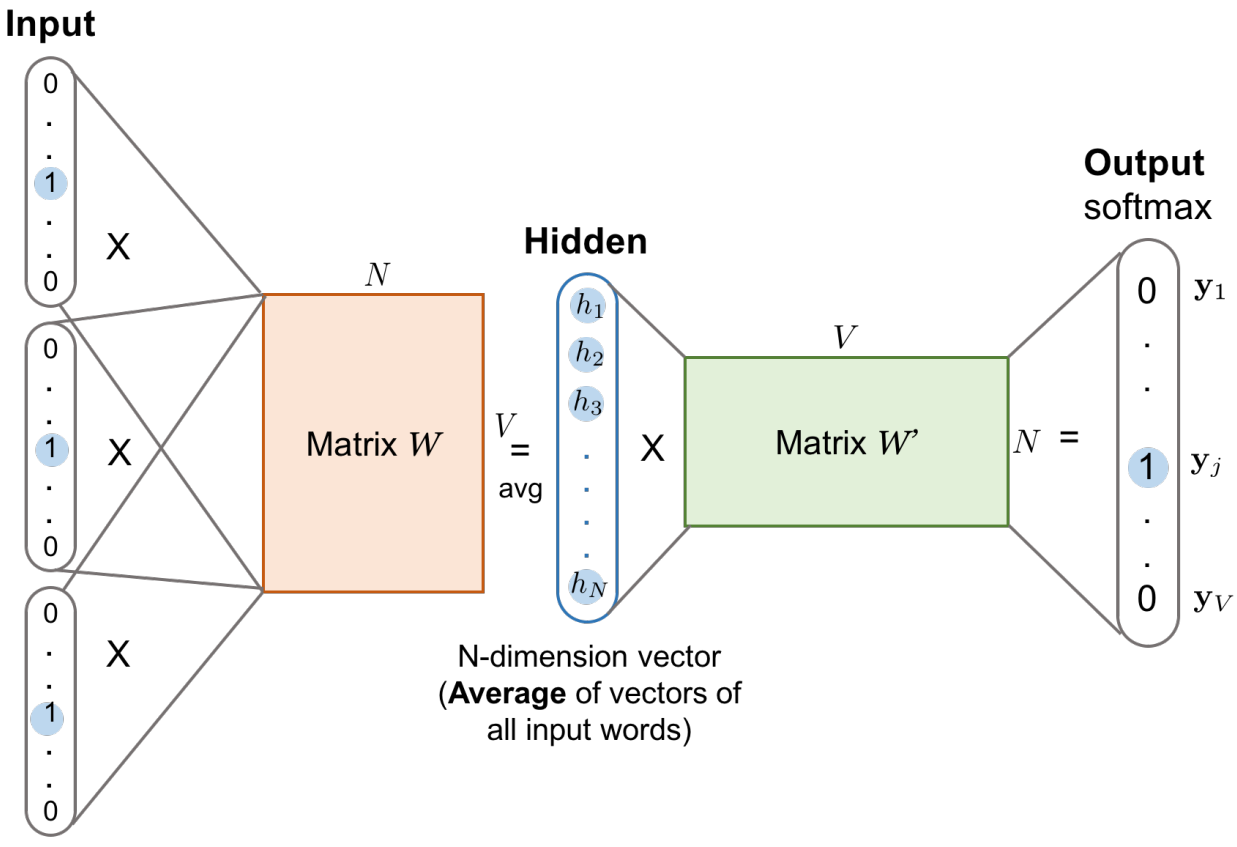
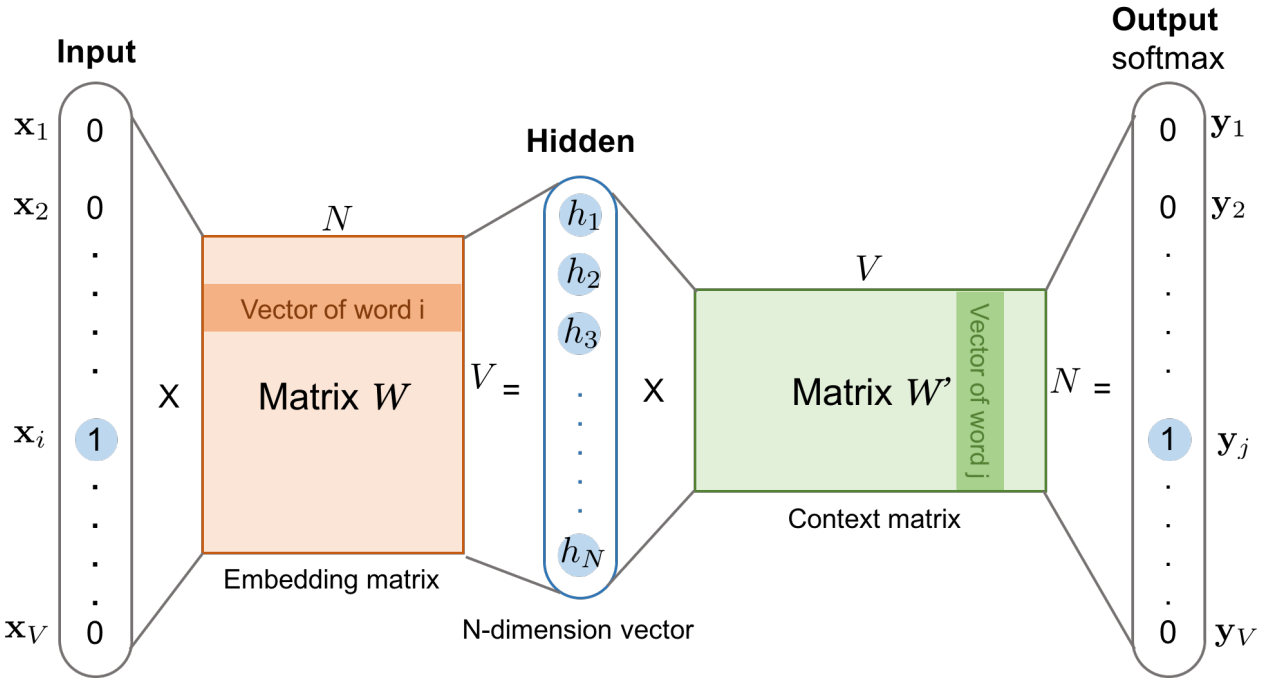
Roteiro

Auxiliar

Embedding - CBoW

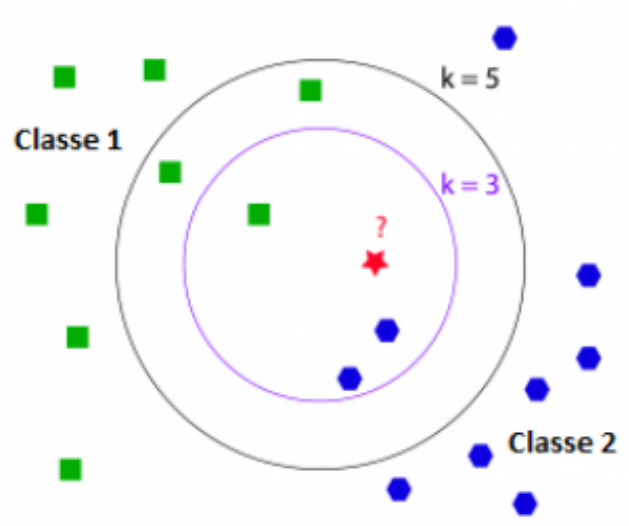


Embedding - Skip-gram



CLASSIFICADORES (*KNN (K-Nearest Neighbor)*)

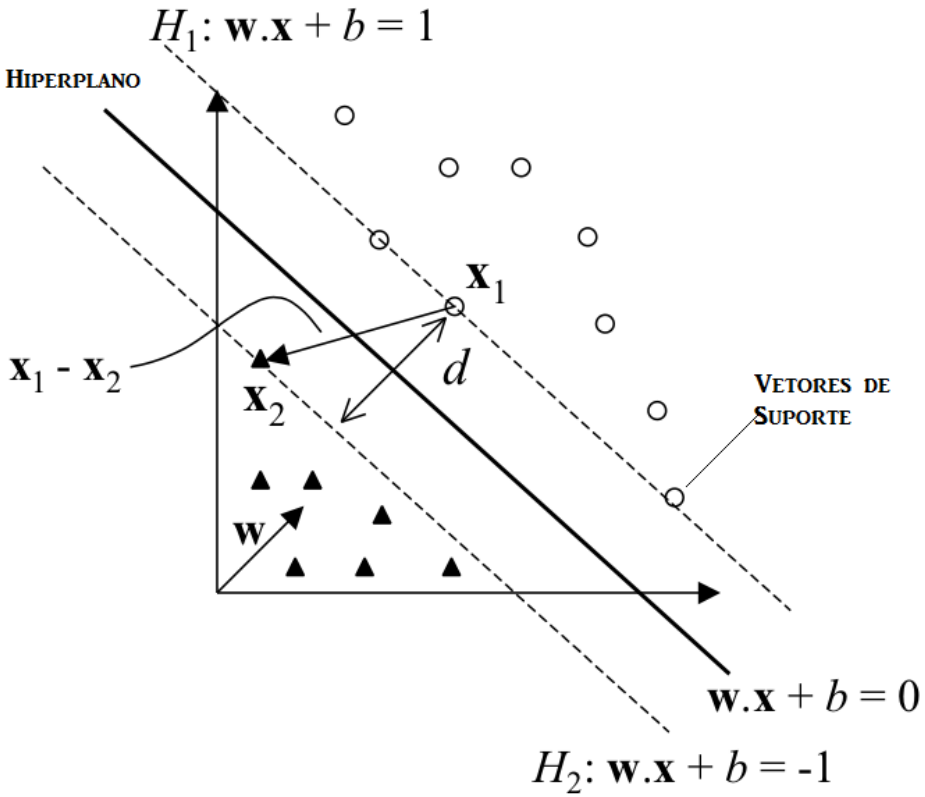
Classifica novas amostras de acordo com as K (5) amostras do conjunto de treinamento mais próximas a essas novas amostras. O KNN usa uma medida de distância (Euclidiana) para definir a semelhança (proximidade) de uma amostra com outra. [1]



[1] DUDA, R. O.; HART, P. E.; STORK, D. G. Pattern classification and scene analysis 2nd ed. ed: Wiley Interscience, 1995.

