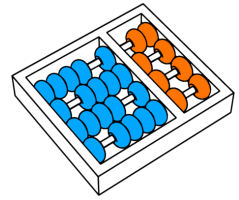


Introdução ao Processamento Digital de Imagem
MC920 / MO443

Prof. Hélio Pedrini

Instituto de Computação

UNICAMP

http://www.ic.unicamp.br/~helio

2º Semestre de 2019

Roteiro

Tratamento dos Dados

- Leitura dos arquivos
  - HTML
  - XML
  - CSV
  - JSON
  - JPG, PNG, TIFF, GIF, BMP
  - MPEG-4, AVI, MOV, WMV
  - WAV, MPEG-3, WMA, OGG, FLAC
- Dados: coleção de amostras, entidades, objetos, instâncias.
- Atributo: propriedade ou característica das amostras.
  - Exemplos: temperatura, preço, cor, área, altura.
- Valores de atributos: números ou símbolos assinalados a um atributo.

Tratamento dos Dados

Forma comum de representação de dados:

$$\mathbf{D} = \left[\begin{array}{c|cccc} & X_1 & X_2 & \ldots & X_d \\ \hline \mathbf{x}_1 & x_{11} & x_{12} & \ldots & x_{1d} \\ \mathbf{x}_2 & x_{21} & x_{12} & \ldots & x_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_n & x_{n1} & x_{n2} & \ldots & x_{nd} \end{array}\right]$$

em que $\mathbf{x}_i$ é a i-ésima linha e uma d-tupla dada por:

$$\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{id})$$

e X_j denota a j-ésima coluna e uma n-tupla dada por:

$$X_j = (x_{1j}, x_{2j}, \ldots, x_{nj})$$

Tratamento dos Dados

Dependendo do domínio da aplicação:

- linhas são chamadas de
  - entidades
  - instâncias
  - exemplos
  - registros
  - transações
- colunas são chamadas de
  - atributos
  - propriedades
  - características
  - dimensões
  - campos

Tratamento dos Dados

Exemplo: Base de dados `Iris`:

Iris setosa

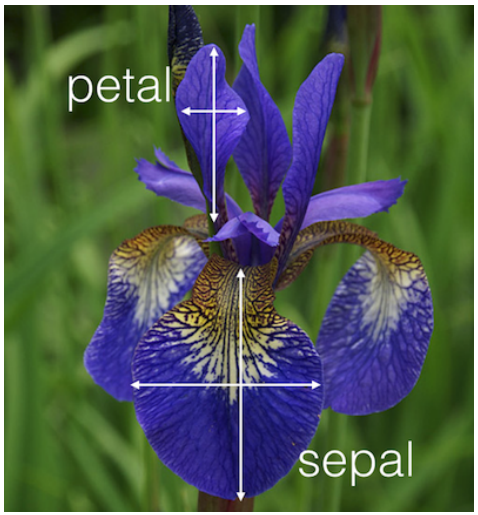
Iris versicolor

Iris virginica

- Dados multivariados apresentados pelo estatístico e biólogo inglês Ronald Fisher em 1936.

Tratamento dos Dados

- A base de dados consiste em 150 amostras, sendo 50 amostras de cada uma das três espécies da flor Íris (`Iris setosa`, `Iris virginica` e `Iris versicolor`).
- Quatro atributos foram medidos de cada amostra: comprimento e largura (em centímetros) das sépalas e pétalas.
- Fisher desenvolveu um modelo discriminante linear baseado na combinação dos atributos para distinguir as espécies.
- Dados disponíveis no *Machine Learning Repository*:
  - `https://archive.ics.uci.edu/ml/datasets/iris`



Tratamento dos Dados

Exemplos de amostras da base:

$$\mathbf{D} = \left[\begin{array}{c|cccc|c} & \textbf{Comprimento da Sépala} & \textbf{Largura da Sépala} & \textbf{Comprimento da Pétala} & \textbf{Largura da Pétala} & \textbf{Classe} \\ & X_1 & X_2 & X_3 & X_4 & Y \\ \hline \mathbf{x}_1 & 5.9 & 3.1 & 4.2 & 1.5 & \text{Iris-versicolor} \\ \mathbf{x}_2 & 6.9 & 3.9 & 4.9 & 1.5 & \text{Iris-versicolor} \\ \mathbf{x}_3 & 6.6 & 2.2 & 4.6 & 1.3 & \text{Iris-versicolor} \\ \mathbf{x}_4 & 4.6 & 3.2 & 1.4 & 0.2 & \text{Iris-setosa} \\ \mathbf{x}_5 & 6.0 & 2.2 & 4.0 & 1.0 & \text{Iris-versicolor} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{x}_{149} & 7.7 & 3.8 & 6.7 & 2.2 & \text{Iris-virginica} \\ \mathbf{x}_{150} & 5.1 & 3.4 & 1.5 & 0.2 & \text{Iris-setosa} \end{array}\right]$$

Tratamento dos Dados

Diferentes tipos de atributos:

- Nominal: valores não numéricos e não ordenados.
  - estado civil, cor dos olhos, profissão.
- Ordinal: valores não numéricos e ordenados.
  - grau de instrução, altura (alto, médio, baixo), dias da semana.
- Intervalar: valores numéricos e ordenados, com uma diferença entre esses valores.
  - datas de calendário, temperaturas em graus Celsius.
- Razão: valores numéricos, em que faz sentido calcular a proporção entre valores.
  - idade, salário, preço, volume de vendas, distâncias.

Tratamento dos Dados

Qualidade dos dados: a coleta dos dados é suscetível a anomalias, erros ou inconsistências.

- Algumas medidas que definem qualidade dos dados:
  - Confiabilidade
  - Completude
  - Consistência
  - Integridade
  - Coerência temporal
  - Relevância
  - Interpretabilidade
  - Acessibilidade

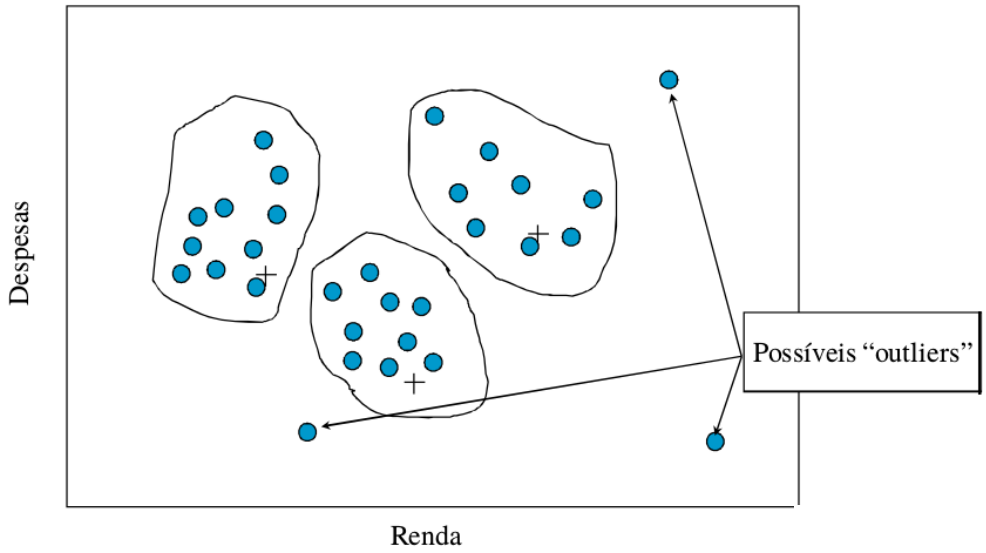
Tratamento dos Dados

- Dados faltantes:
  - ausência de observações resultantes de perda de informação.
  - indisponibilidade durante a coleta.
  - falha humana ou do dispositivo.
  - valores incompletos, atributos incompletos.
  - atributo pode não ser aplicável a todos os casos (por exemplo, salário anual não aplicável a crianças).
- Exemplo: ocupação = " "
- Potencial problema: a quantidade de dados menor do que a esperada pode comprometer a confiabilidade dos resultados da análise, eventualmente produzindo resultados tendenciosos ou específicos para uma determinada população.
- Possíveis soluções:
  - Estimar dados ou valores faltantes.
  - Ignorar valores faltantes durante a análise.
  - Preencher com medidas estatísticas (média, mediana ou moda dos demais atributos).

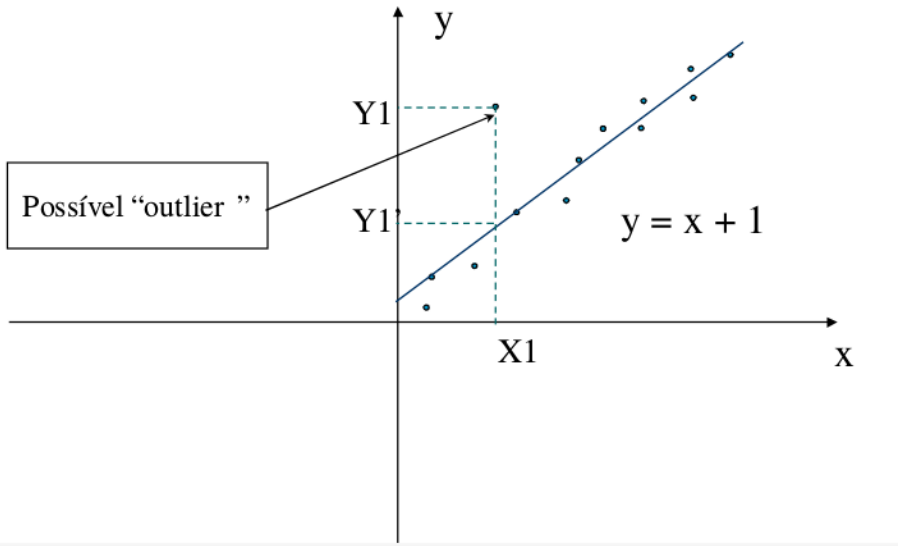
Tratamento dos Dados

- Dados discrepantes:
  - ruídos: modificação ou distorção de valores originais.
  - *outliers*: dados com atributos consideravelmente diferentes da maioria dos demais no conjunto de dados.
  - erros de resposta.
  - sensores defeituosos.
  - problemas no processo de coleta, entrada, transmissão de dados.
- Exemplo: idade = -10
- Potencial problema: valores resultantes podem ser subestimados ou superestimados.
- Possíveis soluções:
  - Excluir casos.
  - Substituir os valores errôneos.

Tratamento dos Dados



Análise de Agrupamentos



Regressão Linear

Tratamento dos Dados

- Valores repetidos:
  - dados duplicados ou que são quase duplicatas de outros.
  - podem ser gerados quando fontes heterogêneas de dados são unidas.
  - falhas de atualização da fonte de dados.
- Exemplo: mesma pessoa com múltiplos endereços de e-mail.
- Potenciais problemas: aumento do custo computacional e comprometimento dos resultados (causado ao estimar as distribuições de probabilidades) se muitas amostras duplicadas forem utilizadas no treinamento.
- Possíveis soluções:
  - Excluir duplicatas.
  - Aplicar técnicas de seleção de características ou redução de dados.

Tratamento dos Dados

Redução dos dados:

- É intuitivo pensar que, quanto maior a quantidade de objetos e atributos, mais informações estão disponíveis para o processo de análise.
- Entretanto, o aumento do número de objetos e do número de atributos (dimensão do espaço) pode fazer com que os dados disponíveis se tornem esparsos ou tornar o processamento muito complexo.
- Algumas técnicas de redução de dados são:
  - Seleção de atributos ou características.
  - Transformação ou codificação de atributos.
  - Redução do número de dados.

Tratamento dos Dados

Seleção de atributos ou características:

- Atributos irrelevantes, pouco relevantes ou redundantes são detectados e removidos.
- Atributos com valores constantes para todos os dados ou que variam pouco (variância baixa) são candidatos à remoção.
- Técnica para seleção de atributos individuais: informação mútua, que mede a dependência entre duas variáveis aleatórias. Filtragem de características com baixa informação mútua.
- Objetivos principais:
  - melhorar a acurácia do classificador.
  - tornar classificador mais rápido, especialmente no treinamento.

Tratamento dos Dados

Seleção de atributos ou características por heurísticas:

- Seleção *forward*: a busca é iniciada sem atributos e os mesmos são adicionados um a um. Cada atributo é adicionado sequencialmente e o conjunto resultante é avaliado segundo um critério. O atributo que produz o melhor critério é incorporado.
- Seleção *backward*: a busca é iniciada com o conjunto completo de atributos e os mesmos são suprimidos um de cada vez. Cada atributo é suprimido sequencialmente e o conjunto resultante é avaliado segundo um critério. O atributo que produz o melhor critério é finalmente suprimido.

Tratamento dos Dados

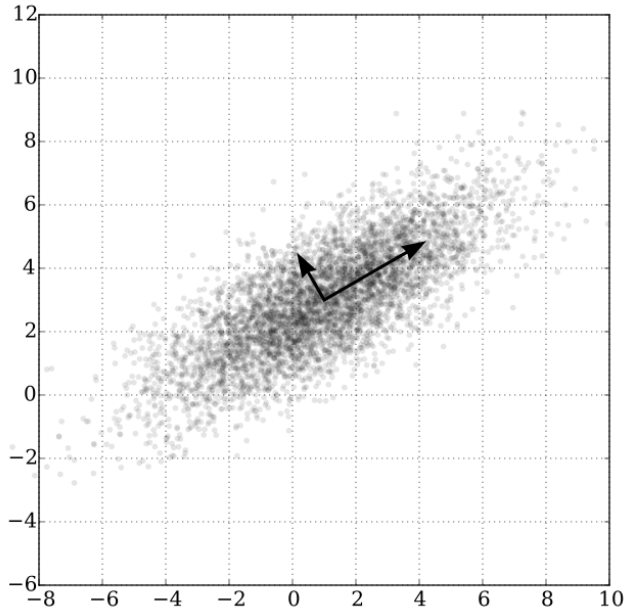
Transformação ou codificação de atributos:

- Análise de componentes principais: método estatístico que converte um conjunto de dados, com atributos possivelmente correlacionados, em um conjunto com atributos linearmente descorrelacionados, chamados de componentes principais.
- O número de componentes principais é menor ou igual ao número de atributos da base.
- A primeira componente principal possui a maior variância (maior variabilidade dos dados), a segunda componente principal possui a segunda variância e assim sucessivamente.
- Um mapeamento linear (chamado de projeção) dos dados em um espaço de dimensão menor é realizado, de forma que a variância dos dados nesse espaço seja maximizada.

Tratamento dos Dados

Transformação ou codificação de atributos:

- Baseia-se no cálculo da matriz de covariância dos dados e de seus autovetores.
- Os autovetores que correspondem aos maiores autovalores (as componentes principais) podem ser usados para reconstruir uma grande fração da variância dos dados originais.



Tratamento dos Dados

Redução do número de dados:

- Pode afetar tanto os atributos quantos os objetos, ou seja, a redução pode ser realizada tanto na quantidade de objetos quanto na dimensionalidade dos objetos da base.
- Amostragem: técnica estatística para selecionar um subconjunto de objetos de uma base que seja representativo de toda ela para efeitos de análise.
- Diferentes estratégias de amostragem: aleatória sem reposição, aleatória com reposição, por grupos, estratificada.

Tratamento dos Dados

Amostragem:

- Obter o conjunto completo dos dados de interesse é muito caro ou consome tempo demais.
- Princípio básico: amostras representativas podem funcionar tão bem quanto o conjunto completo.
- Amostras são representativas se elas têm aproximadamente as mesmas propriedades que o conjunto original de dados.

Tratamento dos Dados

Amostragem:

- Amostragem sem reposição: à medida que cada item é selecionado, ele é removido da população.
- Amostragem com reposição: objetos não são removidos da população quando são selecionados para compor a amostra (mesmo objeto pode ser escolhido mais de uma vez).
- Amostragem por grupos: divide os dados em várias partições e então seleciona aleatoriamente amostras de um subconjunto de partições.
- Amostragem estratificada: a seleção de amostras mantém a proporção de dados de cada classe.

Tratamento dos Dados

Transformação dos dados:

- Além dos problemas de inconsistências, mencionados anteriormente, que as bases de dados podem sofrer, outro tipo comum é a não uniformidade dos atributos, ou seja, alguns atributos podem ser numéricos, outros categóricos, bem como os domínios de cada atributo podem ser muito diferentes.
- Essas situações podem afetar significativamente o processo de análise de dados, tal que elas precisam ser tratadas antes de sua aplicação.
- Principais técnicas:
  - padronização.
  - normalização.

Tratamento dos Dados

Padronização: visa resolver problemas de unidades e escalas dos dados

- Capitalização: dados nominais podem aparecer em minúsculo, maiúsculo ou ambos. Para evitar inconsistências com ferramentas sensíveis a letras minúsculas ou maiúsculas, pode-se padronizar as fontes em uma das formas.
- Caracteres especiais: alguns processos são sensíveis ao conjunto de caracteres (por exemplo, problemas em decorrência de acentuação) utilizado em determinado idioma. Uma simples troca de letras em valores nominais pode evitar eventuais problemas.

Tratamento dos Dados

Padronização:

- Padronização de formatos: uso de alguns tipos de atributos (datas, número de documentos) permite diferentes formatos. Por exemplo, datas podem ser apresentadas como DDMMAAAA ou MMDDAAAA. Para evitar esses problemas, pode-se padronizar o formato de cada atributo da base, especialmente quando diferentes bases são integradas.
- Conversão de unidades: uso de diferentes unidades de medida, por exemplo, centímetros ou metros, quilômetros por hora ou milhas por hora. Como anteriormente, os dados devem ser convertidos e padronizados em uma mesma unidade de medida.

Tratamento dos Dados

Normalização:

- Procura tornar os dados mais apropriados à aplicação de alguma técnica de análise, como as redes neurais artificiais ou métodos baseados em distância.
- A necessidade de normalização pode ser consequência de vários fatores: evitar a saturação dos neurônios em uma rede neural com múltiplas camadas e fazer com que cada atributo dos dados de entrada tenha o mesmo domínio.
- Algumas técnicas: normalização Min-Max; normalização z-score; normalização por escala decimal.

Tratamento dos Dados

Normalização:

- Normalização Min-Max

$$v' = \frac{v - \min_A}{\max_A - \min_A}$$

- Normalização z-score (ou média zero)

$$v' = \frac{v - \mu_A}{\sigma_A}$$

- Normalização por escala decimal

$$v' = \frac{v}{10^j}$$

em que j é menor inteiro tal que $\max(|v'|) < 1$

Tratamento dos Dados

Exemplo de normalização Min-Max:

CPF	Despesa
99999999999	1000
11111111111	2000
22222222222	4000

$$v_1' = \frac{1000 - 1000}{4000 - 1000} = 0$$

$$v_2' = \frac{2000 - 1000}{4000 - 1000} = 0.3333$$

$$v_3' = \frac{4000 - 1000}{4000 - 1000} = 1$$

CPF	Despesa
99999999999	0
11111111111	0.3333
22222222222	1

Tratamento dos Dados

Exemplo de normalização z-score:

CPF	Despesa
99999999999	1000
11111111111	2000
22222222222	4000

$$v_1' = \frac{1000 - 2333.333}{1527.5252} = -0.8729$$

$$v_2' = \frac{2000 - 2333.333}{1527.5252} = -0.2182$$

$$v_3' = \frac{4000 - 2333.333}{1527.5252} = 1.0911$$

CPF	Despesa
99999999999	-0.8729
11111111111	-0.2182
22222222222	1.0911

Tratamento dos Dados

Exemplo de normalização por escala decimal:

CPF	Despesa
99999999999	1000
11111111111	2000
22222222222	4000

$$v_1' = \frac{1000}{10^j} = 0.1$$
$$v_2' = \frac{2000}{10^j} = 0.2$$
$$v_3' = \frac{4000}{10^j} = 0.4$$

CPF	Despesa
99999999999	0.1
11111111111	0.2
22222222222	0.4

Tratamento dos Dados

Conversão de atributos categóricos para numéricos:

- Processo de binarização do atributo categórico (*One Hot Encoding*):

...	Marca	Preço
...	Volkswagen	30.000
...	Toyota	35.000
...	Honda	37.000
...	Honda	41.000

...	Volkswagen	Toyota	Honda	Price
...	1	0	0	30.000
...	0	1	0	35.000
...	0	0	1	37.000
...	0	0	1	41.000

Tratamento dos Dados

Desbalanceamento de classes:

- Situação ocorre quando as classes não estão igualmente representadas.
- Exemplos:
  - classificação binária: em um problema de detecção de fraudes em transações de cartão de crédito, 99% das amostras pertencem à classe negativa (não fraude) e 1% das amostras pertence à classe positiva (fraude).
  - classificação multiclasse: dados de treinamento de três classes, em que 70% das amostras são da classe A, 25% da classe B e 5% da classe C.
- Potencial problema: classificador tende a ter melhor resposta para as classes majoritárias, em detrimento das minoritárias.
- Possíveis soluções: coletar mais dados (nem sempre é possível); alterar a métrica de avaliação; balancear a base de dados.

Tratamento dos Dados

Detecção de fraude:

- Classificadores terão uma tendência a dar respostas negativas para transações com fraude, ou seja, alto número de falsos negativos.
- Problema: custo de um falso negativo é maior do que o custo de um falso positivo.
  - falso negativo: fraude não foi detectada em tempo (prejuízo para a operadora de cartão).
  - falso positivo: transação normal bloqueada (aborrecimento para o usuário do cartão).

Tratamento dos Dados

Diagnóstico médico:

- Casos de pessoas com alguma doença são tipicamente menos comuns do que pessoas saudáveis.
- Em um diagnóstico médico, a classe positiva (pessoa doente) tem uma frequência muito menor do que a classe negativa (pessoas saudáveis).
- Problema: classificadores tenderão a classificar pessoas doentes como supostamente saudáveis (falsos negativos).
  - consequência: diagnóstico tardio e dano ao paciente.

Tratamento dos Dados

Abordagens para tratamento de dados desbalanceados:

- ponderação das classes: fornecer mais peso para a classe positiva nas otimizações do classificador.
  - `class_weight` na biblioteca `scikit-learn`.
- *oversampling*: replicar dados da classe minoritária; se feito aleatoriamente, pode causar *overfitting*.
  - técnica comum: SMOTE (*Synthetic Minority Over-sampling Technique*).
- *undersampling*: reduzir dados da classe majoritária; pode acarretar em perda de informação.
  - técnica comum: agrupamentos (*clustering*), em que grupos de amostras são substituídos por centroides de grupos calculados pelo algoritmo *K-means*.

Tratamento dos Dados

SMOTE (*Synthetic Minority Over-sampling Technique*):

- Selecionar apenas dados da classe minoritária.
- Para cada amostra, escolher k vizinhos mais próximos.
- Criar novas amostras entre o dado em questão e seus vizinhos por meio de interpolação.

Análise dos Dados

Tipos de problemas:

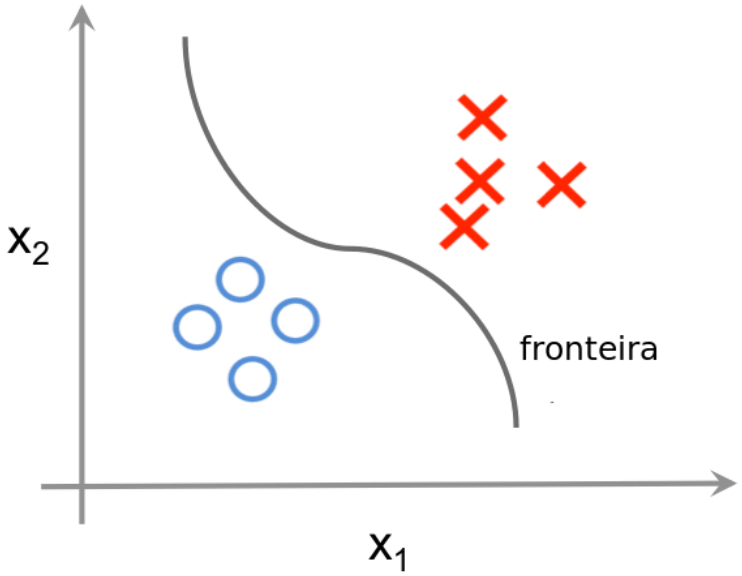
- Classificação
  - Atribuição de um objeto a uma classe.
  - Utilizada para predizer valores discretos.
    - ★ exemplo: classificar um produto como "bom" ou "ruim" em um teste de controle de qualidade.
- Regressão
  - Generalização de uma tarefa de classificação.
  - Utilizada para predizer valores contínuos.
    - ★ exemplo: predizer o valor das ações de uma empresa baseado no desempenho passado e indicadores do mercado de bolsas.
- Agrupamento
  - Organização de objetos em grupos representativos.
  - Utilizado para encontrar grupos de objetos similares.
    - ★ exemplo: organizar formas de vida em uma taxonomia de espécies.

Análise dos Dados

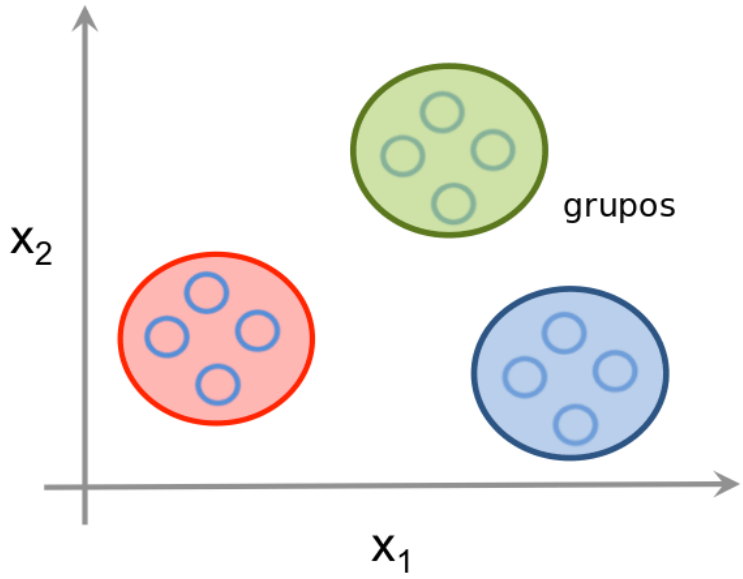
A análise de dados pode envolver modelos preditivos ou descritivos.

- Modelo Preditivo:
  - Realiza a predição acerca de valores dos dados com base em resultados conhecidos de outros dados.
  - Em geral, a modelagem é baseada em dados históricos para fazer a predição (ou previsão) sobre novos dados.
- Modelo Descritivo:
  - Identifica padrões ou relacionamentos em dados.
  - Importante para se compreender os dados.

Análise dos Dados



(a) tarefa preditiva



(b) tarefa descritiva

Análise dos Dados

Modelos Preditivos:

- Em tarefas de predição, o objetivo consiste em encontrar uma função (também denominada de modelo ou hipótese) a partir dos dados de treinamento.
- O modelo é utilizado para prever um rótulo ou valor, que caracterize um novo exemplo, com base nos valores dos seus atributos de entrada.
- Portanto, cada objeto do conjunto de treinamento deve possuir atributos de entrada e de saída.

Análise dos Dados

Modelos Preditivos:

- Estes algoritmos seguem o paradigma da aprendizagem supervisionada.
- O termo supervisionado refere-se à simulação da presença de um supervisor externo, que conhece a saída (rótulo) associada a cada exemplo (conjunto de valores para os atributos de entrada).
- Com base neste conhecimento, o supervisor externo pode avaliar a capacidade da hipótese induzida em predizer o valor de saída para novos exemplos.

Análise dos Dados

Modelos Preditivos:

- Se o atributo de saída é uma classe (valor discreto), o problema é chamado de classificação.
- Se o atributo de saída é um número (valor contínuo), o problema é chamado de regressão.

Análise dos Dados

Modelos Preditivos (Exemplos de Regressão):

- Predição do valor de venda de uma casa.

R$ 200.000,00

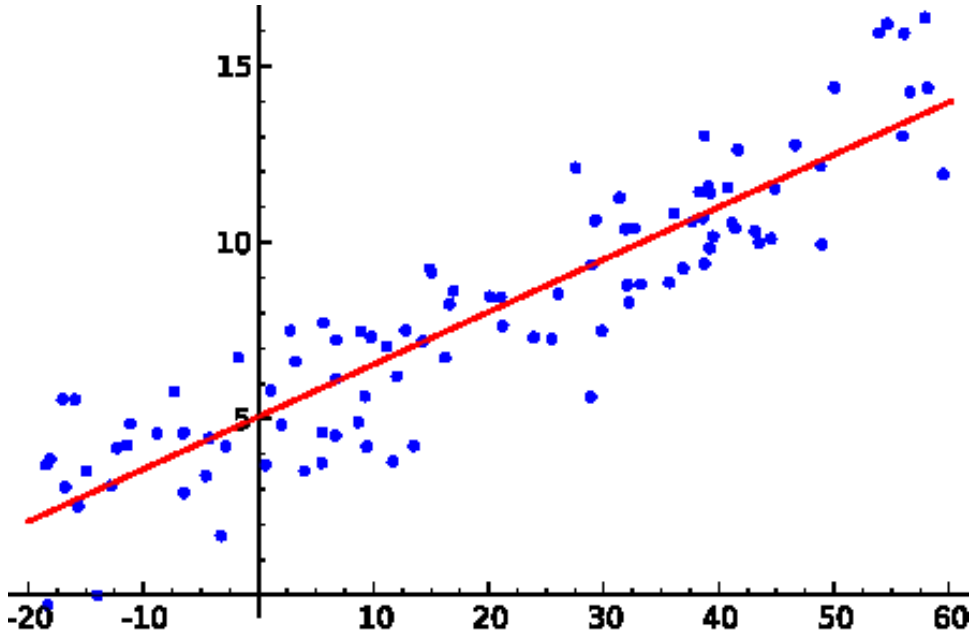
R$ 1.500.000,00

?

- Predição do preço de ações da bolsa de valores.
- Predição da taxa de desemprego em uma cidade.
- Predição do índice de criminalidade em uma região.

Análise dos Dados

Análise de regressão: processo estatístico para estimar as relações entre variáveis.
Auxilia a compreensão do relacionamento entre uma variável dependente e uma ou mais variáveis independentes.



