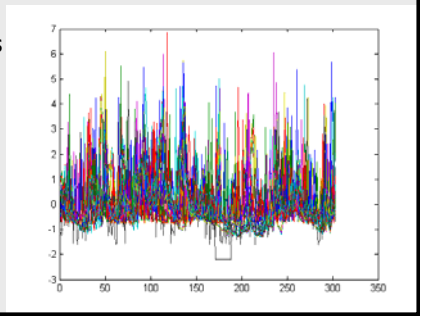
$$\boldsymbol{X} = \boldsymbol{v}_1 \boldsymbol{p}_1^T + \boldsymbol{v}_2 \boldsymbol{p}_2^T + \ldots + \boldsymbol{v}_L \boldsymbol{p}_L^T + \boldsymbol{E}, \quad L \leq k$$

Análise de Componentes Principais

- Na prática:
  - Os *k* vetores de loading *p* (componentes principais) são os autovetores da matriz de covariância de *X*
  - Os autovalores associados λ indicam o percentual da variância da matriz *X* que é explicada pelos autovetores correspondentes, em ordem decrescente
  - Com base nos valores de λ, pode-se selecionar quantos componentes principais são realmente necessários para representar a informação contida nos dados.

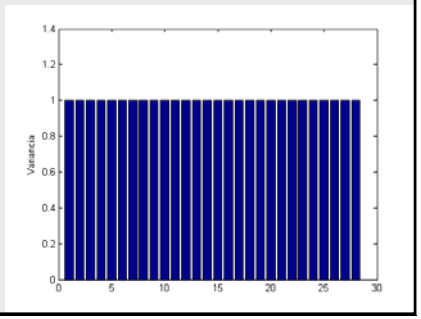
Análise de Componentes Principais

- Exemplo: matriz de dados 303 x 28
  - *k* = 28 variáveis
  - *n* = 303 amostras
  - Variâncias unitárias



Análise de Componentes Principais

- Exemplo: matriz de dados 303 x 28
  - *k* = 28 variáveis
  - *n* = 303 amostras
  - Variâncias unitárias



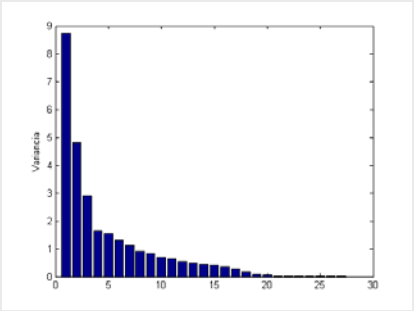
Análise de Componentes Principais

- Exemplo: matriz de dados 303 x 28
  - Matlab: [PC, SCORE, L] = princomp(X);

```
  L       28x1                   224  double array
  P       28x28                 6272  double array
  V       303x28               67872  double array
  X       303x28               67872  double array
```

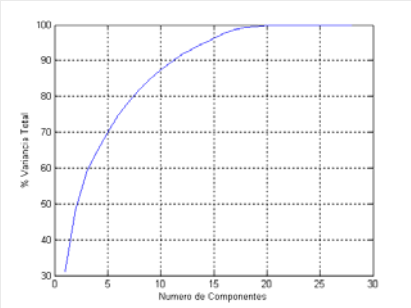
Análise de Componentes Principais

- Exemplo: matriz de dados 303 x 28



Análise de Componentes Principais

- Exemplo: matriz de dados 303 x 28



Referências

- Livro "Data Mining: Practical Machine Learning Tools and Techniques – second edition", Ian H. Witten & Eibe Frank, Elsevier
- Website http://www-stat.stanford.edu/~olshen/manuscripts/selenite/node6.html
- Website http://www.pfc.cfs.nrcan.gc.ca/profiles/wulder/mvstats/transform_e.html
- Apresentação http://www.cis.hut.fi/Opinnot/T-61.6010/s99/presentations/oct27_MK.ppt
- Apresentação http://csnet.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf