



Introdução

Dados: factos, números ou texto que pode ser processado por um computador

Metadados: dados sobre os próprios dados

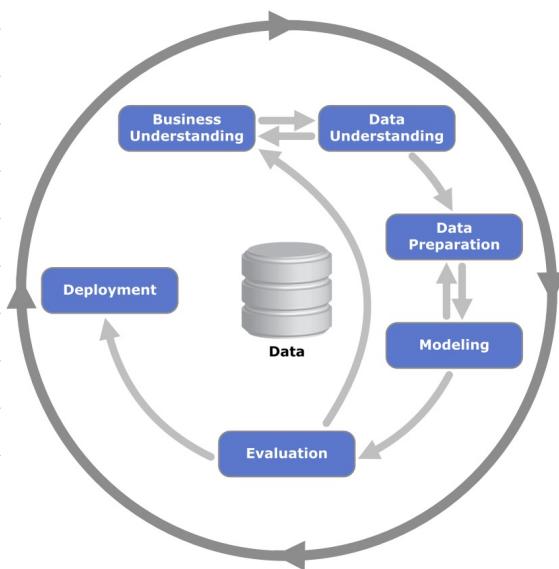
Informação: padrões, associações ou relações entre os dados

↓
Conhecimento

Como avaliar?

1. Correção - probabilidade e risco em testes
2. Generalidade - domínio e condições de validade
3. Utilidade - relevância e poder predictivo
4. Compreensibilidade - simplicidade, clareza e parcimônia
5. Novidade - anteriormente desconhecido e inesperado

Mineração de Dados: processo de descoberta de conhecimento a partir dos dados



CRISP - DM

1. Compreensão do Negócio

- a. Determinar Objetivos do Negócio
- b. Avaliar Situação
- c. Determinar Objetivos de Mineração de Dados
- d. Projetar Plano de Projeto

2. Compreensão dos Dados

- a. Colecionar Dados Iniciais
- b. Descrever Dados
- c. Explorar Dados
- d. Verificar Qualidade dos Dados

3. Preparação dos Dados

- a. Conjunto de Dados
- b. Selecionar Dados
- c. Limpar Dados
- d. Construir Dados
- e. Integrar Dados
- f. Formatar Dados

4. Modelação

- a. Selecionar Técnicas de Modelação
- b. Definir Design de Teste
- c. Construir Modelos
- d. Avaliar Modelos

5. Avaliação

- a. Avaliar Resultados
- b. Rever Processo
- c. Determinar Próximos Passos

Compreensão dos Dados

Atributo / Funcionalidade: propriedade ou característica de um objeto
Objeto / Caso: descrito por uma coleção de atributos

Dados: coleção de objetos de dados descritos por atributos

ESTRUTURA
DEPÊNDENCIA

Atributos	ESCALAS	Nominal: $= \neq$		não há relação entre os valores	
		Ordinal: $<>>$	Intervalar: $+ -$	há uma ordem entre os valores	não existe O absoluto
Categóricos	Nominal				
Numéricos	Intervalar				
	Rácio				

Attributes		Operations				
Type	Scale	$=, \neq$	$<, \leq, >, \geq$	$+, -$	\times, \div	
Numeric	Ratio	✓	✓	✓	✓	
	Interval	✓	✓	✓		
Categorical	Ordinal	✓	✓			
	Nominal	✓				

Numéricos	Discretos	conjunto finito ou infinitamente contável
	Contínuos	conjunto infinito

Dimensionalidade: número de atributos

Especificidade: contagem de presenças únicas

SUMARIZAÇÃO

Frequência: número absoluto/relativo de ocorrências de cada valor

Localização:

1. mínimo
2. máximo
3. moda
4. média — μ_x
5. quantis

Dispersão:

1. intervalo
2. desvio padrão — σ_x
3. variância — σ_x^2
4. intervalo inter-quantil — $IQR = Q_3 - Q_1$
5. covariância — como duas variáveis variam juntas?
6. correlação — como a alteração numa variável impacta outra?

Outliers: valores fora do intervalo $[Q_1 - 1,5 IQR; Q_3 + 1,5 IQR]$

VISUALIZAÇÃO

1. Gráfico Circular
2. Gráfico de Barras
3. Histograma
4. Gráfico QQ
5. Gráfico de Caixa e Bigodes

6. Gráfico de Dispersão
7. Gráfico de Conjuntos Paralelos
8. Correlogramas
9. Gráficos de Séries Temporais
10. Gráficos com Dados Agrupados

Preparação dos Dados

Ruído: modificação dos valores originais

Outliers: objetos de dados com características que não consideravelmente diferentes da maioria dos outros objetos de dados no conjunto de dados

→ RUIDO ou OBJETIVO

Valores em Falta:

1. Completamente Aleatório: o valor em falta é independente dos dados observados e não observados - não há nada de sistemático sobre ele
2. Aleatório: o valor em falta está relacionado com os dados não observados da própria variável - é informativo/não-ignorável
3. Não Aleatório: o valor em falta está relacionado com os dados observados e não com os não observados - pode haver algo de sistemático sobre ele

↓ Soluções:

1. remover observações com valores em falta - considerar só casos completos
2. ignorar valores em falta
3. fazer estimativas para preencher os valores em falta - imputação

→ Duplicados

→ Dados Inconsistentes/Incôertos

PRÉ-PROCESSAMENTO DE DADOS

- Necesidade de "criar" novas variáveis
- Necesidade de seleccionar subconjuntos representativos dos dados

Extracção de Funcionalidades: para sensores, imagens, logs, tráfego e documentos

Limppeza dos Dados: para arrumar o conjunto de dados
↓ CONJUNTO DE DADOS

- Cada valor pertence a uma variável e uma observação
- Cada variável contém todos os valores de uma certa propriedade medida entre todas as observações
- Cada observação contém todos os valores das variáveis medidas para o caso respetivo

↓ TABELA

- Cada linha representa uma observação
- Cada coluna representa um atributo medido para cada observação

Como lidar com valores em falta?

1. Remover todos os casos com um valor desconhecido
2. Preencher os desconhecidos com a imputação do valor mais comum
3. Preencher com o valor mais comum nos casos que não são semelhantes àquele com desconhecidos
4. Preencher com a interpolação linear los valores próximos em tempo/espaço
5. Explorar eventuais correlações entre variáveis
6. Não fazer nada

Como lidar com valores incorretos?

1. Deteção de Inconsistências: técnicas de integração de dados
2. Conhecimento do Domínio: auditoria dos dados que usa conhecimento/restricções
3. Métodos Centrados nos Dados: métodos estatísticos para detectar outliers

Transformação dos Dados: mapear um conjunto interno de valores de um atributo para um novo conjunto de valores tal que cada valor antigo pode ser identificado com um dos valores novos

1. Normalização:

- a. Escala Min-Max: $y_i = \frac{x_i - \min_x}{\max_x - \min_x}$, $y_i \in [0, 1]$ - não robusto quando há outliers
- b. Padronização: $y_i = \frac{x_i - \mu_x}{\sigma_x}$ - numa distribuição normal, $y_i \in [-3, 3]$, $\mu_x = 0$, $\sigma_x = 1$

2. Numérico → Categórico:

- a. Binarização: se o atributo só tem 2 valores nominais possíveis, então pode ser transformado num atributo binário
- b. One-Hot Encoding: se o atributo tem K valores nominais possíveis, então pode ser transformado em K atributos binários

3. Discretização: converter um atributo contínuo num atributo ordinal de variáveis numéricas

- a. Não Supervisionada: encontrar quebras nos valores dos dados
 - i. Igual Largura: pode ser afetada pela presença de outliers
 - ii. Igual Freqüência: mesmo número de valores em cada intervalo
- b. Supervisionada: usar etiquetas das classes para encontrar quebras

Engenharia de Funcionalidades: processo de utilizar conhecimento do domínio dos dados para criar funcionalidades que podem ajudar a resolver o problema

1. Expressar relações conhecidas entre variáveis existentes
2. Expressar dependências de casos conhecidos
 - a. representar valores relativos em vez de valores absolutos $y_t = \frac{x_t - x_{t-1}}{x_{t-1}}$
 - b. suas variáveis cujos valores são o valor da mesma variável em tempos anteriores

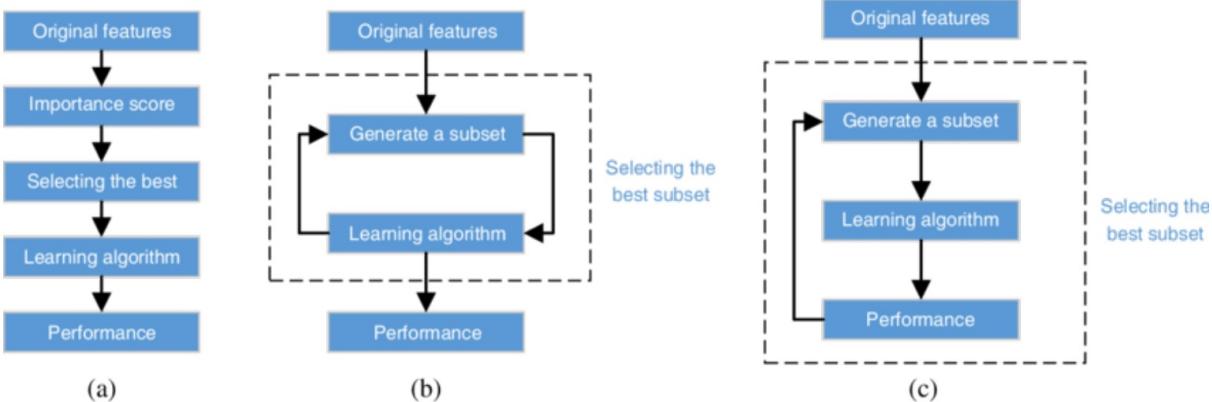
Amostragem dos Dados: amostrar os casos do conjunto de dados original para obter um conjunto de dados muito menor — usar uma amostra vai funcionar quase tão bem como usar o conjunto de dados inteiro se a amostra for representativa, isto é, se tiver aproximadamente as mesmas propriedades (de interesse) do conjunto de dados original

1. Aleatória: há igual probabilidade de selecionar qualquer item particular
 - a. Sem Reposição: cada objeto selecionado é removido da população
 - b. com Reposição: objetos selecionados não são removidos da população
2. Estatística: dividir os dados em várias partícipes e sortear amostras aleatórias de cada partição
3. Incremental: começar com uma amostra pequena e incrementar o seu tamanho até não existir ganho no desempenho do modelo

Redução da Dimensionalidade: para evitar a maldição da dimensionalidade () reduzir o tempo e a memória, facilitar a visualização e reduzir ruído/irrelevância

1. Seleção de Funcionalidades
 - a. Descartar Funcionalidades Irrelevantes: contêm informação inútil para a tarefa
 - b. Descartar Funcionalidades Redundantes: duplicam informação contida outros atributos
2. Análise do Componente Principal: encontrar uma projeção num novo conjunto de eixos que captura a maior quantidade de variabilidade nos dados — encontrar m-númeras combinações lineares que melhor capturam a variabilidade nos dados

MÉTODOS DE SELEÇÃO DE FUNCIONALIDADES



- a. Filtagem: seleciona funcionalidades independentemente da tarefa de mineração
remove funcionalidades com baixa variação, alta correlação e ordena por medida de relevância
- b. Embrulho: seleciona funcionalidades tomando em consideração a tarefa de mineração
procura o subconjunto ótimo de funcionalidades
- i. Iterativo:
 - I. Seleção para a Frente: seleciona um atributo, adiciona, repete
 - II. Eliminação para Trás: seleciona um atributo, remove, repete
 - ii. Recurssivo: remove recursivamente os atributos do conjunto atual
- c. Embedido: a seleção está construída no algoritmo que produz o modelo para a tarefa de mineração de dados

Modelação Descritiva

Objetivo: descrever/visualizar ou encontrar estrutura no que foi observado

Semelhança: medida numérica de quão semelhantes dois objetos são - maior $[0, 1]$ quando os objetos são mais parecidos

Diferença: medida numérica de quão diferentes dois objetos são - menor quando $[0, +\infty[$ os objetos são mais parecidos

Desigualdade Triangular:

1. $d(x_i, x_j) \geq 0$
2. $d(x_i, x_j) = 0 \Leftrightarrow x_i = x_j$
3. $d(x_i, x_j) = d(x_j, x_i)$
4. $d(x_i, x_j) \leq d(x_i, x_k) + d(x_k, x_j)$

Distância Euclidiana: $d(x_i, x_j) = \sqrt{\sum_{a=1}^m (x_i^a - x_j^a)^2}$ - $\mu = 2$

Distância de Manhattan: $d(x_i, x_j) = \sum_{a=1}^m |x_i^a - x_j^a|$ - $\mu = 1$

Distância de Minkowski: $d(x_i, x_j) = \sqrt[\mu]{\sum_{a=1}^m |x_i^a - x_j^a|^\mu}$