Análise dos Dados

Regressão Linear

- Relacionamentos entre uma ou mais variáveis independentes são realizados com uma abordagem linear.
  - Simples: uma única variável independente.
  - Múltipla: diversas variáveis independentes.

Regressão Não-Linear

- Dados observados são modelados por uma função que é uma combinação não linear dos parâmetros do modelo e depende de uma ou mais variáveis independentes

Análise dos Dados

Modelos de regressão envolvem as seguintes variáveis:

- parâmetros desconhecidos: β
- variáveis independentes: $\mathbf{X}$
- variável dependente: $\mathbf{y}$

Análise dos Dados

Modelo de regressão linear simples:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \qquad i = 1, \ldots, n$$

em que ϵ_i é um erro aleatório.
Na forma matricial:

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_1 + \epsilon_1 \\ y_2 &= \beta_0 + \beta_1 x_2 + \epsilon_2 \\ \ldots &= \ldots \\ y_n &= \beta_0 + \beta_1 x_n + \epsilon_n \end{aligned} \quad \Longrightarrow \quad \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} \quad \Longrightarrow \quad \mathbf{y} = \mathbf{X}\beta + \epsilon$$

Modelo de regressão linear múltipla:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots \beta_p x_{ip} + \epsilon_i \qquad i = 1, \ldots, n$$

Análise dos Dados

Modelos Preditivos (Exemplos de Classificação):

- Classificação de transações de cartão de crédito como legítimas ou fraudulentas.
- Categorização de regiões (água, vegetação, edificação) usando dados de satélite.
- Classificação de lesões de pele como malignas ou benignas.
- Classificação de vídeos como conteúdo sensível ou não.

Análise dos Dados

Modelos Descritivos:

- Em tarefas de descrição, o objetivo consiste em explorar ou descrever um conjunto de dados.
- Os algoritmos utilizados nestas tarefas não levam em conta o atributo de saída.
- Por esse motivo, diz-se que estes algoritmos seguem o paradigma de aprendizagem não supervisionada.
- Por exemplo, uma tarefa descritiva de agrupamento de dados tem por meta encontrar grupos de objetos semelhantes no conjunto de dados.

Análise dos Dados

As tarefas que utilizam modelos descritivos são genericamente divididas em:

- Agrupamento: em que os dados são agrupados de acordo com sua semelhança.
- Regras de Associação: que consistem em encontrar padrões frequentes de associações entre os atributos de um conjunto de dados.
- Detecção de Anomalias: que consiste na identificação de padrões em dados com um comportamento diferente do esperado.
- Redução de Dimensionalidade: cujo objetivo é encontrar uma descrição simples e compacta (sumarização) de um conjunto de dados.

Análise dos Dados

Aprendizado de Máquina:

- Ciência de se programar computadores para que aprendam a partir dos dados.

Aprendizado:

- Supervisionado: dados de treinamento têm rótulos conhecidos.
- Não-Supervisionado: procura-se por padrões no conjunto de dados.

Modelos Preditivos e Descritivos:

- Conforme mencionado anteriormente, técnicas de aprendizado supervisionado utilizam modelos preditivos, enquanto técnicas de aprendizado não supervisionado utilizam modelos descritivos.

Análise dos Dados

Desafios do Aprendizado Supervisionado:

- Dados
  - quantidade insuficiente de dados de treinamento.
  - dados de treinamento não representativos.
  - dados com baixa qualidade.
- Modelos
  - muito especializados (*overfitting*).
  - muito genéricos (*underfitting*).

Análise dos Dados

Overfitting: significa que o modelo tem um bom desempenho nos dados de treinamento, mas não generaliza em dados de teste.
Possíveis causas:

- poucas amostras de treinamento.
- dados de treinamento não representativos.
- ocorrência de ruídos.

Possíveis soluções:

- simplificar o modelo: reduzir o número de parâmetros, reduzir o número de atributos nos dados de treinamento.
- obter mais dados de treinamento.
- melhorar a qualidade dos dados de treinamento (por exemplo, reduzir ruído).

Análise dos Dados

Underfitting: significa que o modelo não se ajusta aos próprios dados de treinamento.
Possíveis causas:

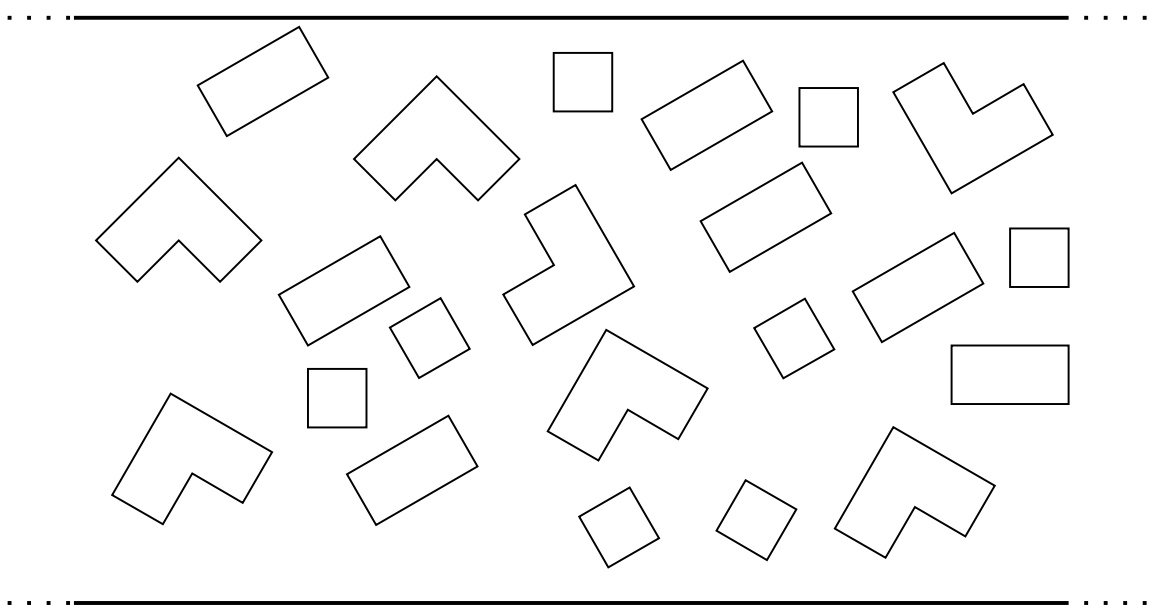
- dados não representativos ou esparsos.

Possíveis soluções:

- selecionar um modelo mais poderoso, com maior número de parâmetros.
- fornecer características mais representativas dos dados.

Análise dos Dados

Exemplo: classificação automática de objetos



Visão superior de uma esteira de rolagem.

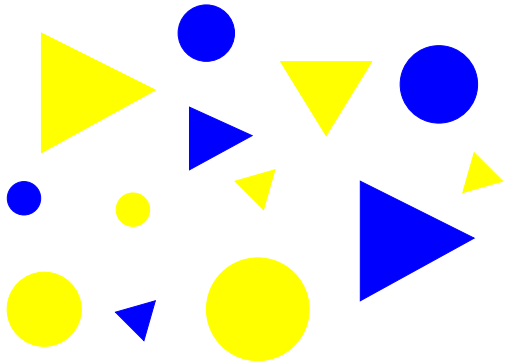
O objetivo da classificação é encontrar um mapeamento, a partir da forma geométrica de cada objeto, para o conjunto $Y = \{C_1, C_2, C_3\}$.

Análise dos Dados

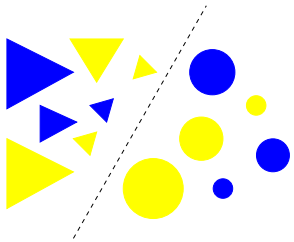
- A classificação visa determinar um mapeamento para relacionar *características* ou *propriedades* extraídas de amostras com um conjunto de rótulos.
- Amostras com características semelhantes devem ser mapeadas ao mesmo rótulo.
- Quando se atribui um mesmo rótulo a amostras distintas, diz-se que tais elementos pertencem a uma mesma *classe*, esta caracterizada por compreender elementos que compartilham propriedades em comum.
- Cada classe recebe um dentre os rótulos $C_1, C_2, \ldots, C_m$, em que m denota o número de classes de interesse em um dado problema.

Análise dos Dados

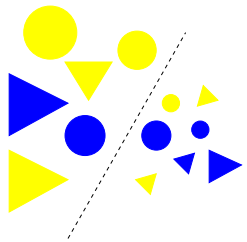
- possíveis características: forma, tamanho, cor, textura.



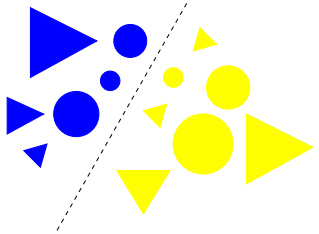
- separação em diferentes classes.



forma



tamanho



cor

Análise dos Dados

Conjunto de características:

- dependência do domínio / informação a priori.
- custo de extração.
- preferencialmente baixa dimensionalidade.
- características discriminativas:
  - valores semelhantes para padrões similares (baixa variabilidade intraclasse).
  - valores diferentes para padrões diferentes (alta variabilidade interclasse).
- características invariantes com respeito à rotação, escala, translação, oclusão, deformações.
- robustez em relação a ruído.
- baixa correlação entre características.
- normalização, padronização.

Análise dos Dados

Classificação supervisionada:

- Quando o processo de classificação considera classes previamente definidas.
- Uma etapa denominada *treinamento* deve ser executada anteriormente à aplicação do algoritmo de classificação para obtenção dos parâmetros que caracterizam cada classe.
- O conjunto formado por amostras previamente identificadas (rotuladas) chama-se *conjunto de treinamento*, no qual cada elemento apresenta dois componentes, o primeiro composto de medidas responsáveis pela descrição de suas propriedades e o segundo representando a classe a qual ele pertence.

Análise dos Dados

Há várias técnicas de aprendizado de máquina supervisionado.

