Dados representativos e transferência de conhecimento

Esta aula constitui um tópico adicional da disciplina. Trata-se de um conteúdo opcional. Sua atividade não valerá nota e não precisa ser enviada.

Importância de se ter uma base de dados representativa

Importância de se ter uma base de dados representativa

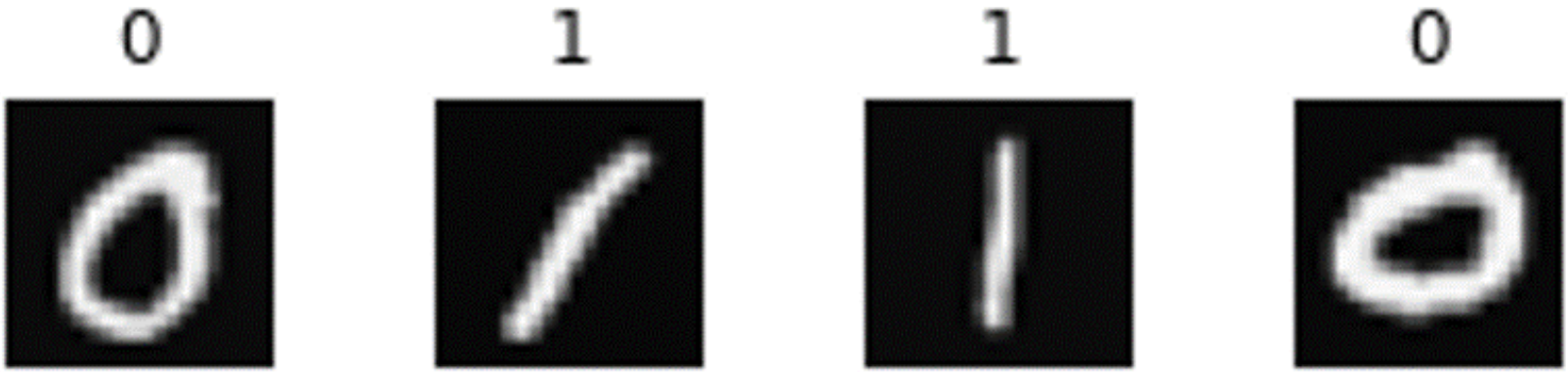
Busque sempre garantir que a quantidade de dados que você tem explica de forma suficiente o problema que está sendo modelado (porém, não basta apenas quantidade, é preciso ter diversidade nas amostras).

Importância de se ter uma base de dados representativa

Busque sempre garantir que a quantidade de dados que você tem explica de forma suficiente o problema que está sendo modelado (porém, não basta apenas quantidade, é preciso ter diversidade nas amostras).

Exemplo:

Você deseja construir um modelo que seja capaz de reconhecer dígitos 0/1 escritos à mão e o seu conjunto de dados possui apenas as 4 amostras abaixo. Será que esse conjunto de amostras contempla uma variedade suficiente de dígitos 0/1 escritos à mão? Basta apenas duplicar essas 4 amostras 1000 vezes, por exemplo?

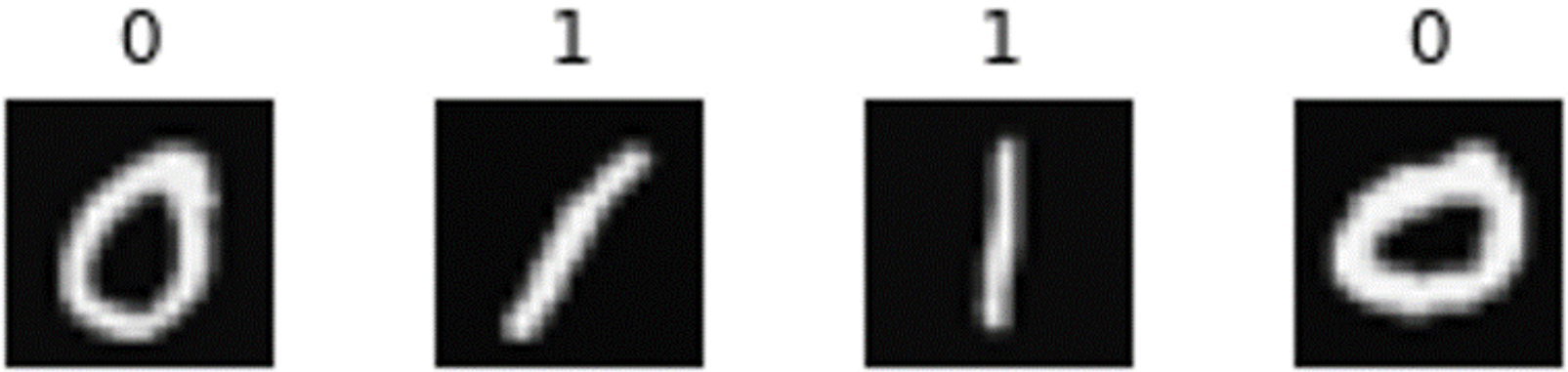


Importância de se ter uma base de dados representativa

Busque sempre garantir que a quantidade de dados que você tem explica de forma suficiente o problema que está sendo modelado (porém, não basta apenas quantidade, é preciso ter diversidade nas amostras).

Exemplo:

Você deseja construir um modelo que seja capaz de reconhecer dígitos 0/1 escritos à mão e o seu conjunto de dados possui apenas as 4 amostras abaixo. Será que esse conjunto de amostras contempla uma variedade suficiente de dígitos 0/1 escritos à mão? Basta apenas duplicar essas 4 amostras 1000 vezes, por exemplo?



Quando a base de dados não é representativa, é possível que ocorra $J_{cv} >> J_{trein}$. Ou seja, o modelo se ajusta aos dados de estimação, mas isso não é suficiente para explicar o problema como um todo. Por isso, o modelo performa mal para novas amostras.

Análise dos erros

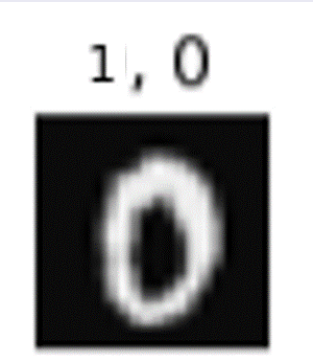
Após treinar um modelo, busque sempre **analisar os erros cometidos pelo modelo**. A análise dos erros pode prover diversas informações relevantes, por exemplo:

- Percebe-se que o modelo erra mais para uma determinada categoria. Existe diversidade suficiente nas amostras de exemplo para essa categoria? $\rightarrow$ **Solução possível:** Obter mais amostras para essa categoria!

Análise dos erros

Após treinar um modelo, busque sempre **analisar os erros cometidos pelo modelo**. A análise dos erros pode prover diversas informações relevantes, por exemplo:

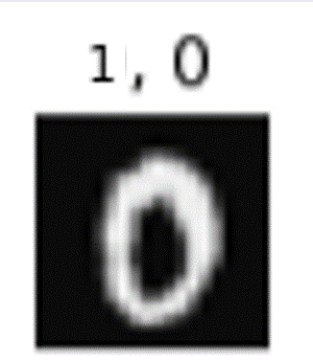
- Percebe-se que o modelo erra mais para uma determinada categoria. Existe diversidade suficiente nas amostras de exemplo para essa categoria? $\rightarrow$ **Solução possível:** Obter mais amostras para essa categoria!
- Existem amostras rotuladas de forma equivocada e o modelo classificou corretamente?



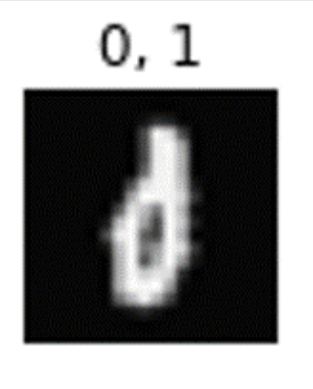
Análise dos erros

Após treinar um modelo, busque sempre **analisar os erros cometidos pelo modelo**. A análise dos erros pode prover diversas informações relevantes, por exemplo:

- Percebe-se que o modelo erra mais para uma determinada categoria. Existe diversidade suficiente nas amostras de exemplo para essa categoria? $\rightarrow$ **Solução possível:** Obter mais amostras para essa categoria!
- Existem amostras rotuladas de forma equivocada e o modelo classificou corretamente?