CLASSIFICADORES (*SVM (Support Vector Machines)*)

Se baseia na margem de separação das classes, onde o objetivo do treinamento é encontrar um hiperplano separador ótimo, aquele em que a distância de separação entre as classes é máxima - hiperplano de margem máxima [2]. *Kernel*: RBF.



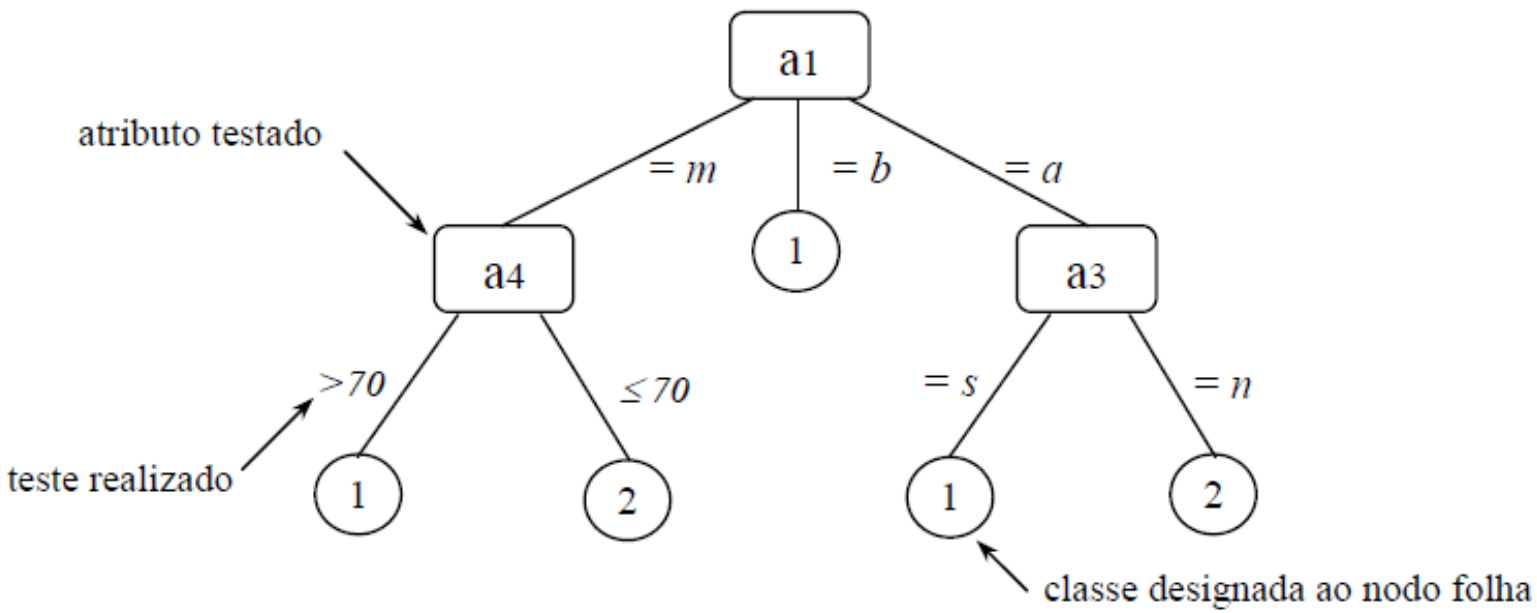
[2] DUTRA, L. P. Detecção das doenças olho de boi e mancha de sarna em maçãs utilizando máquina de vetores de suporte. 2017.