Distância de Chebyschev: diferença máxima entre qualquer atributo dos dois pontos - $\mu = \infty$

Distância Heterogênea: $d(x_i, x_j) = \sum_{a=1}^m S_a(x_i^a, x_j^a)$

a categórica? $S_a(x_i^a, x_j^a) = \begin{cases} 0, & \text{se } x_i^a = x_j^a \\ 1, & \text{caso contrário} \end{cases}$

a numérica? $S_a(x_i^a, x_j^a) = \frac{|x_i^a - x_j^a|}{\text{max}_a - \text{min}_a}$

$$\text{Coeficiente Geral de Semelhança: } \delta(x_i, x_j) = \frac{\sum_{a=1}^m w_a \delta(x_i^a, x_j^a)}{\sum_{a=1}^m w_a}$$

$\delta(\cdot)$: medida de semelhança - $\forall x_i, x_j : \delta(x_i, x_j) = 1 \Leftrightarrow x_i = x_j \wedge \delta(x_i, x_j) = \delta(x_j, x_i)$

m: número de atributos

x_i^a : valor do atributo índice-a para x_i

w_a : peso para o atributo a

A GRUPAMENTO

Objetivo: obter o agrupamento "natural" dos dados - encontrar alguma estrutura no conjunto de dados

Agrupamento Particional: divide as observações em K partisões de acordo com algum critério

Agrupamento Hierárquico: gera uma hierarquia de grupos, de 1 a N grupos, sendo N o número de linhas do conjunto de dados

A GRUPAMENTO PARTICIONAL

Objetivo: particionar o conjunto de dados em K grupos ao minimizar ou maximizar critérios pré-especificados

Objetivo: minimizar distância intra-grupos e maximizar distância inter-grupos

Compactade do Grupo: quão semelhantes são os casos dentro do mesmo grupo

Separação do Grupo: quão longe está o grupo dos outros grupos

Agrupamento "Hard": um objeto pertence a um só grupo

Agrupamento "Fuzzy": cada objeto tem uma probabilidade associada para pertencer a cada grupo

Centroide: $\bar{x}^{(k)} = \frac{1}{n_k} \sum_{x_i \in C_k} x_i$ ou $\tilde{x}^{(k)}$, $C_k = \{x_1, \dots, x_{n_k}\}$

Objetivo: obter o conjunto de grupos C que minimiza $J(C) = \sum_{j=1}^k \sum_{x_i \in C_j} d(x_i, \bar{x}^{(j)})$

Algoritmo de K-Médias:

1. Inicializar os centros dos K grupos para um conjunto de observações aleatoriamente escolhidas

2. Repeti:

a. Alocar cada observação ao grupo cujo centro está mais próximo

b. Recalcular o centro de cada grupo

3. Até os grupos serem estáveis — não existe decrescimento significativo ou existe um crescimento no critério de minimização $J(C)$

• O algoritmo maximiza a diferença entre grupos, é rápido e escala bem

• O algoritmo não é ótimo, pode obter soluções diferentes e K tem de ser definido

VALIDAÇÃO DOS AGRUPAMENTOS

Avaliação Supervisionada: compara o agrupamento obtido com a informação externa disponível

Avaliação Não Supervisionada: tenta medir a qualidade do agrupamento sem nenhuma informação sobre a estrutura ideal dos dados

1. Coeficientes de Coesão: determinam quão compactos/coesos são os membros do grupo
2. Coeficientes de Separação: determinam quão diferentes são os membros de diferentes grupos

Coeficiente de Silhueta: para cada objeto x_i

1. Obtém a distância média a todos os objetos no mesmo grupo $-a_i$
 2. Para qualquer outro grupo ao qual x_i não pertence, calcular a distância média aos membros desse outro grupo e obtém o valor mínimo dessas distâncias $-b_i$
- $$\rightarrow S_i = \frac{b_i - a_i}{\max(a_i, b_i)} \quad \rightarrow S_i \in [-1, 1]$$

$S_i \approx 1$: dados bem agrupados

$S_i \approx 0$: dados residem entre dois grupos

$S_i \approx -1$: dados provavelmente no grupo errado

Como escolher K?

$$1. K = \sqrt{n}/2$$

2. Método da Silhueta: calcular o coeficiente médio de silhueta para cada K e escolher o melhor valor

3. Método do Cotovelo: calcular a distorção (soma dos quadrados) entre grupos e escolher o K tal que adicionar outro grupo não retorna uma distorção muito menor

Outros Métodos de Agrupamento Particional:

1. PAM: particionamento em torno de centroides - mais robusto contra "outliers"
2. CLARA: amostra aleatoriamente, aplica PAM para obter K centroides, aloca as observações a um deles, calcula a soma das diferenças e retorna a menor

• Os algoritmos semelhantes a K-médias têm problemas com grupos de tamanhos ou densidades diferentes, forma não globular e dados com outliers/ruido

DBSCAN: a densidade de uma observação é estimada pelo número de observações dentro de um certo raio — DBSCAN (E , MinPoints)

Pontos Nucleares: se o número de observações dentro do seu raio é maior do que E

Pontos de Fronteira: se o número de observações dentro do seu raio não alcança E

mas estão dentro do raio de um ponto nuclear

Pontos de Ruído: não têm observações suficientes dentro do seu raio nem estão suficientemente perto de nenhum ponto nuclear

1. Classifica cada observação em uma das três possíveis alternativas
2. Elimina os pontos de ruído da formação dos grupos
3. Todos os pontos nucleares dentro de uma certa distância são alocados no mesmo grupo
4. Cada ponto de fronteira é alocado ao grupo do ponto nuclear mais próximo

• lida com grupos de diferentes formas e tamanhos e resiste a ruído
• densidades variadas e dados multi-dimensionais

AGRUPAMENTO HIERÁRQUICO

Objetivo: obter uma hierarquia de grupos em que cada nível representa uma solução com K grupos

Aglomerativo: "bottom-up" — começa com tantos grupos quantos os casos e em cada nível superior une o par de grupos mais semelhantes num só grupo

Divisivo: "top-down" — começa com um único grupo e em cada nível escolhe o grupo com menor uniformidade para ser dividido em dois

Medidas de Proximidade:

- 1. Single: $\min(d)$
- 2. Complete: $\max(d)$
- 3. Average: \bar{d}

↳ lida com não-elipses
↳ critério local
↳ não considera outliers

↳ enviesado para globulos
↳ critério não-local
↳ não responde a ruído/outliers

Modelação Preditiva

Machine Learning: estudo sistemático de algoritmos e sistemas que melhoram o seu conhecimento ou desempenho com experiência

Objetivo: construir modelos que capturam o conhecimento a partir de casos observados para fazer inferências em casos não observados

Aprendizagem Não Supervisionada: nenhuma etiqueta/valor alvo é associado a cada exemplo — o objetivo é obter uma descrição do conjunto de dados

Aprendizagem Supervisionada: há uma etiqueta/valor alvo associada a cada exemplo — o objetivo é aprender uma função (modelo) que mapeie cada exemplo com a sua variável alvo

Aprendizagem Reforçada: o algoritmo de aprendizagem constrói exemplos a partir de uma regra de regras, depois, um processo iterativo é usado para melhorar (reforçar) o conjunto de exemplos até um critério de avaliação ser suficientemente bom

MODELAÇÃO PREDITIVA

Modelos Preditivos: obtidos na base da assumção de que existe um mecanismo de conhecido que mapeia as características das observações em conclusões

Objetivo: descobrir o mecanismo — obter uma aproximação da função que mapeia os descriptores na variável-alvo

Descritores/Preditores/Variáveis Independentes: conjunto de variáveis que descreve as propriedades (funcionalidades, atributos, preditores) do conjunto de dados.

Alvo/Variável Dependente: o que se quer prover/concluir em relação às observações

Dado um conjunto de variáveis preditoras X e uma variável-alvo Y , existe uma função f tal que $f(X) = Y$ — como f é desconhecida, o objetivo é aprender a melhor aproximação para f , \hat{f} , tal que as variáveis-alvo podem ser obtidas a partir do conjunto de dados de input e, com \hat{f} , fazer previsões

VIES

Underfitting: o modelo é demasiado simples para capturar padrões nos dados
Overfitting: o modelo desempenha-se muito bem nos dados de treino, mas não generaliza bem para dados não vistos
VARIÂNCIA

Usos:

1. Previsão: fazer previsões em relação à variável-alvo de novos casos
2. Compreensibilidade: compreender quais os fatores que influenciam as conclusões

Problema de Classificação: a variável-alvo Y é nominal

Problema de Regressão: a variável-alvo Y é numérica

Os modelos preditivos assumem uma forma funcional para a função desconhecida
Um criterio de preferência permite comparar as diferentes variantes de modelos

Abordagens:

1. Baseada em Distância
2. Probabilística
3. Fórmula Matemática
4. Lógica
5. Otimização
6. Conjuntos de Modelos

CLASSIFICAÇÃO

$$D = \{(x_i, y_i)\}_{i=1}^N$$

- $x_i = (x_{i1}, \dots, x_{ip})$ - valor do vetor de funcionalidades
- $y_i \in Y$ - valor da variável nominal Y

Objetivo: aprender a melhor aproximação da função desconhecida $Y = f(x)$

Abordagem:

1. Assumei uma forma funcional $h_0(x)$ para $f()$, sendo Θ um conjunto de ^{parâmetros} V
2. Assumei um critério de preferência sobre Θ das parametrizações possíveis de Θ
3. Procurar pelo melhor $h()$ de acordo com o critério e o conjunto de dados

Problema de Classificação Binário: a variável-alvo só tem dois valores possíveis ✓
Problema de Classificação Multi-Classe: a variável-alvo tem mais de dois valores ✓

Um-VS-Todos: treinar um modelo para cada classe - K classes $\leftrightarrow K$ classificadores

Um-VS-Um: treinar um modelo para cada par de classes - K classes $\leftrightarrow K(K-1)/2$ classif.

Rácio de Erros: proporção de previsões incorretas - $L_{0/1} = \frac{1}{N} \sum_{i=1}^N I(\hat{y}_i, y_i)$

Exatidão: $1 - L_{0/1}$

Matriz de Confusão: classe real VS. classe prevista

Exatidão: proporção de previsões corretas - $(TP + TN) / (TP + FP + TN + FN)$

Precisão: proporção de previsões positivas corretas - $TP / (TP + FP)$

Lembrança: proporção de exemplos positivos capturados $TP / (TP + FN)$ ↗ TRADEOFF

$F_\beta = \frac{(\beta^2 + 1)}{\beta^2} \times \text{Precisão} \times \text{Leitura}$ $\beta \rightarrow 0$: diminui peso da leitura
 $\beta \rightarrow \infty$: diminui peso da precisão

Rácio de Falsos Positivos: $FPR = FP / (TN + FP)$

Curva ROC: trade-off entre TPR e FPR conforme varia o threshold de discriminação

↳ $0 \leq AUC \leq 1$ - quanto maior AUC melhor distingue

Fraud Detection (CC4036)

Test

2022/2023
DCC - FCUP

Name: _____ Student Nr: _____

-
- Duration: 1h.
 - Multiple choice questions
 - **Mark your answers with a circle.**
 - In case of a mistake, cancel the answer with a cross and do a circle over the new answer.
 - Each correct answer scores 0.5.
 - Incorrect answers will decrease your score in 0.25
 - This is an individual exam. Any attempt to communicate with a third party is regarded as fraud. Cell phones or other communicating devices are forbidden, as well as access to the internet.

1. In a project, two goals were proposed: 1. predict whether a loan application is going to be successfully paid or not; 2. better understand customers. The direct approach(es) to address this(goal)s are:

agrupamento

- ✓ classification for goal 1 and clustering for goal 2.
 regression for goal 1 and clustering for goal 2.
 clustering for goal 1 and classification for goal 2.
 classification for both.

2. It is common to reduce the dimensionality of the data. This operation aims to:

- ✓ decrease the number of variables that describe the examples.
 remove the number of examples and variables with missing values.
 decrease the number of examples.
 take a smaller sample of the data.

3. The MAE (mean absolute error) measure is easier to read than the MSE (mean squared error) measure because:

- (a) the scale of values is the same as the dependent variable.
(b) the scale of values is not the same as the dependent variables.
(c) places more emphasis on larger errors.
(d) is represented on a scale that is not [-1,1].

4. Which of the following cannot be varied in the generation of homogeneous ensembles:

- (a) algorithm.
- (b) data.
- (c) hyperparameter values.
- (d) model.

5. Linear regression is:

- (a) not very sensitive to overfitting due to the linear nature of the model.
- (b) not very sensitive to overfitting due to the high variance of the models.
- (c) very sensitive to overfitting due to the linear nature of the model.
- (d) very sensitive to overfitting due to the high variance of the models.

6. The single link proximity measure can be used in:

- (a) DBSCAN clustering algorithm.
- (b) hierarchical clustering algorithms.
- (c) k-means algorithm.
- (d) in any clustering algorithm.

7. In Principal Component Analysis (PCA), each component is a:

- (a) cluster.
- (b) visualization of the data set.
- (c) linear or a non-linear combination of the original attributes.
- (d) linear combination of the original attributes.

8. Each iteration of the k-Means algorithm performs the following operation:

- (a) choose the number of clusters.
- (b) find the k-nearest neighbours.
- (c) allocate each observation to the cluster with its k-nearest centroid.
- (d) allocate each observation to the cluster with the closest centroid.

9. Feature selection methods used before modeling:

- (a) can eliminate both redundant and variable variables without useful information for the model.
- (b) eliminate only redundant variables.
- (c) eliminate only variables with no useful information for the model.
- (d) eliminate redundant or variable variables with no useful information for the model, but the same method never eliminates both types of variables. WRAPPER

✓ 10. Feature extraction is the process that:

- (a) removes noise and outliers.
- (b) reduces data dimensionality. *SELECTION*
- (c) incorporates domain knowledge to create new features. *ENGINEERING*
- (d) obtains features from raw data.

✓ 11. The difference between classification and regression is:

- (a) in the dependent variables, implying the use of different evaluation measures.
- (b) in the independent variables, implying the use of different evaluation measures.
- (c) in all variables (independent and dependent), although it is possible to use the same evaluation measures in both cases.
- (d) in the dependent variables, although it is possible to use the same evaluation measures in both cases.

✓ 12. To measure the location of a variable with extreme values, one should use the:

- (a) standard deviation.
- (b) inter-quartile range.
- (c) mean.
- (d) median.

✓ 13. Suppose we have the following set of values for two variables

$$(X, Y) = \{(1, 50), (2, 40), (3, 32), (4, 24), (5, 18), (6, 12), (7, 8), (8, 4), (9, 2), (10, 0)\}$$

and we want to measure how correlated they are. We can say that:

- (a) the Pearson correlation coefficient will indicate a stronger correlation than the Spearman rank correlation coefficient.
- (b) the Pearson correlation coefficient will indicate a weaker correlation than the Spearman rank correlation coefficient.
- (c) the Pearson and the Spearman rank correlation coefficients will indicate the same measure of correlation.
- (d) the Pearson correlation coefficient will indicate that there is no correlation; only the Spearman rank correlation coefficient will indicate that.

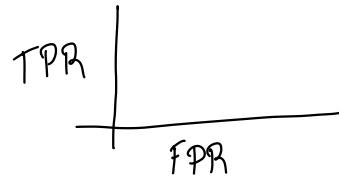
14. Extreme Gradient Boosting (XGBoost) is a:

- (a) variance reduction method.
- (b) bias reduction method.
- (c) bias and variance reduction method.
- (d) neither bias nor variance reduction method.



15. The Area Under Curve (AUC) is an evaluation metric that:

- (a) gives equal importance to positive and negative classes.
- (b) combines in a single measure the true and false positive ratios.
- (c) is able to substitute the confusion matrix.
- (d) evaluates the accuracy.



16. The k value in the k-nearest neighbours algorithm should be selected taking into account the following:

- (a) If $k=1$, it is more sensitive to noise in the labels (i.e. incorrectly labelled examples).
- (b) If k is large (e.g. $k \geq 100$), it is more sensitive to noise in the labels (i.e. incorrectly labelled examples).
- (c) If $k=1$, it is more sensitive to the number of attributes.
- (d) If k is large (e.g. $k \geq 100$), it is more sensitive to the number of attributes.

Data Mining Study Test

Q1. What is the main goal of descriptive analytics?

- a) Predict future outcomes
- b) Describe or summarize data
- c) Perform real-time analysis
- d) Identify anomalies in data

Q2. Which of the following is NOT a characteristic of Big Data?

- a) Volume
- b) Variety
- c) Velocity
- d) Validity

Q3. In CRISP-DM, what is the main goal of the 'Business Understanding' phase?

- a) Understand the data
- b) Understand business objectives
- c) Build a predictive model
- d) Evaluate the model's performance

Q4. What is the difference between supervised and unsupervised learning?

- a) Supervised learning involves labeled data
- b) Unsupervised learning involves labeled data
- c) Supervised learning does not require data
- d) Unsupervised learning requires test data

Q5. Which of the following is a distance measure commonly used in clustering?

- a) Euclidean distance
- b) Manhattan distance

c) Minkowski distance

d) All of the above

✓ Q6. Which of the following describes the **Ratio scale of attributes**?

- a) There is an **absolute zero**
- b) Values vary within an interval
- c) Values have no relationship
- d) There is an arbitrary zero

✓ Q7. What is one of the main goals of **clustering** in descriptive modeling?

- a) Predict the output values
- b) **Find natural groupings of data**
- c) Increase dimensionality
- d) Reduce the data quality issues

✓ Q8. What is the function of **data normalization**?

- a) It reduces the number of variables
- b) **It standardizes different scales**
- c) It converts numerical data to categorical data
- d) It removes irrelevant data

✓ Q9. What is a common problem in **high-dimensional data sets**?

- a) High correlation
- b) **Curse of dimensionality**
- c) Lack of resolution
- d) Noise in the data

Q10. What is the difference between **precision and recall**?

- a) Precision is the number of correct positive results, recall is false negatives

- b) Precision is the number of false positives, recall is the number of true negatives
- c) Precision is the number of correct positive results, recall is the true positive rate
- d) Precision is false positive rate, recall is false negative rate

Q11. Which of the following is a common method for handling missing data?

- a) Removing the entire dataset
- b) Using imputation techniques
- c) Ignoring the missing values
- d) Converting numerical to categorical data

Q12. What is overfitting in predictive modeling?

- a) A model that performs well on unseen data
- b) A model that is too simple
- c) A model that is too complex and performs well only on training data
- d) A model with too few variables

Q13. Which is the first step in the CRISP-DM process?

- a) Data Understanding
- b) Business Understanding
- c) Data Preparation
- d) Modeling

Q14. What is the purpose of PCA (Principal Component Analysis)?

- a) To reduce the dimensionality of data
- b) To increase the number of variables
- c) To add noise to the data
- d) To visualize missing data

Q15. Which one of the following techniques is used for binary classification?

a) K-means

b) Decision Trees

c) DBSCAN

d) Hierarchical Clustering

Q16. What is the silhouette coefficient used for?

- a) To measure model accuracy
- b) To evaluate clustering cohesion and separation
- c) To reduce dimensionality
- d) To correct missing values

New Data Mining Study Test

Q1. What are the three key characteristics of Big Data?

- a) Volume, Variety, Velocity
- b) Volume, Validity, Variance
- c) Velocity, Value, Volume
- d) Variety, Verification, Volume

Q2. Which type of attribute scale involves order but no defined distances?

- a) Nominal
- b) Ordinal
- c) Interval
- d) Ratio

Q3. In CRISP-DM, what is the purpose of 'Data Understanding'?

- a) Collecting initial data
- b) Cleaning the data
- c) Transforming the data
- d) Verifying business objectives

Q4. What type of learning is used when there is no target variable?

- a) Supervised learning
- b) Unsupervised learning
- c) Reinforcement learning
- d) Predictive learning

Q5. Which of the following measures how far apart clusters are in a clustering method?

- a) Cluster cohesion
- b) Cluster separation

c) Distance matrix

d) Intra-cluster distance

Q6. What kind of data quality issue is represented by an outlier?

a) Noise

b) Missing data

c) Duplicate data

d) Inconsistent data

Q7. Which technique is used to reduce the number of variables in a dataset?

a) Dimensionality reduction

b) Data normalization

c) Data cleaning

d) Feature engineering

Q8. What is the primary use of confusion matrices?

a) To show the relationship between independent variables

b) To calculate the precision and recall

c) To track missing values

d) To display the predictions of a clustering algorithm

Q9. What is an example of a supervised learning algorithm?

a) K-means

b) Hierarchical clustering

c) Decision tree

d) PCA

Q10. Which type of clustering method generates a hierarchy of clusters?

a) K-means

b) Agglomerative clustering

c) DBSCAN

d) Fuzzy clustering

Q11. What is the 'curse of dimensionality'?

a) The issue of missing data

b) The issue of too few attributes

c) The difficulty in analyzing high-dimensional data

d) The difficulty in clustering small datasets

Q12. What is a common way to deal with missing values?

a) Discard the dataset

b) Use a default value

c) Impute based on similar cases

d) Ignore the missing data

Q13. What is the objective of predictive modeling?

a) Describing past data

b) Predicting the outcome for unseen data

c) Summarizing the data

d) Cleaning the data

Q14. What is the purpose of 'Feature Engineering'?

a) To reduce the number of variables

b) To create new variables based on domain knowledge

c) To clean the data

d) To remove redundant features

Q15. Which type of evaluation is used for classification models?

a) Silhouette coefficient

b) Confusion matrix

c) Elbow method

d) Sum of squared errors

Q16. What does PCA aim to achieve?

a) To cluster the data

b) To reduce the number of dimensions

c) To normalize the data

d) To predict future data

Third Data Mining Study Test

Q1. What is the purpose of the 'Data Preparation' phase in CRISP-DM?

- a) To collect initial data
- b) To clean and format data
- c) To evaluate the model
- d) To summarize data insights

Q2. Which method is used to handle imbalanced data in classification?

- a) Random sampling
- b) Stratified sampling
- c) Data normalization
- d) Overfitting

Q3. What does data 'sparsity' refer to?

- a) Data with low variance
- b) Data with missing values
- c) Data with mostly zero values
- d) Data with high resolution

Q4. What is the goal of feature selection in data preparation?

- a) To create new features
- b) To remove irrelevant features
- c) To normalize data
- d) To clean noisy data

Q5. What is the difference between parametric and non-parametric models?

- a) Parametric models are simpler
- b) Non-parametric models assume a fixed structure

c) Parametric models have a fixed number of parameters

d) Non-parametric models are always better

Q6. Which technique is often used to handle high-dimensional data?

a) K-means clustering

b) PCA (Principal Component Analysis)

c) Decision Trees

d) Random Sampling

Q7. What is the main disadvantage of the k-means algorithm?

a) It requires the number of clusters in advance

b) It works only with categorical data

c) It can handle outliers easily

d) It always provides an optimal solution

Q8. What does overfitting imply in machine learning?

a) The model captures noise in the training data

b) The model generalizes well to new data

c) The model is too simple

d) The model does not work on test data

Q9. Which technique is used for anomaly detection?

a) DBSCAN

b) k-means

c) PCA

d) Random forests

Q10. What is a binary classification problem?

a) A problem with more than two classes

b) A problem where the target variable has two possible outcomes

c) A problem without any target variable

d) A clustering problem

Q11. Which distance metric is commonly used in hierarchical clustering?

a) Cosine distance

b) Euclidean distance

c) Minkowski distance

d) Hamming distance

Q12. What is a common evaluation metric for classification problems?

a) Sum of squared errors

b) Accuracy

c) Silhouette coefficient

d) Elbow method

Q13. Which method deals with non-linear decision boundaries?

a) Linear regression

b) Decision trees

c) Logistic regression

d) PCA

Q14. What does the term 'outlier' refer to?

a) A data point that follows the general pattern

b) A data point significantly different from others

c) A missing value

d) A duplicate data point

Q15. What does the elbow method help determine?

a) The number of clusters in k-means

b) The correct class label in classification

c) The number of features to keep

d) The size of the training data set

Q16. Which problem arises when a model performs well on training data but poorly on unseen data?

a) Underfitting

b) Overfitting

c) Noise

d) Data leakage

Teste 1: Introdução à Mineração de Dados

- ✓ 1. What is the main objective of data mining?
a) To collect raw data
(b) To discover patterns and knowledge from data
c) To organize data in tables
d) To visualize data
- ✓ 2. Which of the following is a key challenge in Big Data?
a) Lack of storage
(b) High dimensionality and complexity
c) Not enough data sources
d) Too simple to analyze
- ✓ 3. What are the three Vs of Big Data?
(a) Volume, Variety, Velocity
b) Volume, Value, Validity
c) Variety, Visualization, Velocity
d) Variety, Volume, Verification
- ✓ 4. What is the purpose of the CRISP-DM model?
a) To clean data
(b) To provide a blueprint for data mining
c) To transform raw data
d) To create machine learning algorithms
- ✓ 5. Which phase in CRISP-DM focuses on defining business objectives?
a) Data Preparation
b) Modeling
(c) Business Understanding
d) Evaluation
- ✓ 6. Which of the following is a common application of data mining in telecommunications?
a) Cross-selling
(b) Response scoring
c) Drug development
d) Stock forecasting

- ✓ 7. In the Knowledge Discovery in Data (KDD) process, what follows data selection?
 a) Data cleaning
b) Data integration
c) Data transformation
d) Data modeling
- ✓ 8. Which of these fields contributes to data mining?
 a) Statistics
b) Data structures
c) Software engineering
 d) All of the above
- ✓ 9. What is the primary focus of data mining?
 a) Discovering hidden knowledge in large datasets
b) Collecting data from multiple sources
c) Creating dashboards
d) Designing relational databases
- ✓ 10. Which of the following is NOT a data mining definition?
 a) Extracting actionable information from large datasets
 b) Building visualizations of data
c) Summarizing data in novel ways
d) Finding relationships in observational data sets
- ✓ 11. Which of these are common domains for data mining?
 a) Market segmentation, fraud detection, medical diagnosis
b) Web development, software testing
c) Graphic design, social media marketing
d) Industrial design, architecture
- ✓ 12. What is metadata?
 a) Data about other data
b) Large sets of unstructured data
c) Numerical information from social networks
d) Information visualization techniques

✓ 13. Which phase of CRISP-DM involves assessing data quality?

- (a) Data Understanding
- b) Deployment
- c) Business Understanding
- d) Evaluation

✓ 14. Which of the following statements is true about data mining?

- a) It always provides accurate predictions
- (b) It is the process of discovering patterns in large datasets
- c) It involves primarily collecting data
- d) It focuses only on text analysis

✓ 15. What are the five phases of CRISP-DM?

- (a) Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, Deployment
- b) Collection, Cleaning, Exploration, Testing, Evaluation
- c) Business Analysis, Data Collection, Model Building, Testing, Review
- d) Data Selection, Transformation, Modeling, Reporting, Analysis

✓ 16. What does data exploration involve?

- a) Verifying data quality
- b) Creating business goals
- c) Designing dashboards
- (d) Analyzing patterns and relationships in data

Teste 2: Compreensão de Dados

1. What is an attribute in a data set?
 a) A property of an object
b) A specific case in the dataset
c) A variable with only one possible value
d) A numeric variable with no missing data
2. What type of data involves explicit or implicit relationships between cases?
 a) Nondependency-oriented data
 b) Dependency-oriented data
c) Ordinal data
d) Continuous data
3. Which attribute scale involves ordered values but undefined intervals?
 a) Nominal
 b) Ordinal
c) Interval
d) Ratio
4. What does the interquartile range (IQR) measure?
 a) The mean
 b) The spread of the middle 50% of the data
c) The number of extreme values
d) The mode of the dataset
5. Which type of graph is typically used to show the distribution of continuous data?
 a) Barplot
b) Pie chart
 c) Histogram
d) Scatter plot
6. What does a Pearson correlation coefficient indicate?
 a) The linear relationship between two variables
b) The mean of two variables
c) The variability in one variable
d) The ratio of two variables

- ✓ 7. What is an outlier in data?
a) A value that fits the general pattern
(b) A value that deviates significantly from others
c) A missing value
d) A duplicate data entry
- ✓ 8. Which characteristic of data refers to the number of attributes in the data set?
a) Sparsity
(b) Dimensionality
c) Resolution
d) Frequency
- ✓ 9. What is data summarization?
a) The process of collecting new data
(b) The process of representing data compactly
c) The process of removing missing values
d) The process of clustering data
- ✓ 10. Which type of attribute can take only distinct or separate values?
a) Continuous
b) Discrete
c) Ordinal
(d) Nominal
- ✓ 11. In which data type can measurements take on any value within a given range?
a) Categorical
b) Discrete
(c) Continuous
d) Nominal
- ✓ 12. Which type of data analysis focuses on summarizing one variable at a time?
(a) Univariate
b) Multivariate
c) Bivariate
d) Cluster analysis

- ✓ 13. What is the purpose of data visualization?
a) To create predictive models
(b) To help detect patterns and trends
c) To build decision trees
d) To normalize the data
- ✓ 14. What type of visualization is best for comparing proportions?
a) Scatter plot
b) Pie chart
c) Box plot
(d) Bar chart
- ✓ 15. Which attribute scale involves numeric values with a true zero point?
a) Interval scale
b) Ordinal scale
c) Nominal scale
(d) Ratio scale
- ✓ 16. Which type of data tends to have gaps or missing values?
(a) Time series data
b) Cross-sectional data
c) Discrete data
d) Nominal data

Teste 3: Preparação de Dados

1. **What is the primary goal of data preparation?**
a) To collect new data
(b) To clean, format, and transform the data for analysis
c) To summarize results
d) To model data

2. **Which of the following is a common data quality issue?**
a) Well-structured datasets
(b) Missing or inconsistent values
c) Sufficient data volume
d) Data without noise

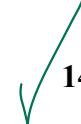
3. **What does data transformation involve?**
a) Creating new models
(b) Converting raw data into meaningful formats
c) Reducing the volume of data
d) Removing irrelevant features

4. **What is the purpose of data cleaning?**
a) To find patterns in the data
(b) To remove noise, missing, and incorrect values
c) To perform predictive analysis
d) To sample large datasets

5. **Which of the following refers to the reduction of the number of features in a dataset?**
a) Feature extraction
b) Data cleaning
(c) Dimensionality reduction
d) Sampling

6. **How can missing values in a dataset be handled?**
a) Ignore missing values
b) Use imputation techniques
c) Remove incomplete records
(d) All of the above

- ✓ 7. What is a common method for reducing the number of features in a dataset?
a) Random Sampling
(b) Principal Component Analysis (PCA)
c) Decision trees
d) K-means clustering
- ✓ 8. What is feature engineering?
a) Adding irrelevant features to the model
(b) Using domain knowledge to create useful features
c) Automatically generating predictions
d) Reducing the number of samples
- ✓ 9. Which method of handling missing values uses central tendency measures?
a) Data cleaning
(b) Imputation
c) Binarization
d) Sampling
- ✓ 10. What is binarization in data transformation?
a) Creating new variables
b) Reducing data dimensionality
(c) Converting categorical values to binary format
d) Removing outliers
- ✓ 11. What is the 'curse of dimensionality'?
a) The issue of having too few features
(b) The difficulty in analyzing high-dimensional data
c) Lack of enough data for analysis
d) Insufficient variability in data
- ✓ 12. Which of the following is a technique for reducing data dimensionality?
a) Data sampling
(b) PCA (Principal Component Analysis)
c) Clustering
d) Cross-validation
- ✓ 13. What is outlier detection in data cleaning?
a) Identifying data points that fit the general pattern
b) Removing duplicate records
(c) Identifying data points that deviate significantly from other data points
d) Filling missing values



14. What does **noise** in a dataset refer to?

- a) Data points that do not affect the analysis
- b) Irrelevant or random data that can affect the model
- c) Missing values
- d) Duplicate records



15. What is **stratified sampling**?

- a) Randomly selecting samples
- b) Sampling data to reflect the original proportions of different groups
- c) Sampling with replacement
- d) Sampling based on cluster analysis



16. How does **feature extraction** help in data preparation?

- a) By creating new models
- b) By removing irrelevant features
- c) By transforming raw data into features useful for modeling
- d) By removing missing values

Teste 4: Modelagem Descritiva

1. **What is the main goal of descriptive modeling?**
a) To predict future outcomes
 b) To describe patterns in existing data
c) To clean and prepare data
d) To create decision trees
2. **Which of the following is often associated with descriptive modeling?**
 a) Clustering
b) Regression
c) Classification
d) Reinforcement learning
3. **What is a similarity measure?**
a) A metric for how different two data objects are
 b) A metric for how alike two data objects are
c) A metric for identifying missing data
d) A method for reducing data dimensionality
4. **Which distance metric is commonly used in clustering algorithms?**
a) Manhattan distance
 b) Euclidean distance
c) Minkowski distance
d) All of the above
5. **Which clustering algorithm uses a partition-based approach?**
 a) DBSCAN
 b) k-means
 c) Agglomerative hierarchical clustering
 d) PAM
6. **What is the purpose of clustering in data analysis?**
 a) To find natural groupings within the data
b) To predict target variables
c) To standardize data
d) To fill missing values



7. What does the **elbow method** help to determine?

- (a) The optimal number of clusters in k-means
- b) The error rate of a model
- c) The distance between clusters
- d) The correlation between variables



8. Which clustering method does not require the number of clusters to be specified?

- a) k-means
- (b) **DBSCAN**
- c) Hierarchical clustering
- d) All clustering methods require specifying the number of clusters



9. What is the main **disadvantage of k-means clustering**?

- a) It is too slow for large datasets
- (b) **It is sensitive to outliers**
- c) It can only handle categorical data
- d) It requires labeled data



10. Which method can handle clusters of varying shapes and sizes?

- a) k-means
- (b) **DBSCAN**
- c) PCA
- d) Linear regression



11. What does the **silhouette coefficient** measure?

- a) The separation between clusters
- b) The compactness of clusters
- c) The accuracy of a model
- (d) **Both a and b**



12. In clustering, what is an **agglomerative method**?

- a) A top-down approach
- (b) **A method that starts with individual points and merges them into clusters**
- c) A method that starts with one cluster and splits it into smaller clusters
- d) A technique for normalizing data



13. Which clustering method is more **robust to outliers**?

- a) k-means
- (b) **PAM (Partitioning Around Medoids)**
- c) Agglomerative clustering
- d) DBSCAN



14. Which coefficient combines both cohesion and separation in clustering?

- a) Jaccard coefficient
- (b) Silhouette coefficient**
- c) Cosine similarity
- d) Minkowski distance



15. What is the dendrogram used for in hierarchical clustering?

- a) To display the number of clusters
- (b) To visualize the hierarchical structure of clusters**
- c) To calculate the cluster centroids
- d) To determine missing values



16. What is a prototype in clustering?

- (a) A representative point for each cluster**
- b) The first cluster formed
- c) A method for reducing dimensionality
- d) A distance metric used for clustering

Teste 5: Modelagem Preditiva

1. What is the primary goal of predictive modeling?
a) To summarize the data
(b) To predict unknown outcomes based on known data
c) To cluster the data
d) To perform descriptive analysis

2. Which of the following is a common type of predictive modeling?
a) Clustering
(b) Classification
c) Dimensionality reduction
d) Data cleaning

3. What is the difference between classification and regression?
a) Classification deals with numerical output, while regression deals with categorical output
(b) Classification deals with categorical output, while regression deals with numerical output
c) Classification is unsupervised, regression is supervised
d) Regression involves clustering, classification involves decision trees

4. Which of the following is an evaluation metric for classification models?
a) Mean Squared Error
(b) Accuracy
c) Silhouette coefficient
d) Within-cluster sum of squares

5. What does an ROC curve represent?
(a) Relationship between true positive rate and false positive rate
b) Relationship between precision and recall
c) The distribution of class probabilities
d) The accuracy of a model

6. Which of the following is true about overfitting?
a) It happens when a model is too simple
(b) It occurs when a model performs well on training data but poorly on unseen data
c) It leads to better generalization
d) It happens only in unsupervised learning

- ✓ 7. Which of the following techniques can be used to prevent overfitting?
a) Using more features
(b) Cross-validation
c) Increasing the number of training samples
d) Using a deeper model
- ✓ 8. What does underfitting imply?
a) The model is too complex
(b) The model is too simple to capture patterns in the data
c) The model performs well on both training and test data
d) The model generalizes well to unseen data
- ✓ 9. Which of the following is an example of a classification algorithm?
a) Linear regression
(b) Decision trees
c) k-means
d) Principal Component Analysis
- ✓ 10. What is the purpose of a confusion matrix?
a) To visualize the distribution of clusters
(b) To evaluate the performance of classification models
c) To determine missing values
d) To calculate the number of features
- ✓ 11. Which of the following metrics combines precision and recall?
a) Accuracy
b) ROC curve
(c) F1-score
d) Mean Squared Error
- ✓ 12. Which classification problem deals with two possible class labels?
(a) Binary classification
b) Multiclass classification
c) Regression
d) Clustering
- ✓ 13. Which of the following methods is used for binary classification?
a) k-means
(b) Logistic regression
c) PCA
d) DBSCAN



14. **What is cross-validation?**

- a) A technique used to reduce the dimensionality of data
- (b) A method used to estimate the performance of a model on unseen data**
- c) A type of clustering algorithm
- d) A method for normalizing data



15. **What is the goal of predictive modeling in medical diagnosis?**

- a) To cluster patients into groups
- (b) To predict the correct diagnosis for a new patient**
- c) To determine patient satisfaction
- d) To visualize patient data



16. **Which method is often used to evaluate regression models?**

- a) Accuracy
- b) Confusion matrix
- (c) Mean Absolute Error**
- d) F1-score

K-Nearest Neighbors

→ Abordagem baseada em Distância

- Não aprende nenhum modelo a partir dos dados — não aprende uma função para mapear as variáveis preditivas na variável-alvo
- É um algoritmo de aprendizagem baseado nas instâncias — aprende por analogia (medida de semelhança entre casos)
- Aplicável a qualquer problema com dados suficientes, porque não faz nenhuma suposição sobre a forma funcional a aproximar.