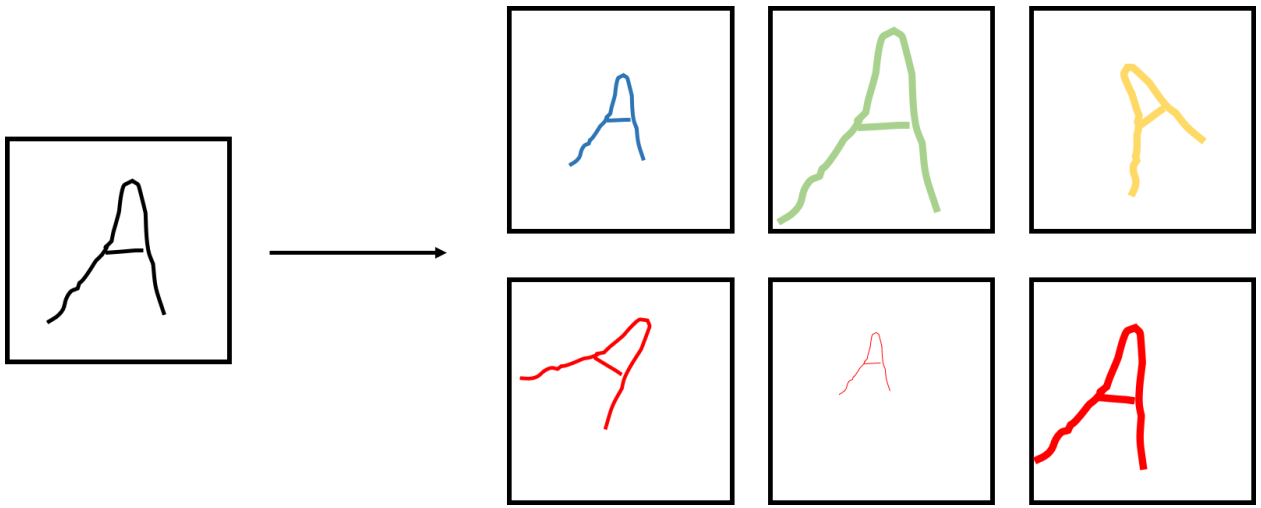


- Tratam-se de erros "aceitáveis"?



Aumentando a base de dados artificialmente

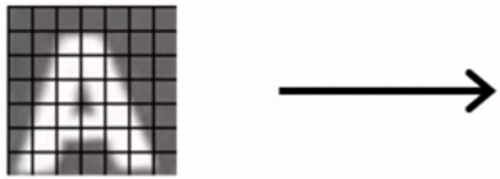
Consiste em criar novas amostras a partir de amostras já existentes (com o objetivo de aumentar a diversidade):

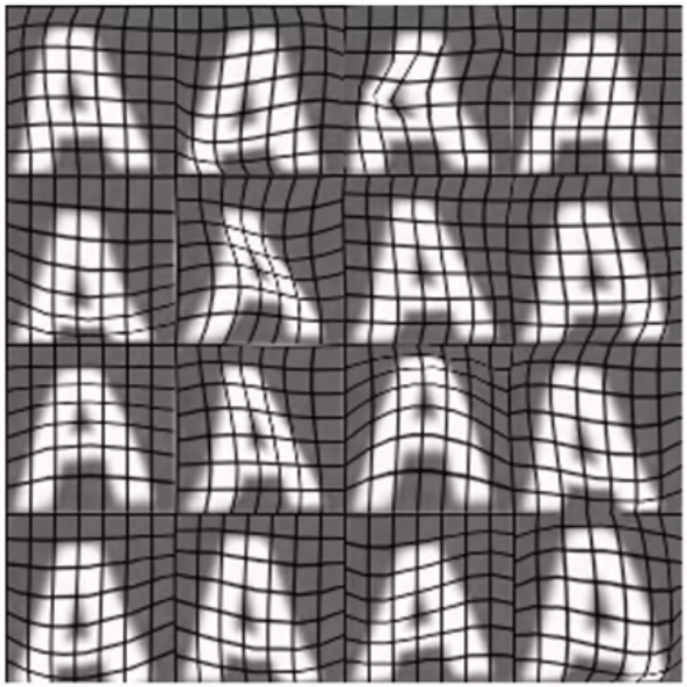


Aumentando a base de dados artificialmente

Consiste em criar novas amostras a partir de amostras já existentes (com o objetivo de aumentar a diversidade):

Exemplo de distorções possíveis:





Aumentando a base de dados artificialmente

Consiste em criar novas amostras a partir de amostras já existentes (com o objetivo de aumentar a diversidade):

Em um sistema de reconhecimento de voz, poderíamos gravar uma pessoa falando, por exemplo: "Alexa, timer de 10 minutos". Em seguida, poderíamos sobrepor a esse áudio diversos tipos de ruído de fundo:

- Pessoas conversando
- Trânsito intenso
- Música de fundo

Observação importante: Adicionar ruído aleatório branco aos dados, em geral, não trás benefício. Para que sejam úteis, as amostras criadas precisam ter relação com o que o modelo possivelmente irá encontrar nas amostras de validação cruzada e de teste.

