Fundamentos de Análise de Dados
Semestre 2021-1
– Trabalhos –

Thelmo de Araujo

abril de 2021

1 Avaliação de Aprendizado: Trabalhos

O aprendizado dos alunos será **parcialmente** avaliado por meio de dois trabalhos: o Trabalho Parcial e o Trabalho Final.

A nota do Trabalho Parcial corresponderá a **40% da Média Final**.

A nota do Trabalho Final corresponderá a **60% da Média Final**.

O **Trabalho Parcial** envolverá os conceitos de ortogonalidade: projeções ortogonais, ortonormalização de Gram-Schmidt e decomposição QR. No **Trabalho Final**, será acrescentada a Teoria Espectral: decomposição espectral e a Decomposição de Valor Singular (SVD).

O Trabalho Final deverá conter todo o Trabalho Parcial, abrangendo assim todos os conceitos estudados na disciplina.

Para aplicar tais conceitos, três temas podem ser desenvolvidos:

- Regressão, que pode ser realizada pelo método dos mínimos quadrados e, posteriormente, pela técnica das componentes principais (ver *Principal Component Regression* em [Jolliffe, 2010]).

- Seleção de características (originais) para classificação, utilizando a ortonormalização de Gram-Schmidt como proposto por Stoppiglia et alli [Stoppiglia et al., 2003], e depois usando a similaridade das características originais com os autovetores da matriz de covariância dos dados, como proposto por Dunteman [Dunteman, 1989].

- Redução de dimensionalidade para classificação, utilizando, inicialmente, a decomposição espectral e, posteriormente, a SVD. Aqui, a primeira etapa deve enfatizar o caráter de projeção ortogonal em subespaços gerados pelos autovetores (que podem ser calculados, numa primeira etapa, como provenientes de uma caixa-preta). A fundamentação teórica da decomposição espectral e da SVD ficaria assim para a segunda etapa.

Cada um dos três temas acima deve ser aplicado a cinco bancos de dados, um para cada aluno. Haverá então trabalhos para até 15 alunos.

Em ambas as etapas, isto é, tanto no Trabalho Parcial quanto no Trabalho Final, cada aluno deverá produzir:

1. Um **Jupyter Notebook** com o código em linguagem Python para a solução do problema computacional proposto.
2. Uma **apresentação** utilizando slides (em LaTeX ou Word).
3. Um **relatório escrito**, constando de:
   - Seções típicas de uma artigo ou relatório científico: Resumo, Introdução, Fundamentação Teórica, Metodologia e Experimentos, Resultados, Conclusão, Trabalhos Futuros e Referências Bibliográficas.
   - Questões e suas soluções, compostas geralmente de demonstrações formais (caso geral) de pequenas proposições e verificações numéricas dessas proposições aplicadas ao banco de dados correspondente ao aluno.

As questões teóricas devem constar explicitamente no relatório. Por exemplo:

> **Questão 16.** Mostre que as matrizes gramianas são simétricas.
> **Solução.** Seja A uma matriz gramiana, então existe uma matriz B tal que
>
> $$A = B^T B\,.$$
>
> Assim,
>
> $$A^T = (B^T B)^T = B^T (B^T)^T = B^T B = A\,.$$
>
> Portanto, A é simétrica. □

As questões que envolvem cálculos numéricos devem ser explicitadas no Relatório e em células próprias no Jupyter Notebook. Numa célula do tipo *markdown* deve constar o enunciado da questão, noutra célula *markdown*, o aluno deve explicar sua solução. Por fim, a verificação numérica deve estar numa única célula de *código*.

2 Jupyter Notebook

Cada aluno desenvolverá o trabalho no Jupyter Notebook em linguagem Python.

Esta disciplina requer a participação ativa de cada estudante, que deve estar constantemente com "a matéria em dia".

3 Relatórios Técnicos Parcial e Final

Cada aluno deverá entregar o Relatório Técnico Parcial no dia de sua apresentação do Trabalho Parcial e o Relatório Técnico Final no dia de sua apresentação do Trabalho Final. Os dias das apresentações serão divulgados durante o semestre.