Método:

1. Escolher o número K e a medida de distância d
2. Para um caso de teste x :
 - a. Encontrar os K casos mais próximos nos dados de treino de acordo com d
 - b. usar os valores da variável alvo desses casos para obter a previsão para x

Classificação: a previsão é a classe maioria

Regressão: a previsão é a média dos valores-alvo

Estimativa Global: procura o K ideal para um dado conjunto de dados

Estimativa Local: tenta estimar o K ideal para cada caso de teste

- A complexidade cresce linearmente com o número de casos
- Tempo de Treino rápido
- Tempo de Teste lento

Aprendizagem Bayesiana

→ Abordagem Probabilística

Teorema de Bayes: $P(Y|X) = \frac{P(X|Y) P(Y)}{P(X)}$

Bayes Ingênuo: assume que os atributos são independentes dada a classe

Assuma-se a função-árvore $f: X \rightarrow Y$ em que cada instância x é descrita por p atributos

O valor mais provável de $f(x)$ é \hat{y} :

$$\hat{y} = \operatorname{argmax}_{y_j \in Y} P(y_j | x_1, \dots, x_p) = \operatorname{argmax}_{y_j \in Y} \frac{P(x_1, \dots, x_p | y_j) P(y_j)}{P(x_1, \dots, x_p)} = \operatorname{argmax}_{y_j \in Y} P(x_1, \dots, x_p | y_j) P(y_j)$$

$$P(x_1, \dots, x_n | y_j) = \prod_i P(x_i | y_j)$$

$$\hat{y} = \operatorname{argmax}_{y_j \in V} P(y_j) \prod_i P(x_i | y_j)$$

Como estimar as probabilidades?

1. Assumir um problema de decisão com p variáveis preditivas
2. Cada variável assume K valores
3. A probabilidade conjunta requer estimar K^p probabilidades
4. Assumir que as variáveis são condicionalmente independentes dada a classe, apenas requer estimar $K \times p$ probabilidades

Atributos Categóricos: a probabilidade é estimada a partir de tabelas de freqüência

Atributos Numéricos: a probabilidade estima-se assumindo uma distribuição normal

Regressão

$$D = \{(x_i, y_i)\}_{i=1}^N$$

- x_i - vetor de funcionalidades com p variáveis preditoras
- $y_i \in \mathbb{R}$ - variável-alvo numérica Y
- $y = f(x)$: função desconhecida

Objetivo: aprender a melhor aproximação da função desconhecida f

Abordagem:

1. aproximar $f()$ por $h_\theta(x)$
2. seguir um critério de preferência sobre o espaço de parametrização θ
3. procurar pelo "melhor" $h()$ de acordo com os critérios e o conjunto de dados

Modelo de Regressão: função que transforma um vetor de valores dos preditores (x) num número real (y), assumindo a relação $y_i = h_\theta(x_i) + \epsilon_i$

$h_\theta(x_i)$: modelo de regressão com o conjunto de parâmetros θ

ϵ_i : erros de observação - resíduos

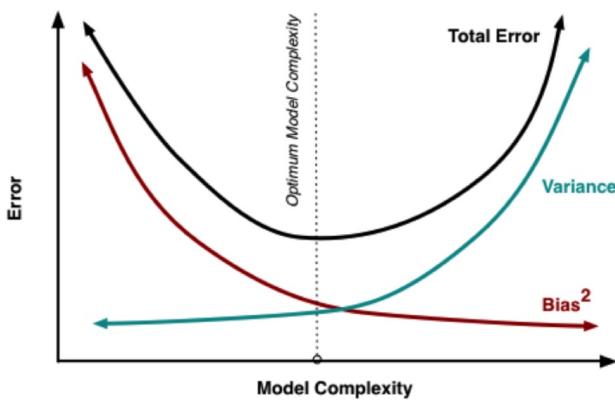
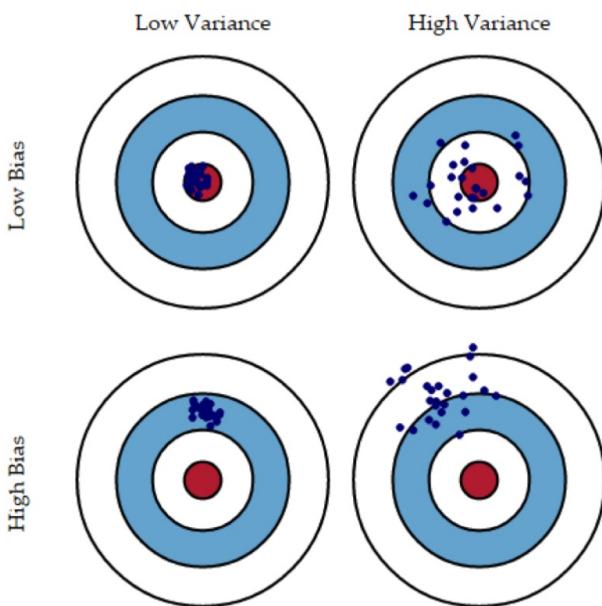
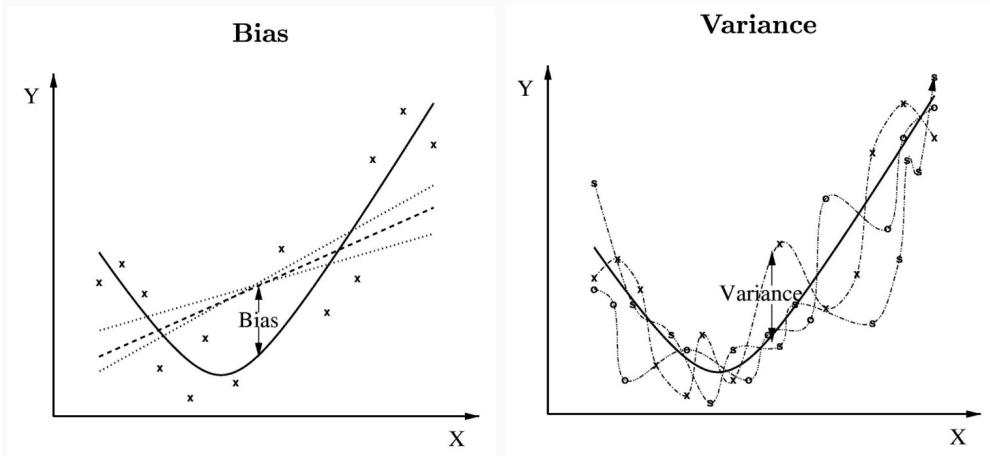
Dado um conjunto de treino $D = \{(x_i, y_i)\}_{i=1}^N$, inicia-se um modelo de regressão $\hat{y} = h_\theta(x)$ e as funções de perda medem a qualidade das previsões

Perda Quadrática: $L(y, \hat{y}) = (y - \hat{y})^2$

Perda Absoluta: $L(y, \hat{y}) = |y - \hat{y}|$

Perda Zero-Um: $L(y, \hat{y}) = 0$ se $y = \hat{y}$, 1 caso contrário

No conjunto de treino, obtém-se a Perda Esperada, isto é, $E[L(y, \hat{y})]$
VIÉS + VARIÂNCIA



MÉTRICAS DE AVALIAÇÃO

\hat{y}_i : previsão do modelo na avaliação para o caso i

y_i : valor verdadeiro da respetiva variável-alvo

Erro Médio Quadrado: $MSE = \frac{1}{N} \sum_{i=1}^N (\hat{y}_i - y_i)^2 \rightarrow RMSE = \sqrt{MSE}$
→ medido numa unidade que é o quadrado da escala original

Erro Médio Absoluto: $MAE = \frac{1}{N} \sum_{i=1}^N |\hat{y}_i - y_i|$
→ medido na mesma unidade da escala original

MÉTRICAS RELATIVAS

\bar{y}_i : valor médio da variável y para o caso i

Erro Médio Quadrado Normalizado: $NMSE = \frac{\sum_{i=1}^N (\hat{y}_i - y_i)^2}{\sum_{i=1}^N (\bar{y} - y_i)^2}, 0 \leq NMSE \leq 1$

Erro Médio Absoluto Normalizado: $NMAE = \frac{\sum_{i=1}^N |\bar{y}_i - y_i|}{\sum_{i=1}^N |\bar{y} - y_i|}, 0 \leq NMAE \leq 1$

Coeficiente de Correlação: $\rho_{\hat{y}, y} = \frac{\sum_{i=1}^N (\hat{y}_i - \bar{y})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^N (\hat{y}_i - \bar{y})^2 \sum_{i=1}^N (y_i - \bar{y})^2}}, -1 \leq \rho \leq 1$
→ força da relação $y \leftrightarrow \hat{y}$

Coeficiente de Determinação: $R^2 = 1 - \frac{\sum_{i=1}^N (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^N (y_i - \bar{y})^2}, 0 \leq R^2 \leq 1$
→ nível de variabilidade explicada

Métricas Quadradas: amplificam os erros grandes

Métricas Absolutas: tratam todos os erros da mesma forma - erro "típico"

Métricas Relativas: independentes do domínio

MÉTODOS DE REGRESSÃO LINEAR

→ Fórmula Matemática

Regressão Linear Simples: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$

y_i : variável-alvo

β_0 : interseção em y

x_i : variável preditora

β_1 : declive

ϵ_i : erro

Regressão Linear Múltipla: $y = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p$

Objetivo: encontrar o vetor de parâmetros β que minimiza $SSE = \sum_{i=1}^N (y_i - (\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p))^2$

Como?

$$\beta = (X^T X)^{-1} \cdot X^T \cdot y \quad \text{OU} \quad \text{Singular Value Decomposition (SVD)}$$

Problema da Multicolinearidade: variáveis preditoras altamente correlacionadas aumentam a variância, tornando as previsões do modelo instáveis e altamente dependentes do treino

↓ SOLUÇÃO

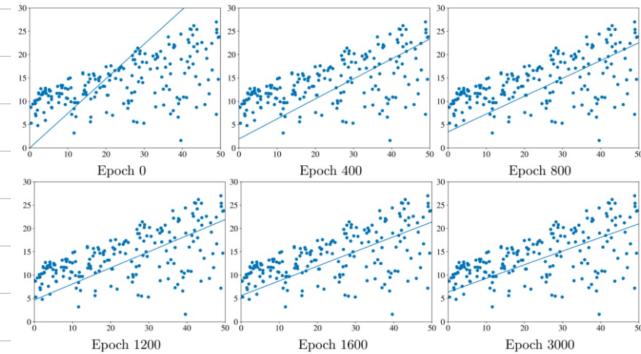
Regularização: afinar o modelo para alcançar um bom trade-off viés-variancia

Encolhimento: adicionar um viés à estimativa da regressão para garantir que os coeficientes não, em média, forem em magnitude

Regressão "Ridge": encolhe os coeficientes usando os menores quadrados ao adicionar o termo de regularização $\lambda \sum_i \beta_i^2$ (norma de L_2) — $\sum_{i=1}^N (y_i - (\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p))^2 + \lambda \sum_i \beta_i^2$

Regressão "Lasso": encolhe os coeficientes usando os menores valores absolutos ao adicionar o termo de regularização $\lambda \sum_i |\beta_i|$ (norma de L_1) — $\sum_{i=1}^N (y_i - (\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p)) + \lambda \sum_i |\beta_i|$

Descida de Gradiente: algoritmo iterativo de otimização para encontrar o mínimo da soma função (a função de erro) — calcula a derivada parcial da função de perda em relação a cada coeficiente e atualiza-os até a perda alcançar um valor pequeno (≈ 0)



1. Batch: calcula o erro para cada exemplo dos dados de treino e só depois atualiza o modelo
2. Estocástico: calcula o erro e atualiza o modelo para cada exemplo dos dados de treino
3. Mini-Batch: divide os dados de treino em pequenos grupos que não usados para calcular o erro e atualizar o modelo

✓ intuitivo; eficiente; reconhecido; simples; eficaz quando a linearidade se verifica
 ✗ assumptions demasiado fortes na forma das funções desconhecidas

Outros:

1. KNN
2. LOESS: combina muitos modelos de regressão como KNN } **NÃO-PARAMÉTRICOS**
3. MARS: estende a regressão linear
4. Support Vector Machines
5. Artificial Neural Networks
6. Random Forests: mistura árvores CART
7. Extreme Gradient Boosting: distribuição otimizada de gradiente por árvores paralelas

Modelos baseados em Árvores

→ Abordagem Lógica

- Os modelos baseados em árvores têm uma natureza recursiva dividir-e-conquistar
- Cada nó é uma partição do espaço de input