Geração sintética de dados

Consiste em criar novas amostras do zero, sem usar amostras já existentes:

Real data

Synthetic data

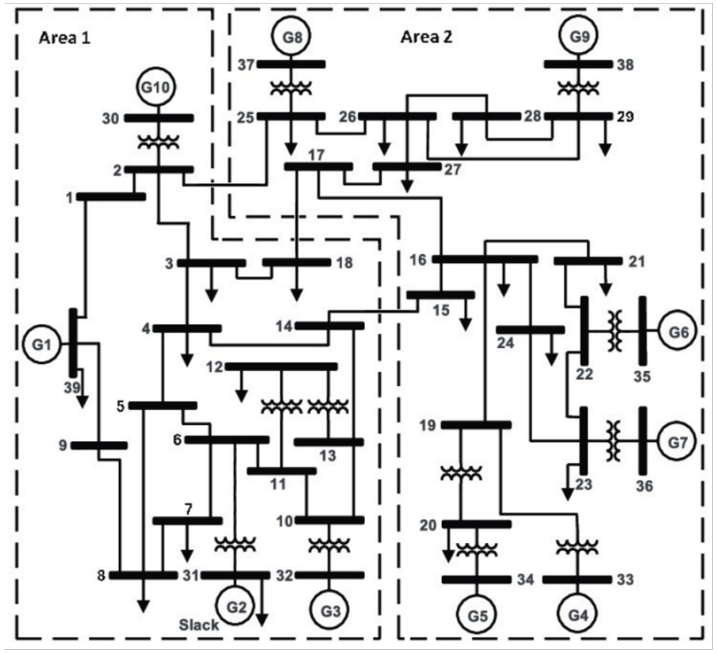
[Adam Coates and Tao Wang]

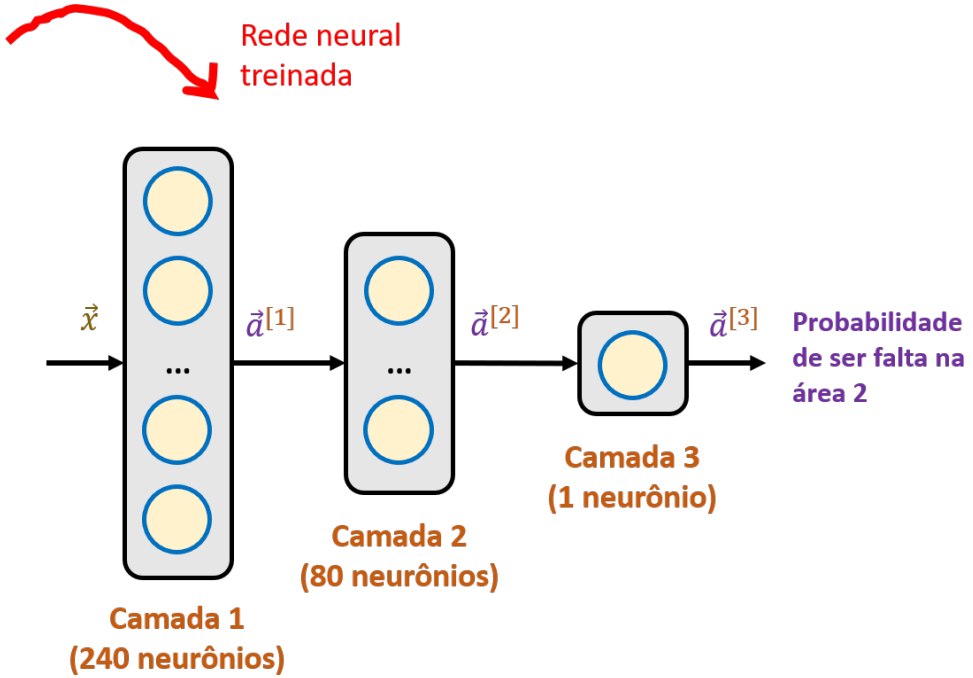
Os dados sintéticos acima apresentados foram criados usando um editor de texto comum. Percebe-se que eles representam bem letras encontradas no "mundo real".

Transferindo conhecimento de uma aplicação para outra

Transferindo conhecimento de uma aplicação para outra

Suponha que você **simulou** o sistema abaixo e treinou um localizador de área sob falta.