Os Relatórios Técnicos deverão ter as seguintes características:

- Ser escrito em LaTeX com a fonte padrão em tamanho 12.
- Conter título, nome da "problema" (por exemplo, Classificação - Written Numbers MNIST) e nome completo do aluno.
- Conter as seguintes seções: Resumo, Introdução e Trabalhos Relacionados, Fundamentação Teórica, Metodologia e Experimentos, Resultados, Conclusão e Trabalhos Futuros, Referências Bibliográficas.
- As fórmulas matemáticas e as tabelas deverão ser escritas em LaTeX (é claro). Não serão aceitos relatórios com fórmulas ou tabelas "coladas" como imagens.
- Teoremas importantes deverão ser enunciados no ambiente LaTeX apropriado (`\begin{theorem} ... \end{theorem}`).
- Proposições e teoremas "menores" deverão ser **demonstrados**.

No Relatório Técnico, serão considerados aspectos de clareza, completude, correção da linguagem, entre outros.

Juntamente com o Relatório Final (arquivos .pdf e .tex), o aluno deverá anexar uma pasta (denominada `tex`) com todos os arquivos necessários para a compilação do arquivo fonte .tex, bem como uma pasta (denominada `python`) com o código fonte do programa em Python.

4 Apresentações Parcial e Final

Nas **últimas semanas de maio** e nas **últimas semanas do semestre letivo**, os alunos farão as respectivas apresentações de seus Trabalhos Parciais e Finais. As datas serão agendadas pelo professor no tempo apropriado.

O aluno deverá apresentar seu trabalho num período de 15 (mínimo) a 20 (máximo) minutos. Excedendo os 20 minutos, o professor interromperá a apresentação.

As Apresentações Parcial e Final deverão ser feitas com slides (de preferência em LaTeX).

Nas Apresentações Parcial e Final, serão considerados aspectos de clareza, completude, correção da linguagem, entre outros. É importante que a apresentação concentre-se nos apectos relevantes para o entendimento do

problema e sua solução; detalhes devem ser deixados para o Relatório Técnico.

5 Trabalhos Parcial e Final: Temas

As subseções a seguir descrevem as tarefas e as questões a serem desenvolvidas em cada um dos três temas descritos na Seção 1: inicialmente, no Trabalho Parcial e, por fim, no Trabalho Final.

5.1 Regressão: O Método dos Mínimos Quadrados

O Método dos Mínimos Quadrados, como vimos, consiste em encontrar uma solução aproximada, $\tilde{\mathbf{x}}$, de um sistema linear inconsistente

$$A\mathbf{x} = \mathbf{b}\,.$$

Isso é feito projetando o vetor $\mathbf{b}$ no espaço-nulo esquerdo de A e resolvendo as equações normais

$$A^T A\tilde{\mathbf{x}} = A^T\mathbf{b}\,.$$

Pode-se substituir A por sua decomposição QR, $A = QR$, e obter novas equações normais:

$$R\tilde{\mathbf{x}} = Q^T\mathbf{b}\,, \tag{1}$$

sob determinadas condições em A.

Tarefas

As seguintes tarefas devem ser realizadas como Trabalho Parcial e incluídas nos Relatórios Parcial e Final:

1. Dividir o banco de dados em conjunto de treinamento e conjunto de teste (proporções comumente utilizadas são 60%/40% e 70%/30%.)
2. Calcular a decomposição QR da matriz de dados de **treinamento** X, isto é
$$X = QR\,.$$
3. Resolver as equações normais em (1) para X.
4. Aplicar os coeficientes de regressão para o conjunto de teste e comparar o resultado com os valores reais de teste utilizando um gráfico: os valores reais no eixo das abcissas, e os valores estimados no eixo das ordenadas. Acrescentar a reta a 45° (estimativa ideal) para facilitar a visualização.
5. Calcular a raiz quadrada do erro quadrático médio (RMSE).

Questões

Os enunciados e as soluções das questões abaixo devem estar nos Relatórios Parcial e Final.

1. Mostrar que, usando a decomposição QR de A (isto é, $A = QR$), a equação normal $A^T A\tilde{\mathbf{x}} = A^T\mathbf{b}$ pode ser escrita como $R\tilde{\mathbf{x}} = Q^T\mathbf{b}$.

2. Qual é a condição sobre a matriz de dados A, no item anterior, para que a matriz R seja invertível? Demonstrar sua afirmação.

3. Verificar numericamente que $Q^TQ = I$, para o respectivo banco de dados.

5.2 Regressão: Componentes Principais

Uma outra maneira de realizar a regressão de dados é utilizando as componentes principais (ver *Principal Component Regression*, em [Jolliffe, 2010]).

Partindo da matriz de dados de treinamento *centralizados* X, com m linhas (amostras) e n colunas (características), calcula-se a matriz de covariância (a menos de um fator $1/m$):

$$\mathrm{cov}(X) = X^T X\,.$$

Pode-se usar a função `np.linalg.eigh()` do NumPy para calcular a decomposição espectral de X^TX:

$$X^T X = Q\Lambda Q^T\,, \tag{2}$$

ou seja:

$$\mathbf{w}, Q = \texttt{np.linalg.eigh}(X^T X)\,,$$

obtendo-se os autovalores $\mathbf{w}$ e seus autovetores associados nas colunas de Q. **Atenção**: os autovalores (e seus autovetores associados) devem ser ordenados em ordem **decrescente**.

Usando a matriz Q em (2), consideremos a matriz Z das coordenadas das amostras em relação à base de autovetores dada por:

$$X^T = QZ^T\,. \tag{3}$$

Consideremos ainda o sistema de equações possivelmente inconsistente:

$$X\boldsymbol{\beta} = \mathbf{y}\,,$$

sendo $\mathbf{y}$ o vetor com os valores reais provenientes do banco de dados.

A mudança de variável $X = ZQ^T$ nos dá o sistema equivalente:

$$Z\boldsymbol{\gamma} = Z(Q^T\boldsymbol{\beta}) = (ZQ^T)\boldsymbol{\beta} = X\boldsymbol{\beta} = \mathbf{y}\,, \tag{4}$$

definindo $\boldsymbol{\gamma} = Q^T \boldsymbol{\beta}$. A equação normal do sistema em (4) é

$$Z^T Z \tilde{\boldsymbol{\gamma}} = Z^T \mathbf{y}\,, \tag{5}$$

cuja solução é dada por:

$$\tilde{\boldsymbol{\gamma}} = (Z^T Z)^{-1} Z^T \mathbf{y}\,. \tag{6}$$

Tarefas

As seguintes tarefas devem ser realizadas como Trabalho Final e incluídas no Relatório Final:

1. Dividir o banco de dados em conjunto de treinamento e conjunto de teste (proporções comumente utilizadas são 60%/40% e 70%/30%.)

2. Calcular $\tilde{\boldsymbol{\gamma}}$, usando a Equação (6), para o respectivo banco de dados, resolvendo assim o problema de regressão. Mostrar o erro (RMSE) e o gráfico como na Subseção 5.1.

3. Comparar os erros de regressão obtidos pelo método de mínimos quadrados usando a decomposição QR e usando a decomposição espectral.

4. Consideremos $\hat{Q}$ a matriz cujas colunas são as r primeiras colunas de Q e $\hat{Z}$ a matriz das coordenadas dos dados em relação à base do subespaço gerados pelas r primeiras colunas de Q. Podemos **definir**:

   $$\hat{X}^T = \hat{Q}\hat{Z}^T\,. \tag{7}$$

   Resolver o problema de regressão $Z\boldsymbol{\gamma} = \mathbf{y}$ calculando

   $$\tilde{\boldsymbol{\gamma}} = \hat{\Lambda}^{-1}\hat{Z}^T\mathbf{y}\,.$$

   com o respectivo banco de dados. A matriz $\hat{\Lambda}$ é dada pela Equação (8), mostrada abaixo. Calcular e erro de regressão e compará-lo com os erros de regressão obtidos anteriormente.