Nó Interno: com um teste sobre o valor da variável preditora

Nó Folha: contém o valor da variável-alvo

Cada caminho desde a raiz até uma folha é um conjunto de testes lógicos que define uma região no espaço de preditores — a previsão para um novo caso de teste é obtida ao seguir um caminho desde a raiz até uma folha de acordo com os seus valores preditores

Árvores de Classificação e Regressão (CART): árvores de particionamento recursivo binário com testes lógicos em cada nó — o critério de preferência usado tem impacto na forma como o melhor teste para cada nó é selecionado e na forma como a árvore evita "overfitting" dos dados de treino

Testes de Divisão para Preditores Numéricos: variável contínua A

1. Ordenar o conjunto $V_{A,D}$ — valores de A que ocorrem nos dados D
2. Avaliar todos os testes $A \leq k$ em que x toma como valores todos os pontos médios entre cada valor sucessivo no conjunto ordenado

Testes de Divisão para Preditores Categóricos: variável categórica A

1. Avaliar todas as combinações possíveis de subconjuntos de valores em $V_{A,D}$ — valores de A que ocorrem nos dados D

ÁRVORES DE REGRESSÃO

Que valor põe nas folhas? O atributo escolhido em cada divisão é aquele que minimiza algum critério de estimativa de erro

Em Árvores de Regressão "Least Squares", quer-se minimizar $\text{Err}(t) = \frac{1}{n_T} \sum_{D_t} (y_i - k_t)^2$

D_t: partição de casos no nó t

n_T: cardinalidade da partição

k_t: constante que minimiza o erro — valor médio da variável-erro (\bar{y}_t)

Como encontrar o melhor teste de divisão? Comparar os erros nas partições da não-expansão e expansão do nó t e escolher a divisão que minimiza o erro, isto é, $s^* = \arg\min_s \text{Err}(t) - \text{Err}(s, t)$

ÁRVORES DE CLASSIFICAÇÃO

Que valor põe nas folhas? A classe majoritária dos casos que estão nessa partição

Índice Gini: $\text{Gini}(D) = 1 - \sum_{i=1}^c p_i^2$ → probabilidade da classe i estimada pela frequência observada

mede a impureza de um conjunto de dados em relação ao conjunto das classes às quais os seus exemplos pertencem — quanto maior for a probabilidade de uma dada classe em relação às outras, mais pura é a partição

Como encontrar o melhor teste de divisão? Comparar a impureza nas partições de não-expansão e expansão do nó t e escolher a divisão que maximiza a redução de impureza, isto é, $s^* = \arg\max_s \text{Gini}(t) - \text{Gini}(s, t)$, $\text{Gini}(s, t) = \frac{n_L}{n_T} \text{Gini}(t_L) + \frac{n_R}{n_T} \text{Gini}(t_R)$

Quando parar de crescer árvores? Apesar de os resultados globais melhorarem conforme a árvore cresce, conforme a árvore desce as decisões de divisão não são tomadas baseadas em conjuntos cada vez mais pequenos e, por isso, potencialmente menos fiáveis — para evitar "overfitting", deve encontrar-se o tamanho ótimo

Pré-Poda: parar de crescer a árvore se a estimativa de qualidade indicar que não vale a pena continuar

Ex: mínimo de casos num nó/folha; profundidade máxima

Pós-Poda: crescer uma árvore excessivamente grande e usar algum procedimento estatístico para podar ramos não fiáveis de acordo com estimativas de erro

CART:

1. crescer uma árvore demasiado grande
2. gerar uma sequência de sub-árvores
3. usar validação cruzada para estimar o erro
 - a. critério de erro-complexidade para árvores de regressão
 - b. critério de custo-complexidade para árvores de classificação
4. usar a regra X-SE para selecionar a melhor sub-árvore

- modelos preditivos interpretáveis
- variáveis preditoras numéricas e categóricas
- não necessita de escalamento de funcionalidades
- seleciona de variáveis e lida com valores em falta
- eficiente

- suscetível a pequenas variações
- modelos instáveis

Avaliação Empírica

Hiperparâmetro: parâmetro cujo valor controla o processo de aprendizagem

Ajustamento de Hiperparâmetros: metodologia experimental para evitar overfitting

1. Dividir os dados de treino: treino + validação
2. Com base no desempenho do modelo, encontrar os melhores hiperparâmetros
3. Treinar um modelo com hiperparâmetros ótimos em todos os dados de treino
4. Testar o modelo no conjunto "holdout"

"Grid Search":

1. definir grade
2. aprender e avaliar os modelos para todas as combinações possíveis
3. escolher o melhor

Pesquisa Aleatória:

1. definir o domínio
2. gerar combinações aleatoriamente
3. aprender e avaliar os modelos para as combinações
4. escolher o melhor

METODOLOGIAS DE AVALIAÇÃO

Tarefa Preditiva: aprender uma aproximação para uma função desconhecida $y = f(x)$ dado um conjunto de valores (de treino) $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$ com valores conhecidos $y = f(x)$

Critério de Avaliação de Desempenho: métrica do desempenho preditivo

Estimativa de Resubstuição: estimativa do desempenho do modelo obtida ao avaliá-lo no conjunto de dados de treino — não fiável e demasiado otimista

• O melhor a fazer é avaliar o modelo em novas amostras da distribuição

Objetivo: obter uma estimativa fiável do erro de previsão esperado de um modelo numa distribuição de dados desconhecidos — para ser fiável, deve ser baseada em casos nunca antes vistos (conjunto de teste)

→ "Os dados usados para avaliar (ou comparar) quaisquer modelos não podem ser vistos durante o desenvolvimento do modelo"!

Metodologia Experimental: colecta uma série de resultados e fornecer como estimativa a média destes resultados, juntamente com o seu erro padrão — obter várias estimativas do erro de previsão de um modelo $e = \{e_1, \dots, e_K\}$ de modo que seja possível calcular uma média amostral do erro de previsão $\bar{e} = \frac{1}{K} \sum_i e_i$ e o respetivo erro padrão desta estimativa, baseado no erro padrão amostral de e , S_E : $S_E(\bar{e}) = \frac{S_e}{\sqrt{K}} = \frac{\sqrt{\frac{1}{K-1} \sum_{i=1}^K (e_i - \bar{e})^2}}{\sqrt{K}}$

Método Holdout: divide aleatoriamente a amostra de dados disponível em dois subconjuntos: um para treinar o modelo e outro para o testar/avaliar (70/30) — preferível para amostras de dados muito grandes e não para amostras pequenas

Subamostragem Aleatória: repete o processo holdout várias vezes ao selecionar aleatoriamente as partições de treino e de teste

K-fold Cross Validation: K repetições de treino em parte dos dados e treino nos restantes

$$\hat{e}_{cv} = \bar{e} \pm SE(\bar{e})$$

• cada exemplo vai ser usado pelo menos uma vez para treino e outra para teste

K-fold Stratified Cross Validation: se for esperado que o algoritmo de aprendizagem seja sensível à distribuição da variável-alvo — Cada "fold" tem aproximadamente a mesma distribuição

Leave One Out Cross Validation: em cada iteração, é deixado um único caso de fora do conjunto de treino — n-fold CV

Método Bootstrap: treinar um modelo numa amostra aleatória de tamanho n com reposição do conjunto de dados original (de tamanho n) e testar o modelo nos casos que não foram usados nos dados de treino — repetir muitas vezes ²⁰⁰

Estimativa Bootstrap: média dos resultados das repetições — otimista por causa da sobreposição entre os casos de treino e de teste

Alternativa: $\hat{e}_{.632} = .632 \hat{e}_0 + .368 \hat{e}_n$ estimativa de resubstituição — otimista estimativa bootstrap leave-one-out — pessimista

• mais apropriado para conjuntos pequenos de dados

COMPARAÇÃO DE MODELOS

Modelos Comparáveis: têm de se referir aos mesmos conjuntos de treino e de teste

Objetivo: confirmar se um dado algoritmo tem melhor desempenho do que outro

Teste Estatístico da Hipótese Nula: Testa se algum resultado é improvável de ocorrido por sorte

Hipótese Nula H_0 : não há diferença entre um conjunto de modelos — a verdadeira diferença é 0 e quaisquer diferenças de desempenho são atribuídas ao acaso

• A hipótese nula é rejeitada se o resultado do teste de significância tem um valor- p menor do que um certo limite selecionado α para nível de significância

valor- p : probabilidade de observar uma diferença tão grande como a diferença amostral dada H_0

• Se $\text{valor-}p < \alpha$, então H_0 é rejeitada com confiança $(1 - \alpha)$

AMOSTRAS GRANDES

Teste-t Emparelhado: teste paramétrico para comparar duas amostras emparelhadas

Assunções:

1. os dados são emparelhados e provêm da mesma população
2. os dados são coletados de uma porção representativa e aleatória da população
3. a amostra é retirada de uma população com distribuição normal
4. o tamanho da amostra é razavelmente grande

Procedimento: para $\{m_{1i}, m_{2i}\}_{i=1}^N$

1. encontrar a diferença entre pares $\{d_i\}_{i=1}^N$
2. assume-se que a diferença entre duas variáveis normalmente distribuídas também o é
3. H_0 é que as diferenças têm média 0 e desvio padrão desconhecido
4. calcular o valor- p
5. se $\text{valor-}p < \alpha$, então H_0 é rejeitada — o desempenho dos modelos é estatisticamente diferente

AMOSTRAS PEQUENAS

Teste Wilcoxon Signed-Rank: teste não-paramétrico baseado em informações de ranking

- não considera as diferenças qualitativamente (não em magnitude absoluta) e não assume distribuição normal, pelo que os outliers têm menos efeito

Procedimento: para $\{m_{1,i}, m_{2,i}\}_{i=1}^N$

1. encontrar a diferença entre os pares $\{d_i\}_{i=1}^N$
2. registrar o sinal e o valor absoluto da diferença
3. ordenar as diferenças absolutas da menor para a maior
4. re-anexar os sinais das diferenças aos seus respetivos rankings

$$a. R^+ = \sum_{d_i > 0} \text{rank}(d_i) + \frac{1}{2} \sum_{d_i = 0} \text{rank}(d_i)$$

$$b. R^- = \sum_{d_i < 0} \text{rank}(d_i) + \frac{1}{2} \sum_{d_i = 0} \text{rank}(d_i)$$

5. tomar a menor das somas e encontrar o valor-p que rejeita H_0 para o nível α

Múltiplas Comparações em Múltiplas Tarefas:

1. Teste de Friedman: H_0 é todos os modelos são equivalentes e os rankings iguais
2. Se H_0 for rejeitada
 - a. Teste Nemenyi Post-Hoc: comparações emparelhadas entre todos os pares de modelos
 - b. Teste Bonferroni-Dunn Post-Hoc: comparações emparelhadas contra uma base-line

Diagrama de Diferença Crítica: mostra a classificação média de cada modelo e liga diferenças médias de classificações que não são estatisticamente significativas

Quão provável é que o modelo M_1 seja melhor do que M_2 por mais de 1%?