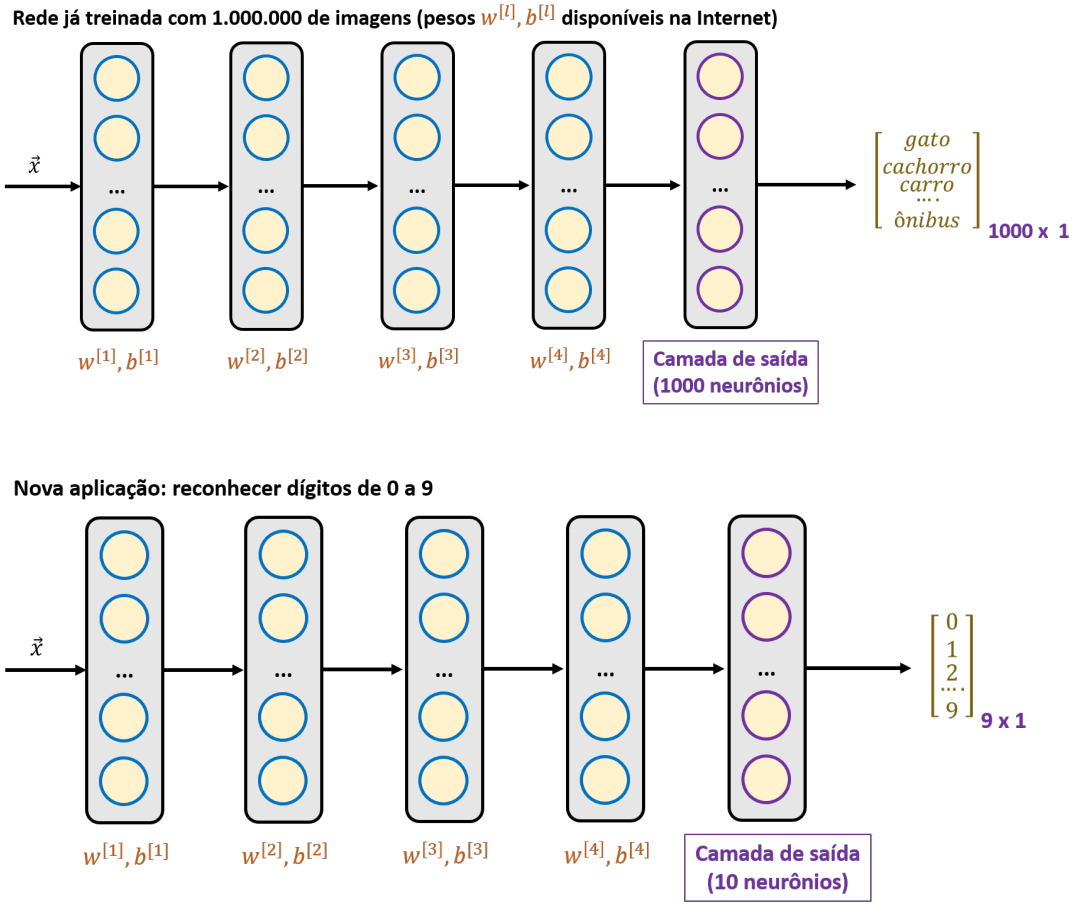




Pergunta:

Seria possível considerar os pesos dessa rede neural já treinada como pesos iniciais para modelagem de um outro sistema onde serão usados dados medidos reais?

Um outro exemplo:



- **Opção 1:** treinar apenas os parâmetros da camada de saída (que precisa ser integralmente substituída)
- **Opção 2:** treinar todos os parâmetros, mas usando parâmetros vindos da rede original como valores iniciais

De olho no código!

Em aulas anteriores, resolvemos o problema de classificação de microchips usando Regressão Logística + Características Polinomiais. Nessa aula, você irá resolver o mesmo problema via Redes Neurais.

Acesse o Python Notebook usando o QR code ou o link abaixo:

`https://colab.research.google.com/github/xaximpvp2/master/blob/main/codigo_aula21_topico_adicional_resolvendo_microchip_usando_RNA.ipynb`

Acesse os dados necessários para rodar o código usando o QR code ou o link abaixo:

`https://ufprbr0-my.sharepoint.com/:t:/g/personal/ricardo_schumacher_ufpr_br/Ee6CfYblcDFEkmfx8FCVXS4B8O-lf5UV3dZunU3R_hY-JQ?e=D1WRIf`

OBS: Para adicionar os dados ao ambiente do Colab Notebook, no menu do canto esquerdo da tela do Colab clique em "Arquivos" e depois "Fazer upload para o armazenamento da sessão". Então carregue os arquivos baixados.

Atividade de aula

Parte 1

Rode todo o código. Certifique-se que você o compreendeu.

Parte 2

1. Treine uma rede neural para o problema e mostre que ela representa um classificador adequado para essa aplicação.