5. Analisar os erros de regressão para diversos valores de r.

Questões

Os enunciados e as soluções das questões abaixo devem estar no Relatório Final.

1. A partir da Equação (3), mostrar que

   $$Z = XQ\,.$$

2. Mostrar que

$$Z^T Z = \Lambda^{-1}\,.$$

3. Calcular, numericamente e de maneira computacionalmente eficiente (isto é, com ordem $\mathcal{O}(n)$), Λ^{-1}, para o respectivo banco de dados.

4. Mostrar que

$$\boldsymbol{\beta} = Q\boldsymbol{\gamma}\,.$$

5. Mostrar que a matriz $r \times r$ definida por

$$\hat{\Lambda} = \hat{Z}^T \hat{Z}$$

é a matriz diagonal

$$\hat{\Lambda} = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_r \end{bmatrix}, \tag{8}$$

sendo $\lambda_1, \dots, \lambda_r$ os r primeiros elementos da diagonal principal da matriz Λ, dada na Equação (2).

6. Mostrar que

$$\hat{X} = \hat{Q}\hat{\Lambda}\hat{Q}^T\,.$$

7. Considerando a versão reduzida da equação normal em (5):

$$\hat{Z}^T \hat{Z}\tilde{\boldsymbol{\gamma}} = \hat{Z}^T \mathbf{y}\,,$$

mostrar que a solução de mínimos quadrados é dada por

$$\tilde{\boldsymbol{\gamma}} = \hat{\Lambda}^{-1}\hat{Z}^T \mathbf{y}\,.$$

8. Mostrar que $\boldsymbol{\gamma} = \hat{Q}^T\boldsymbol{\beta}$ **não** implica $\boldsymbol{\beta} = \hat{Q}\boldsymbol{\gamma}$.

5.3 Classificação: Projeções Ortogonais

Antes do estudo formal da decomposição espectral, os estudantes cujos trabalhos serão sobre o problema de classificação devem mostrar diversas propriedades das matrizes ortogonais e das projeções ortogonais.

Partindo da matriz de dados de treinamento *centralizados* X, com m linhas (amostras) e n colunas (características), calcula-se a matriz de covariância (a menos de um fator $1/m$):

$$A = \mathrm{cov}(X) = X^T X\,.$$

Pode-se usar a função `np.linalg.eigh()` do NumPy para calcular a decomposição espectral de A:

$$A = Q\Lambda Q^T\,,$$

ou seja:

$$\mathbf{w}, Q = \texttt{np.linalg.eigh}(A)\,,$$

obtendo-se os autovalores $\mathbf{w}$ e seus autovetores associados nas colunas de Q. **Atenção**: os autovalores (e seus autovetores associados) devem ser ordenados em ordem **decrescente**.

Tarefas

As seguintes tarefas devem ser realizadas como Trabalho Parcial e incluídas nos Relatórios Parcial e Final:

1. Centralizar a matriz de dados original $\tilde{X}$ obtendo a matriz X.
2. Dividir o banco de dados em conjunto de treinamento e conjunto de teste (proporções comumente utilizadas são 60%/40% e 70%/30%.)
3. Calcular a matriz de covariância dos dados *treinamento* de duas maneiras:
   - $\texttt{np.cov}(\tilde{X})$,
   - $\frac{1}{m-1}X^T X$.

   Comparar os resultados. Atenção: verifique se o parâmetro `rowvar` na função `np.cov()` deve ser `True` ou `False`.
4. Calcular a decomposição espectral da matriz de covariância dos dados de treinamento, isto é, $\mathrm{cov}(X) = Q\Lambda Q^T$, usando a função do NumPy:

$$\texttt{np.linalg.eigh}(\mathrm{cov}(X))\,.$$

Questões

Os enunciados e as soluções das questões abaixo devem estar nos Relatórios Parcial e Final.