Teste de Sinal Bayesiano: compara dois modelos em múltiplos conjuntos de dados ao determinar a distribuição de probabilidade posterior das diferenças de desempenho entre dois modelos

Região de Equivalência Prática: região do espaço de densidade de probabilidade em que os dois classificadores são praticamente equivalentes

Aprendizagem em Domínios Desequilibrados

Assunções:

1. Os casos nos dados de treino não são uniformemente representados
2. Os casos sub-representados são os mais relevantes

Aprendizagem em Domínios Desequilibrados: a variável-alvo tem uma distribuição não-uniforme e os valores na domínio-alvo não são igualmente importantes, mas o foco está nos casos raros

Classificação: foco na classe minoritária

Regressão: foco em valores extremos

O modelo deve ser especialmente exato nos casos mais relevantes e sub-representados

As métricas padrão de desempenho são inadequadas porque assumem que todas as instâncias são igualmente relevantes para o desempenho do modelo

Para prevenir que o modelo seja encorajado para os casos mais frequentes, é necessário considerar métricas de desempenho encorajadas para o desempenho dos casos raros e estratégias de aprendizagem que se focam nesses casos raros

Curva PR: "trade-off" entre "recall" e precisão conforme varia o "threshold" de discriminação para as duas classes - é mais adequado porque não considera TN

AUC-PR: medida de desempenho que indica quão bom o modelo é a distinguir a classe-alvo (positiva)

Ex: G-Mean; MCC; Index of Balanced Accuracy

ESTRATÉGIAS DE APRENDIZAGEM

Estratégias de Pré-Processamento de Dados: alterar a distribuição dos dados para fazer com que o algoritmo padrão se focue nos casos raros e relevantes

- permitem a aplicação de qualquer algoritmo de aprendizagem, obtendo um modelo interpretável e encaixado em direção aos objetivos de aprendizagem
- é difícil mapear a distribuição dos dados numa nova distribuição ótima

1. Alteração da Distribuição: mudar a distribuição dos dados para endreçar o problema de baixa representatividade dos casos mais relevantes

- Sub-Amostragem Aleatória: remove exemplares da classe maioria ou valores comuns dos dados originais, reduzindo o seu tamanho — pode descartar exemplares úteis
 - Sobre-Amostragem Aleatória: adiciona um conjunto de cópias da classe minoritária ou exemplares de valores raros aos dados — possível "overfitting"
 - SMOTE: sobre-amostrar os exemplares da classe minoritária ao gerar novos dados sintéticos, criando novos exemplares ao interolar um exemplo minoritário e um dos seus K vizinhos mais próximos da classe minoritária — reduz os riscos
2. Pesar o Espaço de Dados: atribuir pesos diferentes a diferentes instâncias de dados
— risco de "overfitting" e indisponibilidade de valores reais de custo

Estratégias de Propósito Especial: alterar o algoritmo de aprendizagem para que possa aprender a partir de dados desequilibrados

- incorpora preferências do domínio como critério, e gera modelos interpretáveis
- aplicação resulta, obriga a re-aprendizagem/adaptação e é difícil mapear

Estratégias de Pós-Processamento Preditivo: manipular as previsões do modelo de acordo com as preferências do domínio — "thresholding"; "cost-thresholding"

- dados originais e algoritmo padrão, não necessidade de re-aprendizagem
- os modelos não refletem as preferências do domínio e são menos interpretáveis

Deteção de Anomalias

Outlier: observação que se desvia tanto das outras observações que levanta a suspeita de que foi gerada por um mecanismo diferente

Outliers: pontos de dados individuais que são diferentes dos dados restantes

Clusters: grupos de pontos de dados que são similares

Outlier: associado com ruído

Anomalia: associado com dados incomuns cuja causa deve ser investigada

Uma anomalia pode ser considerada um outlier, mas um outlier não é necessariamente uma anomalia

Tipos de Outliers:

1. Ponto: instância que individualmente ou em grupos pequenos é diferente das restantes
2. Contexto: instância que quando considerada num contexto é diferente das restantes
3. Coletivo: instância que individualmente pode não ser um outlier, mas inspecionado em conjunção com instâncias relacionadas e em relação a todos os dados é-o

Attribui uma Etiqueta: identificação de instância normal ou outlier

Attribui uma Pontuação: probabilidade de ser um outlier

Deteção Não-Supervisionada de Outliers: o conjunto de dados não tem informação sobre o comportamento de cada instância — assume que instâncias com comportamento normal são as mais frequentes

Deteção Semi-Supervisionada de Outliers: o conjunto de dados tem poucas instâncias de comportamento normal ou outlier

ABORDAGENS DE DETEÇÃO NÃO-SUPERVISIONADA DE OUTLIERS

Deteção de Outliers Baseada em Estatísticas: todos os pontos que satisfazem um teste de discordância estatística para algum modelo estatístico são declarados outliers

- solução justificável se as assumções forem verdadeiras; intervalo de confiança
- os dados nem sempre seguem um modelo estatístico
- escolher as melhores hipóteses de teste não é fácil
- capturar interações entre atributos nem sempre é possível
- estimar parâmetros para modelos estatísticos é difícil

Deteção de Outliers Baseada em Proximidade: instâncias normais ocorrem em vizinhanças densas, enquanto outliers ocorrem longe das suas vizinhanças mais próximas

- fundamentalmente guiado pelos dados; não faz assumções sobre a distribuição dos dados
- pode ser difícil distinguir outliers de regiões ruidosas pouco densas
- tem de combinar análise local e global
- o contraste nas distâncias perde-se com a dimensionalidade
- teste computacionalmente complexo

Deteção de Outliers Baseada em Agrupamento: instâncias normais pertencem a grupos grandes e densos, enquanto instâncias outliers não pertencem a grupos, estão longe do grupo mais próximo e/ou formam grupos pequenos de baixa densidade

- facilmente adaptável; teste rápido
- treino computacionalmente dispensável
- pode falhar se pontos normais não criarem grupos
- espaços muito dimensionais podem não ter grupos significativos
- técnicas não otimizadas

iForest: instâncias com atributos-valores distinguíveis não mais prováveis de serem separadas no partitionamento inicial

Objetivo: isolar explicitamente pontos anômalos

Isolamento: separar uma instância do resto das instâncias

0. Parâmetros: número de árvores e tamanho da sub-amostragem

1. Treino: construir um conjunto de árvores de decisão binárias aleatórias induzidas pelos dados (árvores de isolamento) usando sub-amostras do conjunto de treino dado

2. Avaliação: passa instâncias de teste através de árvores de isolamento para obter uma pontuação de outlier para cada instância — a pontuação é relacionada com o comprimento médio do caminho (outliers perto da raiz; pontos normais profundos)

• nem medidas de distância/densidade

• elimina o custo computacional do cálculo de distâncias

• escala bem para conjuntos de dados grandes e com muitas dimensões

• hiper-parâmetros têm de ser afinados

• aleatoriedade — resultados diferentes em execuções diferentes

• amostras grandes podem causar "masking" ou "swamping"

ABORDAGENS DE DETEÇÃO SEMI-SUPERVISIONADA DE OUTLIERS

Proposta: constrói um modelo de previsão para o comportamento normal e classifica quaisquer desvios a este comportamento como outliers

SVM One-Class: obtém uma fronteira esférica, no espaço de funcionalidades, à volta dos dados normais, cujo volume da hiper-esfera deve ser minimizado para minimizar o efeito de incorporar outliers na solução, incluindo só os pontos de treino — qualquer ponto que resida fora da hiper-esfera é um outlier

Redes Neuronais Auto-Associativas: rede baseada num percepção treinada só com dados normais — a rede recria com ruidos dados normais e gera erro alto para outliers

• modelos interpretáveis; aprendizagem exata do comportamento; pode detectar raros outliers
• requer instâncias pré-etiquetadas como normais; possível alto rácio de falsos alarmes

DETEÇÃO DE OUTLIERS CONTEXUAIS

Atributos Contextuais: usados para determinar o contexto (ou vizinhança) da instância - sequencial; espacial; gráfico

Atributos Comportamentais: definem características não-contextuais da instância

- O comportamento do outlier é determinado usando os valores para os atributos comportamentais num contexto específico

Redução a Detecção de Pontos Outliers: segmentar dados usando atributos contextuais e aplicar detecção de pontos outliers em cada contexto usando atributos comportamentais
Utilizar Estrutura nos Dados: construir modelos a partir dos dados usando atributos contextuais para prever o comportamento esperado em relação a um dado contexto

- permite definições naturais de outliers
- deteta outliers difíceis de detectar globalmente
- é difícil identificar um conjunto de bons atributos contextuais
- assume que todas as instâncias normais vão ser comportamentalmente semelhantes

DETEÇÃO DE OUTLIERS COLETIVOS

Um outlier coletivo pode também ser um outlier contextual se analisado em relação a um contexto  Um problema de detecção de outliers coletivos pode ser transformado num problema de detecção de outliers contextuais ao incorporar informação de contexto

- permite definições naturais de outliers
- estruturas não explicitamente definidas; necessidade de extraer funcionalidades; dependência aplicacional; custo computacional

Modelos de Ensemble

→ Abordagem de Ensemble

Ensembles: coleções de modelos que são usados juntos para entregar um certo problema de previsão

↓ Não há um algoritmo globalmente melhor
MAS

- Um conjunto de classificadores é melhor do que classificadores individuais se:
 1. tiverem melhor desempenho do que classificadores aleatórios
 2. tiverem erros não correlacionados
 3. cometerem erros em diferentes regiões do espaço

Como alcançar diversidade?

1. Combinar outputs de formas diferentes
 - a. Regulação: média / soma (pesada)
 - b. Classificação: voto (pesada)
2. Gera modelos diferentes
 - a. Homogêneos: único algoritmo de indução
 - b. Heterogêneos: múltiplos algoritmos de indução
3. Em Ensembles Homogêneos:
 - a. perturbar o conjunto de exemplos de treino
 - b. perturbar o conjunto de atributos
 - c. escolher parâmetros diferentes para o algoritmo de indução
 - d. usar variações do algoritmo de indução

• Os modelos de ensemble podem atuar numa ou em ambas as componentes do erro

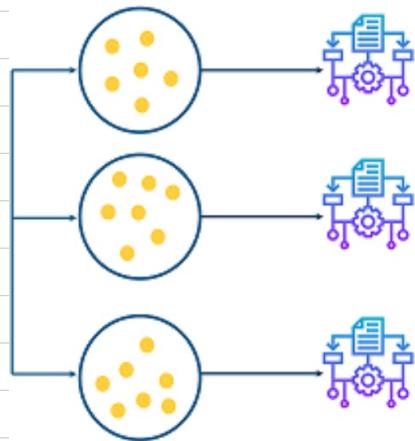
TIPOS DE ENSEMBLES

Modelos Independentes/Paralelos: construir os modelos independentemente de forma a garantir alguma diversidade entre eles - "bagging"

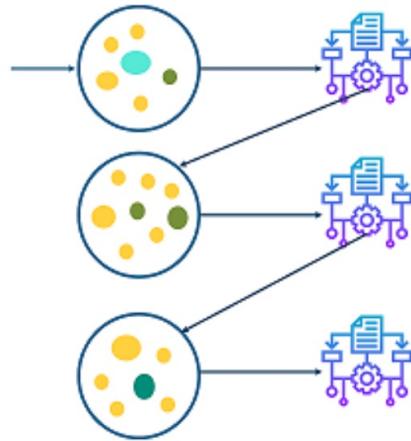
Como alcançar diversidade?

1. conjuntos de treino diferentes
2. preditores diferentes

Modelos Coordenados/Sequenciais: construir um modelo "maior" ao compô-lo de modelos menores e integrados, com uma participação pesada - "boosting"



Bagging - Parallel



Boosting - Sequential

	BAGGING	BOOSTING
AMOSTRAGEM	independente	dependente do erro
AGREGAÇÃO	uniforme	pesada
REDUÇÃO DO ERR	variância	ríss + variância

MODELOS INDEPENDENTES

Bagging: obter um conjunto de K modelos usando diferentes amostras com reposição dos dados de treino — para cada modelo, uma pequena proporção dos exemplos vai ser diferente

- A diversidade entre os K modelos é garantida se os "base learners" tiverem alta variância — algoritmos instáveis e sensíveis a pequenas perturbações
- Fácil de implementar e paralelizar
- O erro diminui devido à redução da componente de variância

Random Forest: conjunto de modelos baseados em árvores em que cada árvore é obtida a partir de uma amostra com reposição dos dados originais e usa seleção aleatória de variáveis durante o crescimento da árvore — gerar amostras de variáveis

Fase de Aprendizagem: desde $t=1$ até T (número de árvores)

1. Retirar uma amostra aleatória com substituição do conjunto de treino D_t
2. Treinar um modelo de árvore $h_t(x)$ em D_t sem podar
3. Em cada separação candidata, usar um subconjunto aleatório de m funcionalidades

Return: $\{h_t(x) \mid 1 \leq t \leq T\}$

Fase de Previsão: prever a classe obtida por voto majoritário ou o valor ao calcular a média do output de cada árvore

"Variable Importance": quais variáveis têm o maior poder preditivo?

→ quanto diminui a exactidão ou cresce o erro médio quadrado quando uma variável é excluída?

→ quanto decresce a impureza quando a variável é escolhida para dividir num nó?

1000...5000

FP

Hiper-Pâmetros: número de árvores & número de atributos a selecionar em cada nó

afinização simples e árvores simples

⇒ menor interpreabilidade do que árvores

MODELOS COORDENADOS

Boosting: criam iterativamente um "strong learner" ao adicionar, em cada iteração, um novo "weak learner" para fazer o "ensemble" — os "weak learners" são adicionados com pesos que refletem o seu poder preditivo

"Weak Learner": modelo que sozinho é incapaz de aproximar corretamente a função de previsão desconhecida

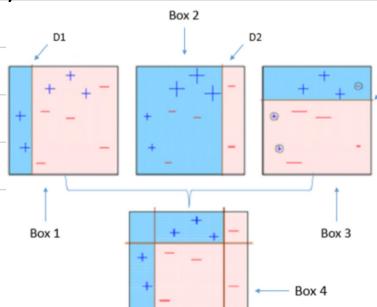
Previsão: voto/média ponderada de cada "learner"

Depois de cada adição, os dados não re-pesados de maneira que os caros ainda mal previstos ganhem mais peso (indicando a probabilidade do exemplo ser selecionado numa amostra uniforme), para que cada novo "weak learner" se focue mais nos erros dos anteriores!