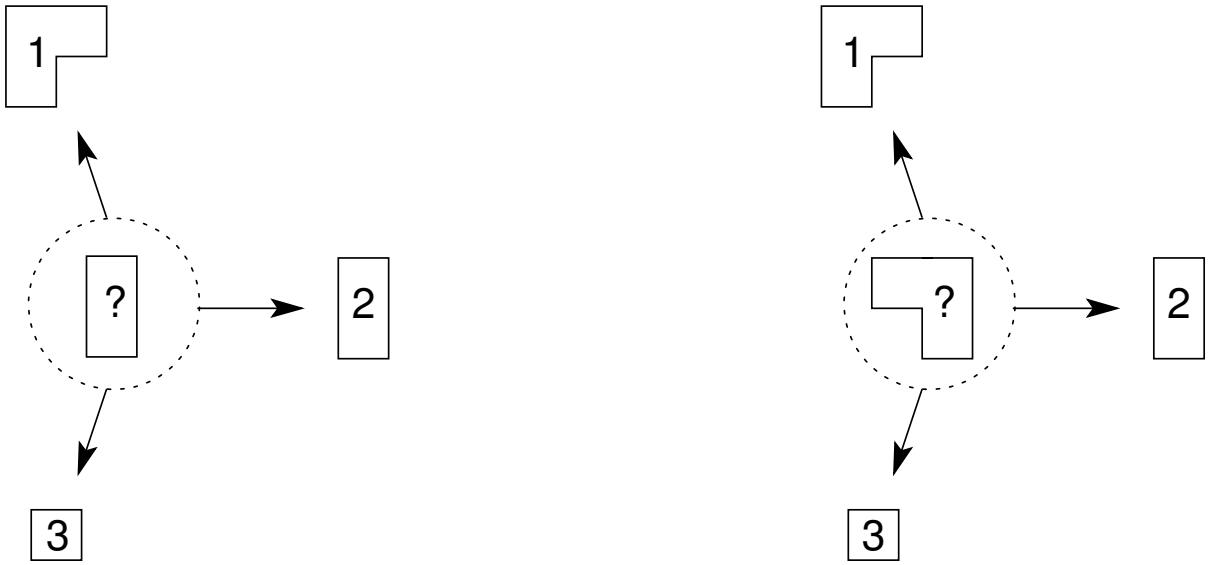
- Exemplos de abordagens:
  - Classificadores Bayesianos.
  - Classificadores pelos Vizinhos mais Próximos.
  - Máquinas de Vetores de Suporte.
  - Florestas Aleatórias.
  - Redes Neurais.

Análise dos Dados

Classificação não supervisionada:

- Quando não se dispõe de parâmetros ou informações coletadas previamente à aplicação do algoritmo de classificação.
- Todas as informações de interesse devem ser obtidas a partir das próprias amostras a serem rotuladas.
- Assim como na classificação supervisionada, amostras que compartilham propriedades semelhantes devem receber o mesmo rótulo na classificação não supervisionada.
- No entanto, diferentemente da classificação supervisionada, as classes não apresentam um significado previamente conhecido, associado aos rótulos.

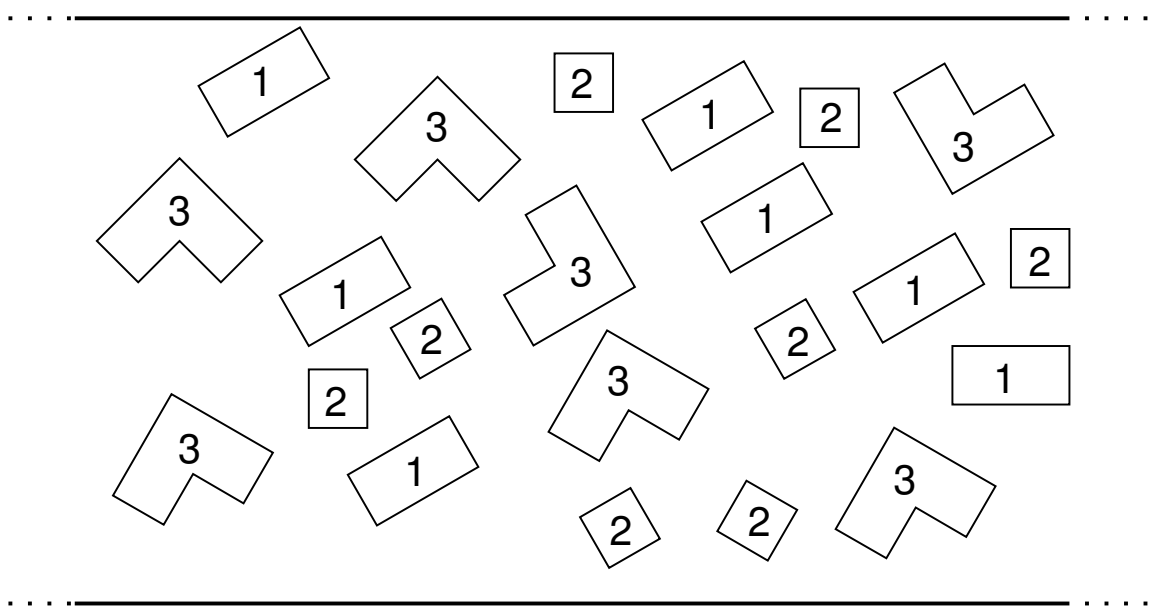
Análise dos Dados



Atribuição de amostras desconhecidas (indicadas pelas linhas tracejadas) a classes.

Análise dos Dados

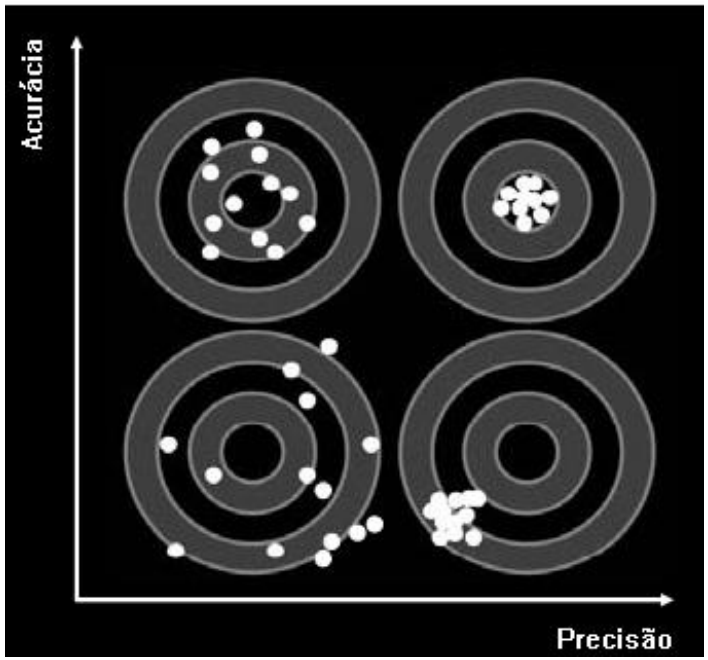
- No exemplo da classificação supervisionada, a classe C_2 é composta de objetos que apresentam formato retangular, entretanto, quando um conjunto de treinamento não se encontra disponível, pode-se afirmar apenas que os elementos que compõem a classe C_2 possuem propriedades semelhantes.



Agrupamento de amostras que possuem propriedades semelhantes.

Avaliação

Acurácia $\times$ Precisão



- Acurácia: indica o grau de concordância que há entre o resultado da medição e o valor verdadeiro da grandeza.
- Precisão: o grau de concordância entre os resultados da medição obtidos com o mesmo procedimento.

Avaliação

Avaliação de Desempenho da Regressão

- Conjunto de observações de pares $D = \{x_i, f(x_i), i = 1, \ldots n\}$, em que f não é conhecida.
- $y_i = f(x_i) \in \mathbb{R}$, em que $f(x_i)$ assume valores em um conjunto ordenado.
- O algoritmo de predição aprende uma aproximação $\hat{f}$ de f.

Avaliação

Avaliação de Desempenho da Regressão

- Erro Absoluto Médio (MAE): soma do valor absoluto das diferenças entre os valores originais e as predições:

$$\text{MAE} = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n}$$

- Exemplo:

$$y_i = [3.0, -0.5, 2.0, 7.0]$$
$$\hat{y}_i = [2.5, 0.0, 2.0, 8.0]$$
$$\text{MAE} = \frac{2}{4} = 0.5$$

Avaliação

Avaliação de Desempenho da Regressão

- Erro Quadrático Médio (MSE): soma dos quadrados das diferenças entre os valores originais e as predições:

$$\text{MSE} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}$$

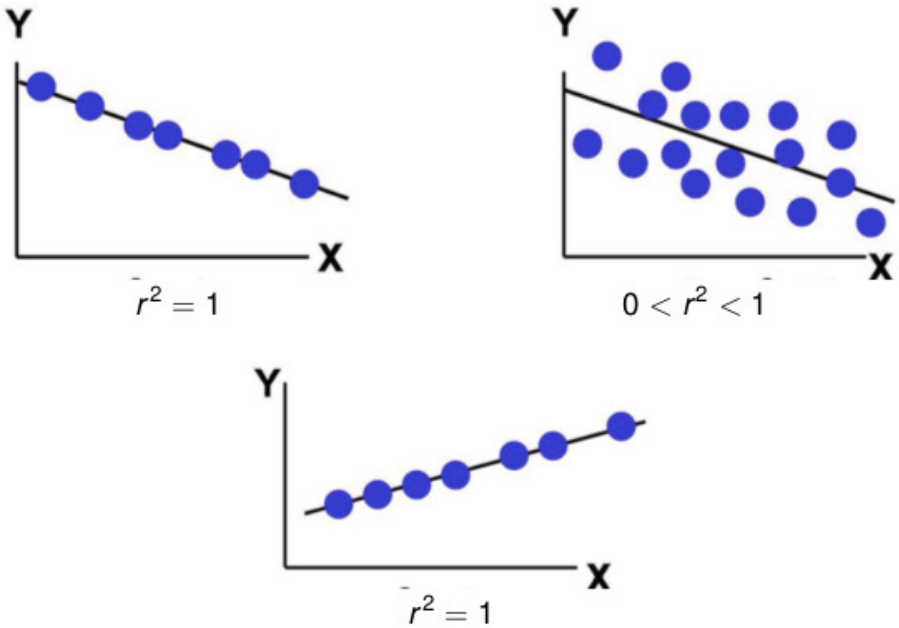
- Exemplo:

$$y_i = [3.0, -0.5, 2.0, 7.0]$$
$$\hat{y}_i = [2.5, 0.0, 2.0, 8.0]$$
$$\text{MSE} = \frac{1.5}{4} = 0.375$$

Avaliação

Avaliação de Desempenho da Regressão

- Coeficiente de Determinação R^2: medida de qualidade do ajuste da função sobre um conjunto de predições e seus valores originais.