1. Demonstrar que uma matriz Q é ortogonal (isto é, quadrada com colunas ortonormais) se, e somente se, $Q^T Q = I$.

2. Verificar, com o respectivo banco de dados, que $Q^T Q \approx I$.

3. Consideremos Z a matriz $m \times n$ em cujas linhas estão as coordenadas das amostras em relação à **base de autovetores**, isto é:

$$X^T = QZ^T ; \tag{9}$$

   e $\hat{Q}$ a matriz $n \times r$, cujas colunas são a r primeiras colunas de Q. Mostrar que a matriz cujas linhas são as coordenadas das amostras em relação à **base de autovetores** do subespaço gerado pelos r primeiros autovetores (associados aos r maiores autovalores) é dada por:

$$\hat{Z} = X\hat{Q} .$$

4. Seguindo a Equação (9), a matriz dos dados projetados é **definida** por:

$$\hat{X}^T = \hat{Q}\hat{Z}^T .$$

   Mostrar que os dados projetados são calculados pela equação

$$\hat{X} = X\hat{Q}\hat{Q}^T .$$

5. Calcular, com o respectivo banco de dados, as matrizes $\hat{Z}$ e $\hat{X}$; verificar que a matriz de projeção

$$\hat{Q}\hat{Q}^T$$

   não é a matriz identidade.

5.4 Classificação: SVD

A Decomposição de Valor Singular (SVD), como visto, encontra os autovetores e (a raiz quadrada dos) autovalores da matriz de covariância dos dados $X^T X$, sem calculá-la explicitamente. Assim, a redução de dimensionalidade para o problema de classificação pode ser realizada utilizando

$$X = USV^T . \tag{10}$$

Tarefas

As seguintes tarefas devem ser realizadas como Trabalho Final e incluídas no Relatório Final:

1. Dividir o banco de dados em conjunto de treinamento e conjunto de teste (proporções comumente utilizadas são 60%/40% e 70%/30%.)

2. Calcular, para o respectivo banco de dados, a SVD da matriz de dados de treinamento centralizados X. Gerar um gráfico de número de valores singulares versus variabilidade acumulada.

3. Selecionar valores apropriados de variabilidade acumulada para reduzir a dimensionalidade do problema de classificação, resolvendo-o para o respectivo banco de dados.

4. Gerar um gráfico de número de valores singulares versus acurácia. Isso deve ser feito de maneira apropriada, não devendo o gráfico ser gerado em tempo superior a 24 horas.

Questões

Os enunciados e as soluções das questões abaixo devem estar no Relatório Final.

1. Considerando uma matriz de dados centralizados X $m \times n$, mostrar que as matrizes $X^T X$ e $X X^T$ possuem os mesmos autovalores não nulos.

2. Verificar numericamente a proposição anterior, calculando, para o respectivo banco de dados, os autovalores de $X^T X$ e $X X^T$.

3. Considerando a SVD $X = U S V^T$, mostrar que as colunas de V são os autovetores de $X^T X$ e que as n colunas de U são os autovetores de $X X^T$ associados aos n maiores autovalores.

4. Verificar a proposição anterior, comparando, para o respectivo banco de dados, a matriz de autovetores de $X^T X$ e a matriz de vetores singulares direitos, isto é, V em $X = U S V^T$. Atenção para com os sentidos dos vetores.

5. Comparar, para o respectivo banco de dados, a matriz Q de autovetores de $X X^T$ e a matriz de vetores singulares esquerdos, isto é, U em $X = U S V^T$. Justifique por que as duas matrizes são diferentes, mas as **submatrizes** $U[:, 0:t]$ e $Q[:, 0:t]$ são iguais (t é o índice do primeiro autovalor nulo).

5.5 Seleção de Características: Ortonormalização de Gram-Schmidt

Uma maneira de reduzir a dimensionalidade num problema de classificação é selecionar um subconjunto das características originais. Um algoritmo guloso será inviável se o número de características for razoavelmente grande.

Stoppiglia et alli [Stoppiglia et al., 2003] sugerem a utilização de projeções em complementos ortogonais.