AdaBoost: processo iterativo (novos modelos não adicionados para formar um ensemble) e adaptativo (em cada nova iteração do algoritmo, os novos modelos não construídos para tentar superar os erros cometidos em iterações anteriores) — em cada iteração os pesos dos caros de heino não ajustados para que os caros erradamente previstos tenham os seus pesos aumentados para fazer com que os novos modelos se focuem em prever-lhos exatamente

→ O peso do "weak model" t é $\alpha_t = \frac{1}{2} \ln \left(\frac{1 - e_t}{e_t} \right)$

Hiper-Parâmetro: número de iterações



Gradient Boosting Machine (GBM): sequencial - não ajusta os pesos dos exemplos em cada iteração, mas treina o novo "learner" para os erros residuais feitos pelo "learner" anterior de modo que o atual seja sempre mais efetivo do que o anterior

Objetivo: minimizar a função de perda ao adicionar "weak learners" usando descida de gradiente

Aprendizagem: construir um modelo aditivo ao adicionar novas árvores para complementar os existentes
minimizar o termo de regularização $\sum_{i=1}^k R(h_i)$ - complexidade das árvores

Previsão: a resposta é a combinação linear ótima de todas as árvores de decisão

Hiper-Parâmetros:

1. rácio de aprendizagem - fator multiplicativo nos erros das árvores subsequentes
↳ quanto mais lento, mais robusto e menos risco de overfitting
2. número de árvores
↳ quanto mais árvores, maior risco de overfitting

Se o rácio de aprendizagem for baixo, não necessárias mais árvores para treinar

Extreme Gradient Boosting (XGB): otimização escalável de GBM

- ↳ penalização inteligente das árvores
- ↳ parâmetro de aleatoriedade para reduzir a correlação entre árvores
- ...

"Feature Importance": quão útil foi cada funcionalidade na construção das árvores "boosted"?
↳ a importância é calculada para uma árvore de decisão única pelo número de vezes que a funcionalidade é selecionada para decisão, pesada pela melhora do modelo em resultado de cada decisão

Support Vector Machines

→ Abordagem de Optimização

Uma fronteira de decisão não-linear no espaço original de funcionalidades X pode ser uma fronteira de decisão linear no espaço estendido de funcionalidades de X : X^2

SVM Linear: dado um conjunto de dados $D = \{ \langle x_i, y_i \rangle \}_{i=1}^N$ em que x_i é um vetor de funcionalidades e $y_i \in Y$ é o valor da variável nominal em $\{-1, +1\}$, cada vetor de funcionalidades x_i é um ponto num espaço multi-dimensional e D é linearmente separável. Existe um hiperplano $h(x) = w \cdot x + b$ que divide o espaço de input tal que

$$g(x) = \operatorname{sgn}(h(x)) = \begin{cases} +1 & \text{se } w \cdot x + b > 0 \\ -1 & \text{se } w \cdot x + b < 0 \end{cases}$$

Objetivo: encontrar a fronteira de decisão que maximiza a margem — o hiper-plano que separa os exemplos das duas classes com a margem máxima generaliza melhor e garante uma melhor execução em dados não observados

Vetores de Suporte: todos os casos que caem nos hiper-planos $H_1: g(x) = w \cdot x + b = +1$ e $H_2: g(x) = w \cdot x + b = -1$ — remover os outros casos não altera a solução

SVM "Hard Margin": funcionam bem em dados linearmente separáveis — não consideram ruídos

SVM "Soft Margin": toleram alguns erros de classificação ("slack variables") para aumentar o tamanho da margem de separação para que os outros pontos possam ser classificados corretamente

Termo de Regularização C: trade-off entre maximizar a margem e minimizar os erros de classificação - C é o custo dos erros de classificação

SVM Não-Linéar: o espaço de input x é mapeado para um espaço de funcionalidades $\phi(x)$ em que as classes são linearmente separáveis

Função Kernel: recebe como input vetores no espaço original e retorna o produto escalar desses vetores no espaço de funcionalidades, nem sempre necessárias as coordenadas dos dados no espaço de funcionalidades!

• Como $K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j)$, em vez de se calcularem os produtos escalares num espaço multi-dimensional, realizam-se operações simples e eficientes no espaço original (nem transformação de funcionalidades)

Teorema de Mercer: qualquer função semi-positiva simétrica é uma Kernel

1. Linear - $K(x_i, x_j) = x_i \cdot x_j$ 2. Polinomial 3. Gausiana 4. Sigmoidal

Como lidar com mais de duas classes?

Resolver várias tarefas de classificação binária, encontrando os vetores de suporte que separam cada classe de todas as outras - um classificador SVM para cada classe

Regressão: encontram a função linear $f(x)$ que aproxima os casos de treino com uma precisão de ϵ - uso da função perda $|\xi|_\epsilon$

:- fundamentos teóricos fortes; solução esparsa; overfitting controlado por "soft margin"
:- a complexidade não depende da dimensionalidade; problema de otimização simples

:- muito susível a hiper-parâmetros; complexidade alta; modelo "black-box"

Redes Neuronais

→ Abordagem de Optimização

Rede Neuronal Artificial: conjunto de unidades (neurônios) conectados em que cada conexão tem um peso associado e cada unidade tem um nível de ativação, bem como meios de atualizar esse nível — aprender é atualizar os pesos das conexões

• Cada unidade tem uma função muito simples que recebe os impulsos de input e calcula o output como uma função desses impulsos — uma combinação linear dos inputs e uma função de ativação (não-linear) $a_i = g(i_{in,i})$

Perceptrão: rede com uma camada de input e uma camada de output
↳ limitado para funções linearmente separáveis — $w_i(t+1) = w_i(t) + \eta(\text{real-previs})x_i$

Redes "Feed-Forward": redes com conexões unidirecionais (desde o input para o output) e sem ciclos em que cada unidade só está conectada a unidades na camada seguinte

Redes Recorrentes: redes com conexões arbitrárias — instáveis; caóticas; convergência lenta

Função de Ativação: usada para determinar o output de cada nó da rede neuronal

Algoritmo de "Backpropagation": como cada unidade é responsável por uma certa fração do erro nos nós de output aos quais está conectada, então o erro é dividido de acordo com o peso da conexão entre as respectivas unidades ocultas e de output, propagando-se para trás — computa o gradiente no espaço de pesos de uma rede neuronal "feed-forward", em relação a uma função de perda

1. Inicializar pesos
2. Para cada exemplo de treino:
 - a. prever o output
 - b. calcular o erro de previsão
 - c. propagar para trás
 - d. atualizar pesos
3. Até convergir — todos os exemplos classificados corretamente ou critério de paragem

Descida de Gradiente Estocástica: atualizar os pesos um exemplo de cada vez, em vez de calcular o gradiente da função de erro completa.

Descida do Gradiente por "Batch": o tamanho da batch é o número de sub-amostras dadas à rede depois das quais acontece a atualização dos pesos

Quando parar de treinar? : cedo \rightarrow não treinada ; tarde \rightarrow overfitting

Critério de Paragem: número máximo de iterações OU erro mínimo

Número de Nós Pártios: poucos \rightarrow underfitting ; muitos \rightarrow overfitting

Rácio de Aprendizagem: define o tamanho dos passos para obter a direção da descida máxima — baixo \rightarrow lento ; alto \rightarrow não-convergência

- dados devem ser padronizados
- valores em falta (0) não influenciam
- o rácio de aprendizagem deve diminuir

- tolerante a ruído; classifica novos padrões; necessita paralelização
- treino lento; "black-box"

Fraud Detection (CC4036)

Test

2022/2023
DCC - FCUP

Name: _____ Student Nr: _____

-
- Duration: 1h.
 - Multiple choice questions
 - **Mark your answers with a circle.**
 - In case of a mistake, cancel the answer with a cross and do a circle over the new answer.
 - Each correct answer scores 0.5.
 - Incorrect answers will decrease your score in 0.25
 - This is an individual exam. Any attempt to communicate with a third party is regarded as fraud. Cell phones or other communicating devices are forbidden, as well as access to the internet.

1. In a project, two goals were proposed: 1. predict whether a loan application is going to be successfully paid or not; 2. better understand customers. The direct approach(es) to address this(goal)s are:

agrupamento

- ✓ (a) classification for goal 1 and clustering for goal 2.
 (b) regression for goal 1 and clustering for goal 2.
 (c) clustering for goal 1 and classification for goal 2.
 (d) classification for both.

2. It is common to reduce the dimensionality of the data. This operation aims to:

- ✓ (b) decrease the number of variables that describe the examples.
 (a) remove the number of examples and variables with missing values.
 (c) decrease the number of examples.
 (d) take a smaller sample of the data.

3. The MAE (mean absolute error) measure is easier to read than the MSE (mean squared error) measure because:

- ✓ (a) the scale of values is the same as the dependent variable.
 (b) the scale of values is not the same as the dependent variables.
 (c) places more emphasis on larger errors.
 (d) is represented on a scale that is not [-1,1].

✓ 4. Which of the following cannot be varied in the generation of homogeneous ensembles:

- (a) algorithm.
- (b) data.
- (c) hyperparameter values.
- (d) model.

✓ 5. Linear regression is:

- (a) not very sensitive to overfitting due to the linear nature of the model.
- (b) not very sensitive to overfitting due to the high variance of the models.
- (c) very sensitive to overfitting due to the linear nature of the model.
- (d) very sensitive to overfitting due to the high variance of the models.

✓ 6. The single link proximity measure can be used in:

- (a) DBSCAN clustering algorithm.
- (b) hierarchical clustering algorithms.
- (c) k-means algorithm.
- (d) in any clustering algorithm.

✓ 7. In Principal Component Analysis (PCA), each component is a:

- (a) cluster.
- (b) visualization of the data set.
- (c) linear or a non-linear combination of the original attributes.
- (d) linear combination of the original attributes.

✓ 8. Each iteration of the k-Means algorithm performs the following operation:

- (a) choose the number of clusters.
- (b) find the k-nearest neighbours.
- (c) allocate each observation to the cluster with its k-nearest centroid.
- (d) allocate each observation to the cluster with the closest centroid.

✓ 9. Feature selection methods used before modeling:

- (a) can eliminate both redundant and variable variables without useful information for the model.
- (b) eliminate only redundant variables.
- (c) eliminate only variables with no useful information for the model.
- (d) eliminate redundant or variable variables with no useful information for the model, but the same method never eliminates both types of variables. WRAPPER



10. Feature extraction is the process that:

- (a) removes noise and outliers.
- (b) reduces data dimensionality. *SELECTION*
- (c) incorporates domain knowledge to create new features. *ENGINEERING*
- (d) obtains features from raw data.



11. The difference between classification and regression is:

- (a) in the dependent variables, implying the use of different evaluation measures.
- (b) in the independent variables, implying the use of different evaluation measures.
- (c) in all variables (independent and dependent), although it is possible to use the same evaluation measures in both cases.
- (d) in the dependent variables, although it is possible to use the same evaluation measures in both cases.



12. To measure the location of a variable with extreme values, one should use the:

- (a) standard deviation.
- (b) inter-quartile range.
- (c) mean.
- (d) median.



13. Suppose we have the following set of values for two variables

$$(X, Y) = \{(1, 50), (2, 40), (3, 32), (4, 24), (5, 18), (6, 12), (7, 8), (8, 4), (9, 2), (10, 0)\}$$

and we want to measure how correlated they are. We can say that:

- (a) the Pearson correlation coefficient will indicate a stronger correlation than the Spearman rank correlation coefficient.
- (b) the Pearson correlation coefficient will indicate a weaker correlation than the Spearman rank correlation coefficient.
- (c) the Pearson and the Spearman rank correlation coefficients will indicate the same measure of correlation.
- (d) the Pearson correlation coefficient will indicate that there is no correlation; only the Spearman rank correlation coefficient will indicate that.



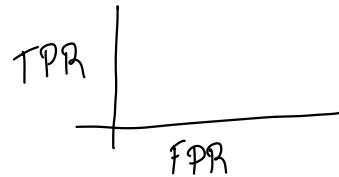
14. Extreme Gradient Boosting (XGBoost) is a:

- (a) variance reduction method.
- (b) bias reduction method.
- (c) bias and variance reduction method.
- (d) neither bias nor variance reduction method.



15. The Area Under Curve (AUC) is an evaluation metric that:

- (a) gives equal importance to positive and negative classes.
- (b) combines in a single measure the true and false positive ratios.
- (c) is able to substitute the confusion matrix.
- (d) evaluates the accuracy.



16. The k value in the k-nearest neighbours algorithm should be selected taking into account the following:

- (a) If $k=1$, it is more sensitive to noise in the labels (i.e. incorrectly labelled examples).
- (b) If k is large (e.g. $k \geq 100$), it is more sensitive to noise in the labels (i.e. incorrectly labelled examples).
- (c) If $k=1$, it is more sensitive to the number of attributes.
- (d) If k is large (e.g. $k \geq 100$), it is more sensitive to the number of attributes.