- $0 \leq r^2 \leq 1$
- $r^2 \approx 0$: modelo pouco adequado
- $r^2 \approx 1$: modelo adequado

Avaliação

Avaliação de Desempenho da Regressão

- $R^2 = \frac{SQ_{\text{exp}}}{SQ_{\text{tot}}} = 1 - \frac{SQ_{\text{res}}}{SQ_{\text{tot}}}$

  em que

$$SQ_{\text{tot}} = \sum_{i=1}^{n} (y_i - \overline{y})^2$$

$$SQ_{\text{res}} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$SQ_{\text{exp}} = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2$$

  $\hat{y}_i$ é o valor estimado (predição) da observação y_i e $\overline{y}_i$ a média das observações.

Avaliação

Avaliação de Desempenho da Classificação

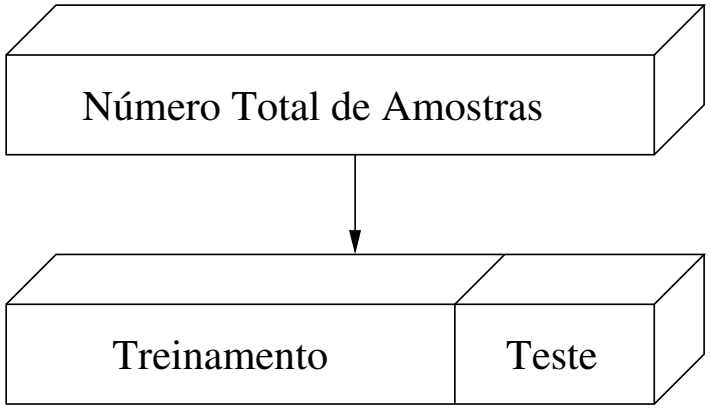
- Em aprendizado supervisionado, a avaliação pode ser realizada comparando-se os resultados preditos pelo classificador com os rótulos conhecidos das classes.
- Exemplo: rótulos conhecidos da referência $\mathbf{y}_{real}$ e predições $\hat{\mathbf{y}}_{C_1}$ e $\hat{\mathbf{y}}_{C_2}$ de dois classificadores C_1 e C_2.

$$\mathbf{y}_{real} = [0, 0, 0, 1, 1, 1, 2, 2, 2, 2, 2]$$
$$\hat{\mathbf{y}}_{C_1} = [0, 0, 2, 1, 1, 2, 1, 2, 2, 2, 2]$$
$$\hat{\mathbf{y}}_{C_2} = [0, 1, 0, 1, 2, 1, 1, 2, 0, 2, 2]$$

Avaliação

Holdout

- Este estimador divide as amostras em uma porcentagem fixa de exemplos p para treinamento e $(1 - p)$ para teste, considerando normalmente $p > 1/2$.



- Uma vez que uma hipótese construída utilizando todas as amostras, em média, apresenta desempenho melhor do que uma hipótese construída utilizando apenas uma parte das amostras, este método tem a tendência de superestimar o erro verdadeiro.

Avaliação

Holdout

- Para tornar o resultado menos dependente da forma de divisão dos exemplos, pode-se calcular a média de vários resultados de *holdout* pela construção de várias partições obtendo-se, assim, uma estimativa média do *holdout*.

Avaliação

Amostragem Aleatória

- Na amostragem aleatória, L hipóteses ($L << n$) são induzidas a partir de cada um dos L conjuntos de treinamento.
- O erro final é calculado como sendo a média dos erros de todas as hipóteses induzidas e calculados em conjuntos de teste independentes e extraídos aleatoriamente.
- Amostragem aleatória pode produzir melhores estimativas de erro que o estimador *holdout*.

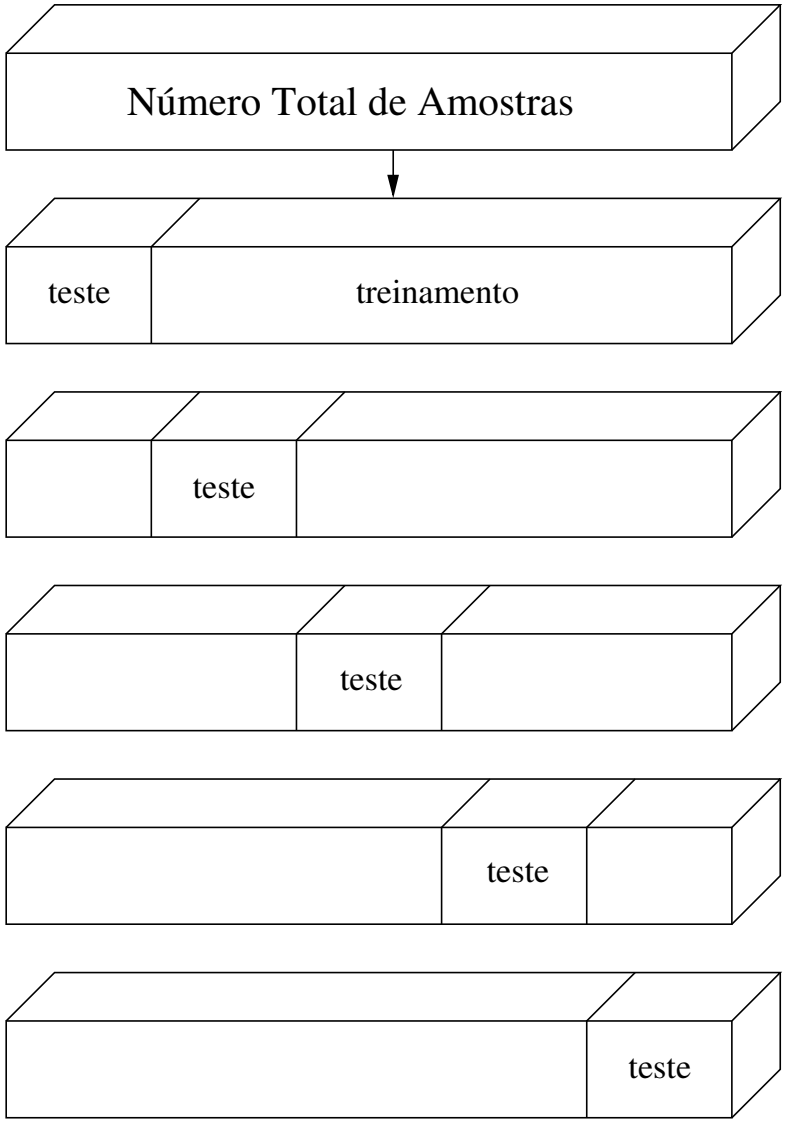
Avaliação

Validação Cruzada

- Este estimador é um meio termo entre os estimadores *holdout* e *leave-one-out*.
- Na validação cruzada com k partições, as amostras são aleatoriamente divididas em k partições mutuamente exclusivas de tamanho aproximadamente igual a n/k exemplos.
- Os exemplos nas $(k-1)$ partições são usadas para treinamento e a hipótese induzida é testada na partição remanescente.
- Este processo é repetido k vezes, cada vez considerando uma partição diferente para teste.
- O erro na validação cruzada é a média dos erros calculados em cada um das k partições.

Avaliação

Validação Cruzada



Avaliação

Validação Cruzada

- Este procedimento de rotação reduz tanto o *bias* inerente ao método de *holdout* quanto o custo computacional do método *leave-one-out*.
- Entretanto, deve-se observar, por exemplo, que na validação cruzada com 10 partições, cada par de conjuntos de treinamento compartilha 80% de exemplos.
- À medida que o número de partições aumenta, esta sobreposição pode evitar que os testes estatísticos obtenham uma boa estimativa da quantidade de variação que seria observada se cada conjunto de treinamento fosse independente dos demais.