Tarefas

As seguintes tarefas devem ser realizadas como Trabalho Parcial e incluídas nos Relatórios Parcial e Final:

1. Dividir o banco de dados em conjunto de treinamento e conjunto de teste (proporções comumente utilizadas são 60%/40% e 70%/30%.)

2. Implementar e executar um algoritmo **guloso** para todas os possíveis subconjuntos de r características para $r = 1$ e $r = 2$, calculando as respectivas acurácias. Destacar, no relatório, quais foram as maiores acurácias obtidas para $r = 1$ e $r = 2$.

3. Implementar e executar o método descrito em [Stoppiglia et al., 2003], com o respectivo banco de dados.

4. Considerando n características originais ordenadas pelo método descrito em [Stoppiglia et al., 2003], calcular a acurácia da classificação para $r = 1, 2, 3, \ldots, n$ e gerar um gráfico com esses resultados.

5. Comparar os resultados do algoritmo guloso com o método no artigo [Stoppiglia et al., 2003], para $r = 1$ e $r = 2$.

Questões

Os enunciados e as soluções das questões abaixo devem estar nos Relatórios Parcial e Final.

1. Considerando n características originais, quantos subconjuntos com r (com $r \leq n$) características existem?

2. Calcular, para o respectivo banco de dados, o número de subconjuntos com r características, para $r = 1, 2, 3, \ldots, n$. Apresentar a solução em forma de tabela.

3. Escrever, em pseudo-código, o algoritmo proposto por Stoppiglia et alli [Stoppiglia et al., 2003].

4. Considerar o k-ésimo passo do algoritmo proposto por Stoppiglia et alli [Stoppiglia et al., 2003], no qual D (com "número de elementos de D" $= |D| = k$) é o conjunto dos índices já ordenados e D^c (com $|D^c| = n - k$) é o conjunto dos índices ainda não ordenados. Formar uma matriz Q com os k vetores de características já ordenadas, projetadas e normalizadas. Dar a dimensão de QQ^T e mostrar que QQ^T é um projetor, mas não é o projetor identidade.

5. Verificar numericamente a questão anterior, para o respectivo banco de dados.

5.6 Seleção de Características: Componentes Principais

A decomposição espectral pode ser usada para a seleção de características originais. A ideia em [Dunteman, 1989] é ordenar os autovalores, e seus respectivos autovetores associados, da matriz de covariância $X^T X$ dos dados de treinamento centralizados em ordem decrescente, ou seja:

$$\lambda_1 \geq \lambda_2 \geq \ldots \lambda_n \quad \text{e} \quad X^T X \mathbf{q}_i = \lambda_i \mathbf{q}_i \, .$$

Depois, verifica-se qual característica original é mais similar a $\mathbf{q}_1$, descobrindo-se qual é a coordenada de $\mathbf{q}_1$ com maior valor absoluto. Assim, a característica original mais relevante é a de índice dado por:

$$j^* = \underset{j \in \{1,\ldots,n\}}{\operatorname{argmax}} \{|q_{j1}|\} \, .$$

Repete-se o procedimento acima para cada autovetor, associando a coordenada de maior valor absoluto à característica do índice correspondente. No caso de um índice repetir-se, escolhe-se a coordenada com segundo maior valor absoluto, e assim sucessivamente. Neste método, as características são selecionadas em **ordem decrescente** de importância, isto é, da mais importante à menos importante.

Uma variação, também sugerida em [Dunteman, 1989], é selecionar as características em **ordem crescente** de importância, ou seja, considera-se a característica original menos importante aquela mais semelhante ao autovetor associado ao **menor** autovalor, e procede-se de maneira similar ao método anterior, encontrando-se a coordenada de maior valor absoluto e associando seu índice à característica correspondente.

Tarefas

As seguintes tarefas devem ser realizadas como Trabalho Final e incluídas no Relatório Final:

1. Dividir o banco de dados em conjunto de treinamento e conjunto de teste (proporções comumente utilizadas são 60%/40% e 70%/30%.)

