

# Projeto

## Introdução à Teoria da Informação

### Prof. Derzu Omaia

### Projeto 2020.2

#### Parte 1:

Implemente um compressor e descompressor utilizando o PPM-C (visto em sala de aula) com o codificador aritmético, ou utilizando o algoritmo LZW. Considere que as mensagens são geradas por fontes com alfabeto  $A = \{0, 1, \dots, 255\}$ . Teste o compressor/descompressor com um corpus de texto em português de 16MB e com um arquivo binário de vídeo.

PPM-C + Aritmético:

O contexto deve ter tamanho máximo  $K$  (parâmetro). O modelo PPM-C alimentará um codificador aritmético. Utilize o mecanismo de exclusão quando necessário. No relatório apresente as curvas de RC x  $K$  e de Tempo de Processamento x  $K$ , para  $K = 0, 1, 2, 3, 4, 5, 6, 7, 8$ . Não é necessário implementar o codificador aritmético, utilize algum já existente.

LZW:

O índice do dicionário deve ser testado com diferentes tamanhos  $K$  bits (parâmetro). Exemplo:  $K=9$ bits tamanho do dicionário:  $2^9=512$ ,  $K=10$ bits tamanho do dicionário  $2^9=1024$ . No relatório apresente as curvas de RC x  $K$  e de Tempo de Processamento x  $K$ , para  $K = 9, 10, 11, 12, 13, 14, 15, 16$  bits. Indique também a quantidade total de índices presentes na mensagem final para cada  $K$ .

Observações:

- Os símbolos do arquivo de teste devem ser lidos no modo binário (números) e não no modo texto (caracteres/strings).
- O codificador deve receber como entrada um arquivo e gerar como saída o arquivo codificado.
- A execução dos experimentos é demorada, evite fazer os experimentos na véspera da entrega pois não dará tempo.

#### Parte 2:

Implementar um reconhecedor de padrões baseado em PPM ou LZW. Utilize um banco de dados previamente rotulado, na etapa de treinamento gere um modelo PPM (árvore) ou LZW (dicionário) para cada categoria do banco de dados selecionado.

Organize o banco de dados em amostras de treino e classificação utilizando a técnica de validação cruzada. Isto é, para cada categoria do banco de dados, selecione todas as

amostras - 1 para treinamento e 1 amostra para classificação, a seleção dessas amostras deve ser aleatória. Para classificação utilize o algoritmo dos K-Vizinhos mais Próximos (K-NN, *k-nearest neighbors*), com  $k=1$ , utilizando como métrica de distância o tamanho do arquivo comprimido. No caso do LZW, outra possibilidade de métrica de distância é a quantidade de índices utilizado pelo LZW, antes da serialização para bytes. A Figura 1 apresenta esse processo de validação cruzada para um banco de dados de texturas de imagens.

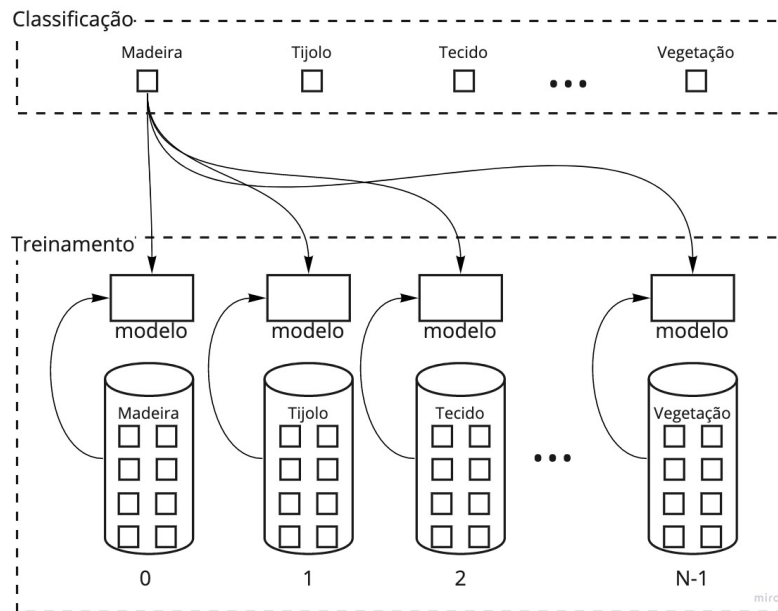


Figura 1 – Esquema de validação cruzada.

O treinamento consiste em gerar o dicionário de cada categoria do banco de dados. A classificação/testes consiste em comprimir 1 amostra (que não foi utilizada na geração do modelo) em todos os modelos/dicionários. Durante a compressão da amostra de teste o dicionário deve permanecer estático. A amostra de teste deve ser atribuída ao modelo que proporcionou a melhor compressão.

No caso do PPM, teste com contexto variável,  $K = 0, 1, 2, 3, 4, 5, 6, 7, 8$ . Para o LZW, teste com dicionários de tamanho  $2^K$ , com  $K = 9, 10, 11, 12, 13, 14, 15, 16$ . Faça um relatório e neste apresente as curvas de Taxa de acerto x  $K$ , e de Tempo de Processamento x  $K$ , para.

Sugestões de bancos de dados para serem utilizados:

1. Iris:
  - a. Banco de dados: *Iris Database Palacký University*
  - b. Utilizar as 40 das 64 pessoas disponíveis no banco. Cada pessoa possui 6 fotos de sua íris.
2. Face
  - a. Banco de dados: [ORL Database of Faces](https://www.dropbox.com/s/mnhfhhb1i51loknk/orl_faces.zip?dl=0).  
([https://www.dropbox.com/s/mnhfhhb1i51loknk/orl\\_faces.zip?dl=0](https://www.dropbox.com/s/mnhfhhb1i51loknk/orl_faces.zip?dl=0))

- b. Utilizar as 40 pessoas disponíveis no banco. Cada pessoa possui 10 fotos de sua face.
- 3. Instrumentos Musicais
  - a. [IRMAS Data Base](https://www.dropbox.com/s/gebspyfkse1bkju/IRMAS-TrainingData_red.zip?dl=0).  
([https://www.dropbox.com/s/gebspyfkse1bkju/IRMAS-TrainingData\\_red.zip?dl=0](https://www.dropbox.com/s/gebspyfkse1bkju/IRMAS-TrainingData_red.zip?dl=0))
  - b. 10 instrumentos musicais disponíveis no banco. Cada instrumento possui aproximadamente 100 amostras de áudio.
- 4. Texturas
  - a. [Brodatz Data Base](https://www.dropbox.com/s/vvzg6xbbcfodn1a/brodatz.zip)  
(<https://www.dropbox.com/s/vvzg6xbbcfodn1a/brodatz.zip>)
  - b. 100 texturas de imagem disponíveis no banco. Cada textura possui apenas 1 amostra. Utilizar as imagens combinadas para classificação.
- 5. Voz
  - a. Escolher o banco.

Observações:

- 1. Utilizar as imagens em escala de cinza.
- 2. Gerar um relatório com o resultado dos experimentos.
- 3. Grupos de até 3 pessoas.

### **Entrega:**

**Parte 1: 20/05/2021**, no horário da aula. Cada dia de atraso reduz em 10% a nota máxima.

**Parte 2: 15/06/2021**, no horário da aula. Cada dia de atraso reduz em 10% a nota máxima.