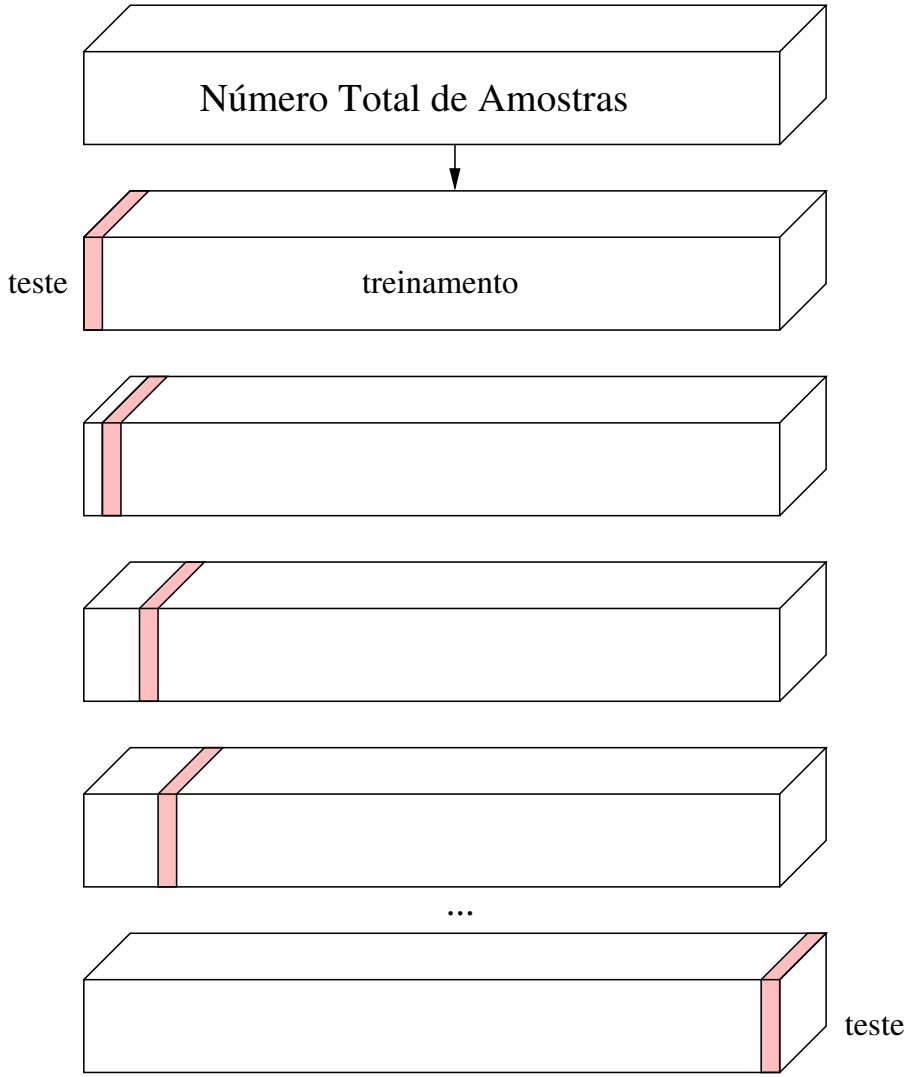
Avaliação

Leave-One-Out

- O estimador *leave-one-out* é um caso especial de validação cruzada.
- É computacionalmente dispendioso e frequentemente é usado em amostras pequenas.
- Para uma amostra de tamanho n, uma hipótese é induzida utilizando $(n-1)$ exemplos; a hipótese é então testada no único exemplo remanescente.
- Este processo é repetido n vezes, cada vez induzindo uma hipótese deixando de considerar um único exemplo.
- O erro é a soma dos erros em cada teste dividido por n.

Avaliação

Leave-One-Out



Avaliação

Bootstrap

- No estimador *bootstrap*, a ideia básica consiste em repetir o processo de classificação um grande número de vezes.
- Diferentemente da validação cruzada (que usa amostragem sem reposição), a técnica de *bootstrap* usa amostragem com reposição para formar o conjunto de treinamento
- Estimam-se então valores, tais como o erro ou *bias*, a partir dos experimentos replicados, cada experimento sendo conduzido com base em um novo conjunto de treinamento obtido por amostragem com reposição do conjunto original de amostras.

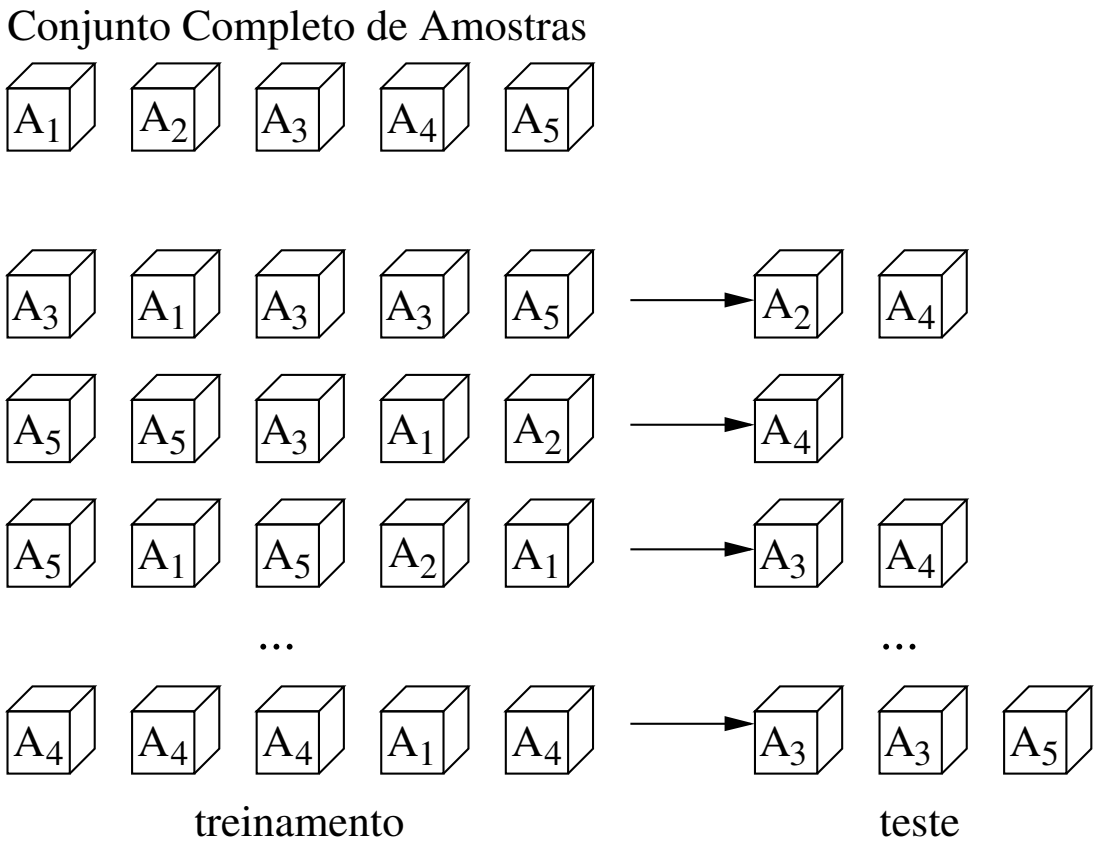
Avaliação

Bootstrap e0

- Há muitos estimadores *bootstrap*, sendo o mais comum denominado *bootstrap e0*.
- Um conjunto de treinamento *bootstrap* consiste em n amostras (mesmo tamanho do conjunto original) amostradas com reposição a partir do conjunto original de amostras.
- Isto significa que algumas amostras A_i podem não aparecer no conjunto de treinamento *bootstrap* e algumas A_i podem aparecer mais de uma vez.
- As amostras remanescentes (aquelas que não aparecem no conjunto de treinamento *bootstrap*) são usadas como conjunto de teste.

Avaliação

Bootstrap e0



Avaliação

Matriz de Confusão ou Tabela de Contingência

- A matriz de confusão apresenta o número de linhas e colunas equivalente ao número de classes do problema, em que um elemento a_{ij} indica o número de amostras atribuídas à classe C_i dado que a classe correta é a C_j. Dessa maneira, os elementos contidos na diagonal principal da matriz denotam o número de amostras classificadas corretamente.

$$\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1m} \\ a_{21} & a_{22} & \dots & a_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mm} \end{bmatrix}$$

Avaliação

Matriz de Confusão ou Tabela de Contingência

- Exemplo: resultado da classificação de 300 amostras pertencentes a seis classes diferentes.

$$\begin{bmatrix} 40 & 4 & 0 & 0 & 7 & 0 \\ 3 & 45 & 0 & 0 & 1 & 0 \\ 0 & 0 & 29 & 1 & 3 & 18 \\ 0 & 0 & 4 & 43 & 0 & 14 \\ 7 & 1 & 1 & 0 & 39 & 0 \\ 0 & 0 & 16 & 6 & 0 & 18 \end{bmatrix}$$

$$\text{taxa de erro} = 1 - \frac{214}{300} \approx 0.287$$

Avaliação

Matriz de Confusão ou Tabela de Contingência

- Estrutura construída com:
  - verdadeiros positivos (VP): valores positivos que o sistema julgou positivos (acerto).
  - falsos negativos (FN): valores positivos que o sistema julgou negativos (erro).
  - verdadeiros negativos (VN): valores negativos que o sistema julgou como negativos (acerto).
  - falsos positivos (FP): valores negativos que o sistema julgou positivos (erro).

		classe predita	
		positiva	**negativa**
classe real	**positiva**	verdadeiros positivos	falsos negativos
	negativa	falsos positivos	verdadeiros negativos

Avaliação

Medidas derivadas da tabela de contingência

- Acurácia: proporção de predições corretas, sem levar em consideração o que é positivo e o que é negativo. Esta medida é altamente suscetível a desbalanceamentos do conjunto de dados e pode facilmente induzir a uma conclusão errada sobre o desempenho do sistema.

$$\text{Acurácia} = \frac{\text{Total de Acertos}}{\text{Total de Dados no Conjunto}} = \frac{\text{VP} + \text{VN}}{\text{VP} + \text{FP} + \text{VN} + \text{FN}}$$

- Precisão: taxa com que todas as amostras classificadas como positivas são realmente positivas. Nenhuma amostra negativa é considerada.

$$\text{Precisão} = \frac{\text{Total de Acertos Positivos}}{\text{Total de Amostras Preditas como Positivas}} = \frac{\text{VP}}{\text{VP} + \text{FP}}$$

Avaliação

Medidas derivadas da tabela de contingência

- Sensibilidade: taxa com que o sistema classifica como positivas todas as amostras que são verdadeiramente positivas. Nenhuma amostra positiva é desconsiderada. Também conhecida como Revocação.

$$\text{Sensibilidade} = \frac{\text{Total de Acertos Positivos}}{\text{Total de Amostras Positivas}} = \frac{\text{VP}}{\text{VP} + \text{FN}}$$

- Especificidade: taxa com que o sistema classifica como negativas todas as amostras que são verdadeiramente negativas. Também conhecida como Seletividade.

$$\text{Especificidade} = \frac{\text{Total de Acertos Negativos}}{\text{Total de Amostras Negativas}} = \frac{\text{VN}}{\text{VN} + \text{FP}}$$

Avaliação

Medidas derivadas da tabela de contingência

- Medida F1: média harmônica entre as medidas de precisão e revocação, fornecendo um valor único para indicar o desempenho geral do modelo.

$$\text{F1} = 2 \times \frac{\text{Precisão} \times \text{Revocação}}{\text{Precisão} + \text{Revocação}} = \frac{2\text{VP}}{2\text{VP} + \text{FP} + \text{FN}}$$

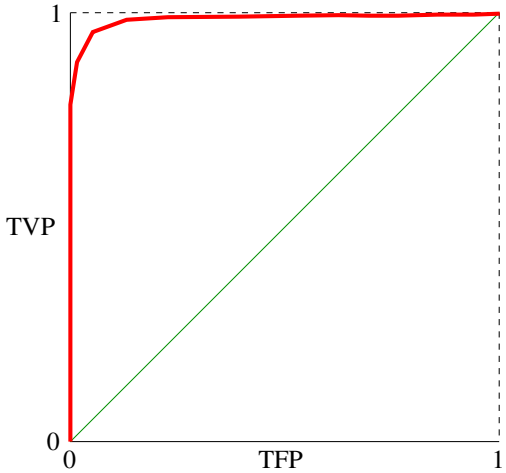
Avaliação

Curva ROC (*Receiver Operating Characteristic*)