2. Implementar e executar os métodos acima com o respectivo banco de dados, ordenando as características originais.

3. Considerando n características originais ordenadas pelos métodos descritos acima, calcular a acurácia da classificação para $r = 1, 2, 3, \ldots$ e gerar um gráfico com esses resultados.

4. Gerar um gráfico mostrando as curvas de acurácia por número de características pelos métodos descritos nesta subseção e na Subseção 5.5.

Questões

Os enunciados e as soluções das questões abaixo devem estar no Relatório Final.

1. Escrever algoritmos em pseudo-código para o dois métodos apresentados nesta subseção (método decrescente e método crescente).

2. Descrever a sua solução para o caso de "empate" na seleção da característica.

3. Explicar por que os métodos aqui descritos podem gerar soluções diferentes.

4. Encontrar uma matriz Q 4×4 **ortogonal** (supondo que as colunas sejam autovetores de alguma matriz simétrica) para a qual os métodos descritos nesta subseção sejam **inconclusivos**. Para isso, é necessário que todas as coordenadas possuam o mesmo valor absoluto.

5. Construir uma matriz diagonal Λ cujos elementos da diagonal principal sejam distintos e estejam em ordem decrescente. Encontrar uma **matriz de correlação** A que possua decomposição espectral $Q\Lambda Q^T$ (com a matriz Q do item anterior). Observação: todos os elementos da diagonal principal de uma matriz de correlação deve ser unitários (isto é, iguais a 1).

6 Bancos de Dados

Regressão

Para o problema de **regressão**, utilizaremos os seguintes bancos de dados:

1. *Airfoil Self-noise* – regressão simples:

    - 1503 amostras;
    - 5 características reais;

- 1 características real para estimar (*self-noise frequency*, em decibéis).

`https://archive.ics.uci.edu/ml/datasets/airfoil+self-noise`

2. *Concrete Compressive Strength* – regressão simples:

   - 1030 amostras;
   - 8 características reais;
   - 1 características real para estimar (*concrete compressive strength*, em MPa).

   `https://archive.ics.uci.edu/ml/datasets/Concrete+Compressive+Strength`

3. *Naval Propulsion Plants* – regressão múltipla (2 variáveis de saída), estimar somente a primeira variável de saída:

   - 11934 amostras;
   - 16 características reais;
   - 2 características reais para estimar, mas estimar somente *GT Compressor decay state coefficient* (remover *GT Turbine decay state coefficient*).

   `http://archive.ics.uci.edu/ml/datasets/condition+based+maintenance+of+naval+propulsion+plants`

   Obs.: Boa oportunidade para rever e aplicar os conceitos de independência linear.

4. *Naval Propulsion Plants* – regressão múltipla (2 variáveis de saída), estimar somente a segunda variável de saída:

   - 11934 amostras;
   - 16 características reais;
   - 2 características reais para estimar, mas estimar somente *GT Turbine decay state coefficient* (remover *GT Compressor decay state coefficient*).

   `http://archive.ics.uci.edu/ml/datasets/condition+based+maintenance+of+naval+propulsion+plants`

   Obs.: Boa oportunidade para rever e aplicar os conceitos de independência linear.

5. *Superconductivity* – regressão simples:

   - 21263 amostras;

- 81 características reais;
- 1 característica real para estimar (*critical temperature*).

`https://archive.ics.uci.edu/ml/datasets/Superconductivty+Data`

Classificação

Para o problema de **classificação**, utilizaremos os seguintes bancos de dados:

1. *Written Numbers MNIST* – reconhecimento de algarismos escritos à mão:

   - 60000 (75%) imagens de treinamento;
   - 20000 (25%) imagens de teste;
   - imagens com 28×28 pixels;
   - 10 algarismos (classes).

   `http://yann.lecun.com/exdb/mnist/`.

   No próprio diretório raiz do banco de dados há programas em Python que extraem os dados no formato `ubyte.gz`.

2. *Sign Language MNIST* – reconhecimento de linguagem americana de sinais:

   - 27455 (79.3%) imagens de treinamento;
   - 7172 (20.7%) imagens de teste;
   - imagens com 28×28 pixels;
   - 24 letras (excluídos o 'J' e o 'Z') (classes).

   `https://www.kaggle.com/datamunge/sign-language-mnist`.

   Um exemplo de como extrair os dados pode ser visto em: `https://analyticsindiamag.com/hands-on-guide-to-sign-language-classification-using-cnn/`.

3. *Cropped Yale B* – reconhecimento facial sob diferentes condições de iluminação:

   - 2424 imagens (amostras);
   - imagens com 168×192 pixels;
   - 38 indivíduos (classes).

   `http://vision.ucsd.edu/~leekc/ExtYaleDatabase/ExtYaleB.html`.

   Obs.: O número de características é maior que o de amostras.

4. *Concrete Crack* – reconhecimento de fraturas em concreto:

   - 40000 imagens;
   - imagens com 227×227 pixels, em 3 canais (RGB);
   - 2 classes (com fratura e sem fratura).

   `https://data.mendeley.com/datasets/5y9wdsg2zt/2`.

   As imagens devem ser baixadas e redimensionadas para, por exemplo, 32×32 pixels. Depois, devem ser convertidas de RGB para níveis de cinza (monocromáticas).

5. *LEGO Bricks* – classificação de peças de LEGO:

   - 40000 imagens;
   - imagens com 400×400 pixels, em 3 canais (RGB);
   - 50 classes.

   `https://www.kaggle.com/joosthazelzet/lego-brick-images`.

   As imagens devem ser baixadas e redimensionadas para, por exemplo, 40×40 pixels. Depois, devem ser convertidas de RGB para níveis de cinza (monocromáticas).

Seleção de Características

Para o problema de **seleção de características**, utilizaremos os seguintes bancos de dados:

1. *Wine Quality - White Wine* – classificação de vinhos em 11 níveis de qualidade (de 0 a 10):

   - 4898 amostras ;
   - 11 características;
   - 7 classes, os níveis de qualidade de 3 a 9.

   `https://archive.ics.uci.edu/ml/datasets/Wine+Quality`.

2. *Wine Quality - Red Wine* – classificação de vinhos em 11 níveis de qualidade (de 0 a 10):

   - 1599 amostras ;
   - 11 características;
   - 6 classes, os níveis de qualidade de 3 a 8.

   `https://archive.ics.uci.edu/ml/datasets/Wine+Quality`.

3. *Abalone* – classificação (ou regressão) para determinar a idade do molusco, que é obtida adicionando-se 1.5 ao número de anéis de crescimento, determinado pela inspeção microscópica de um corte da concha cônica do animal:

   - 4177 amostras;
   - 8 características, sendo uma categórica;
   - O número de classes depende do número máximo e do número mínimo de anéis realmente presentes no banco de dados.

   `https://archive.ics.uci.edu/ml/datasets/Abalone`.

4. *Leaf* – classificação de folhas de plantas:

   - 340 amostras ;
   - 14 características;
   - 30 classes, as espécies de folhas de 1 a 15 e de 22 a 36.

   `https://archive.ics.uci.edu/ml/datasets/leaf`.

   As classes estão na primeria coluna e os espécimens estão na segunda coluna, que não deve ser usada.

5. *Vehicle* – classificação de 4 tipos de veículos:

   - 846 amostras ;
   - 18 características;
   - 4 classes, os tipos de veículo (*bus, opel, saab, van*).

   `https://archive.ics.uci.edu/ml/datasets/Statlog+%28Vehicle+Silhouettes%29`.

Referências

[Dunteman, 1989] Dunteman, G. H. (1989). *Principal Components Analysis*. Quantitative Applications in the Social Sciences. Sage Publications, Newbury Park, CA.

[Jolliffe, 2010] Jolliffe, I. T. (2010). *Principal Component Analysis*. Springer Series in Statistics. Springer, New York, 2 edition.

[Stoppiglia et al., 2003] Stoppiglia, H., Dreyfus, G., Dubois, R., and Oussar, Y. (2003). Ranking a random feature for variable and feature selection. *J. Mach. Learn. Res.*, 3:1399–1414.