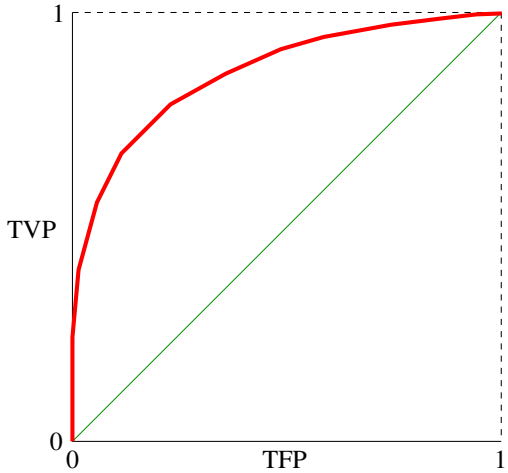
- representação gráfica que ilustra o desempenho de um classificador, mostrando como seu limiar de discriminação varia.
- diferentes valores de limiar podem ser utilizados para gerar pontos que, ligados, formam a curva.
- útil para comparar classificadores.
- obtida pela representação da fração entre os verdadeiros positivos dos positivos totais (TVP = VP / (VP + FN)) com relação à fração dos falsos positivos dos negativos totais (TFP = FP / (FP + VN)), em várias configurações do limiar.
- a taxa TVP também é conhecida como sensibilidade (revocação).
- a taxa TFP também é conhecida como 1 - especificidade.

Avaliação

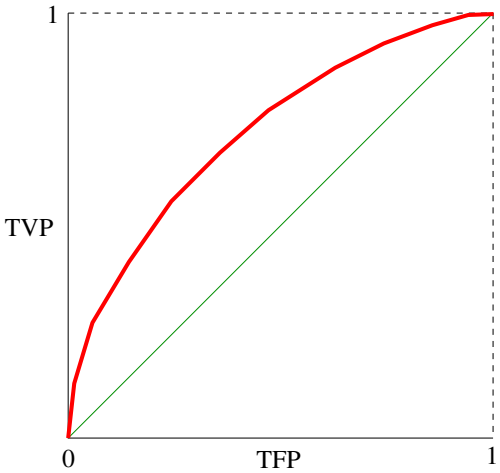
Curva ROC (*Receiver Operating Characteristic*)



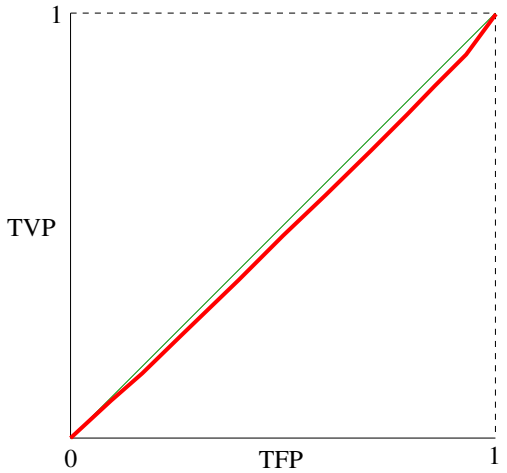
separação satisfatória



separação razoável



separação pobre

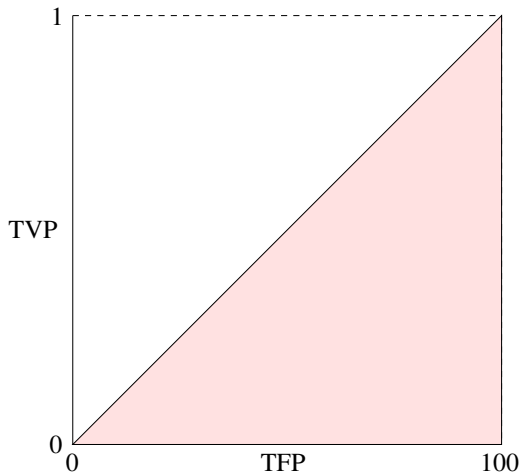


separação aleatória

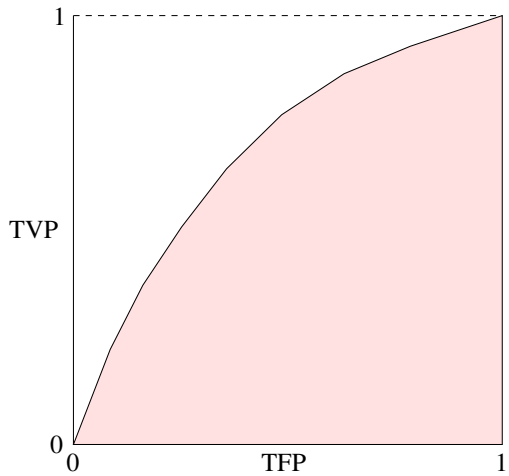
Avaliação

Área sob a Curva ROC (*Receiver Operating Characteristic*)

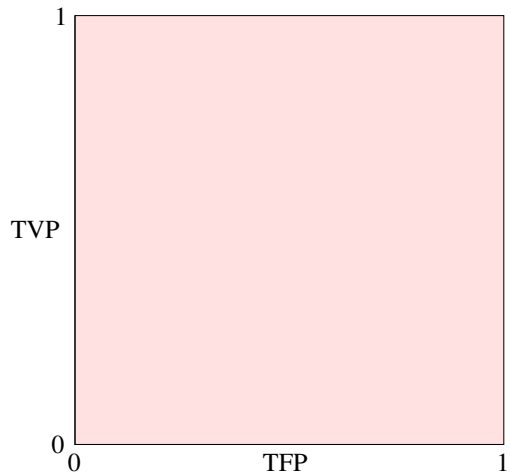
- fornece uma estimativa do desempenho de classificadores.
- gera um valor contínuo no intervalo [0, 1].
- quanto maior o valor da área, melhor.
- adição de áreas de trapezóides sucessivos.



área = 0,5



área = 0,75



área = 1,0

Avaliação

- Exemplo: resultado da classificação de 27 animais (8 gatos, 6 cachorros e 13 coelhos), cuja matriz de confusão é:

		classe predita		
		gato	**cachorro**	**coelho**
classe real	**gato**	5	3	0
	cachorro	2	3	1
	coelho	0	2	11

- Dos 8 gatos, 3 foram classificados como cachorros;
- Dos 6 cachorros, 2 foram classificados como gato e 1 como coelho;
- Dos 13 coelhos, 2 foram classificados como cachorros.

Avaliação

- Matriz de confusão para a classe gato:

5 verdadeiros positivos (gatos que foram corretamente classificados como gatos)	3 falsos negativos (gatos que foram incorretamente classificados como cachorros)
2 falsos positivos (cachorros que foram incorretamente classificados como gatos)	17 verdadeiros negativos (animais restantes, corretamente classificados como não gatos)

Avaliação

Métricas de avaliação com desbalanceamento de classes

- Não utilizar acurácia como métrica, pois falsos negativos e falsos positivos terão o mesmo custo.
- Utilizar acurácia balanceada ou medida F1:
  - $\text{Acurácia Balanceada} = \dfrac{\text{Revocação} + \text{Especificidade}}{2}$
  - $\text{F1} = 2 \times \dfrac{\text{Precisão} \times \text{Revocação}}{\text{Precisão} + \text{Revocação}}$

Avaliação

Exercício: Dada a matriz de confusão a seguir:

		classe predita	
		A	**B**
classe real	**A**	120	30
	B	40	310

Calcule as seguintes métricas:

- verdadeiros positivos
- falsos negativos
- falsos positivos
- verdadeiros negativos
- acurácia
- precisão
- revocação
- F1

Avaliação

Verdadeiros Positivos (VP) = 120

Falsos Negativos (FN) = 30

Falsos Positivos (FP) = 40

Verdadeiros Negativos (VN) = 310

Avaliação

$$\text{Acurácia} = \frac{\text{VP} + \text{VN}}{\text{VP} + \text{FP} + \text{VN} + \text{FN}} = \frac{120 + 310}{120 + 40 + 310 + 30} = 0,86$$

$$\text{Precisão} = \frac{\text{VP}}{\text{VP} + \text{FP}} = \frac{120}{120 + 40} = 0,75$$

$$\text{Revocação} = \frac{\text{VP}}{\text{VP} + \text{FN}} = \frac{120}{120 + 30} = 0,80$$

$$\text{Especificidade} = \frac{\text{VN}}{\text{VN} + \text{FP}} = \frac{310}{310 + 40} \approx 0,89$$

$$\text{F1} = 2 \times \frac{\text{Precisão} \times \text{Revocação}}{\text{Precisão} + \text{Revocação}} = 2 \times \frac{0,75 \times 0,80}{0,75 + 0,80} \approx 0,77$$

Avaliação

Exercício: Dada a matriz de confusão a seguir:

		classe predita		
		A	**B**	**C**
classe real	**A**	6	4	0
	B	3	4	2
	C	0	1	11

Calcule as seguintes métricas:

- verdadeiros positivos
- falsos negativos
- falsos positivos
- verdadeiros negativos
- acurácia
- precisão
- revocação
- F1

Avaliação

Para a classe **A**:

- VP: 6
- FN: 4
- FP: 3
- VN: 15

Para a classe **B**:

- VP: 4
- FN: 5
- FP: 5
- VN: 17

Para a classe **C**:

- VP: 11
- FN: 1
- FP: 2
- VN: 10

$$\text{Acurácia} = \frac{\text{VP} + \text{VN}}{\text{VP} + \text{FP} + \text{VN} + \text{FN}} = 0,76$$

$$\text{Precisão} = \frac{\text{VP}}{\text{VP} + \text{FP}} = \frac{21}{21 + 10} = 0,68$$

$$\text{Revocação} = \frac{\text{VP}}{\text{VP} + \text{FN}} = \frac{21}{21 + 10} = 0,68$$

$$\text{F1} = 2 \times \frac{\text{Precisão} \times \text{Revocação}}{\text{Precisão} + \text{Revocação}} = 0,68$$