CLASSIFICADORES (*DT (Decision Tree)*)

É constituído essencialmente uma série de decisões *if-else*. Os dados vão sendo particionados em subconjuntos e alguma medida de pureza (gini) dos subconjuntos vai sendo avaliada para decidir quando parar. [3]

Algoritmo de construção utilizado:

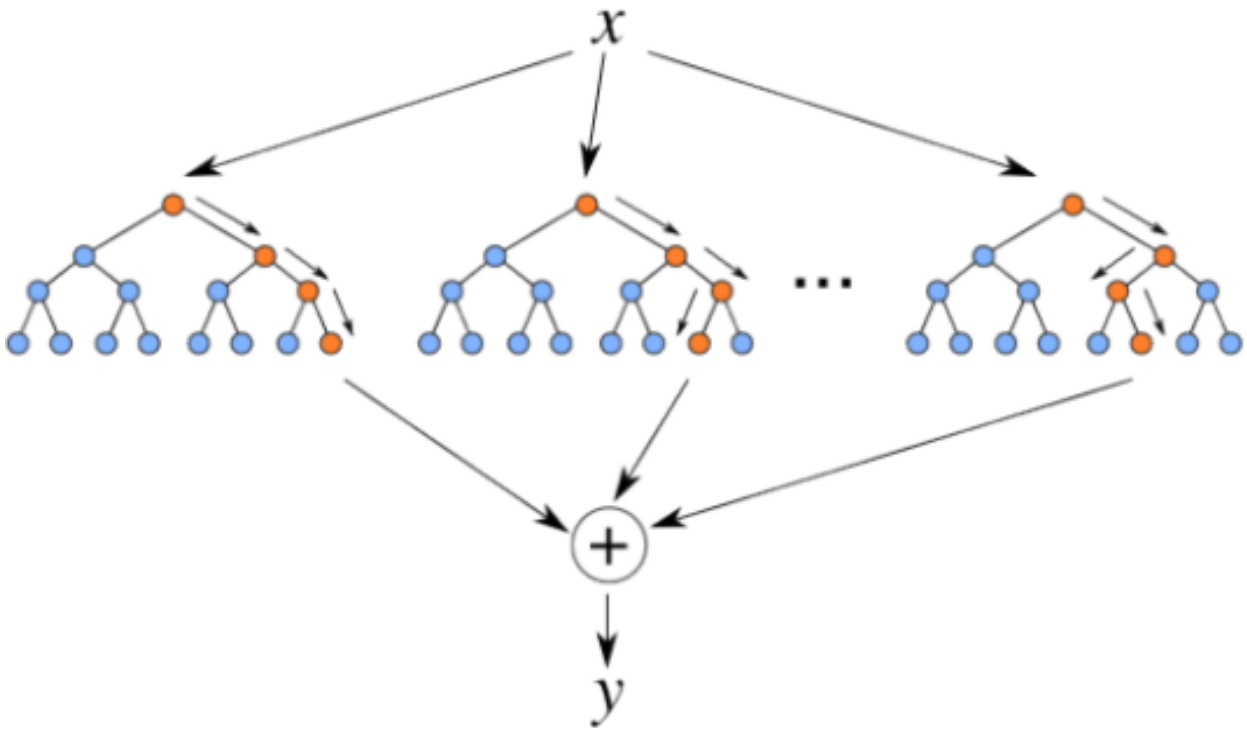
- CART (*Classification and Regression Trees*)



[3] CARACIOLO, M. P. Introdução a AD para classificação e MD. 2009.

CLASSIFICADORES (*RF (Random Forest)*)

Tem o objetivo de efetuar a criação de várias árvores de decisão usando um subconjunto de atributos selecionados aleatoriamente a partir do conjunto original, contendo todos os atributos e que estes possuem um tipo de amostragem chamado de *bootstrap*, a qual é do tipo com reposição, possibilitando assim melhor análise dos dados.

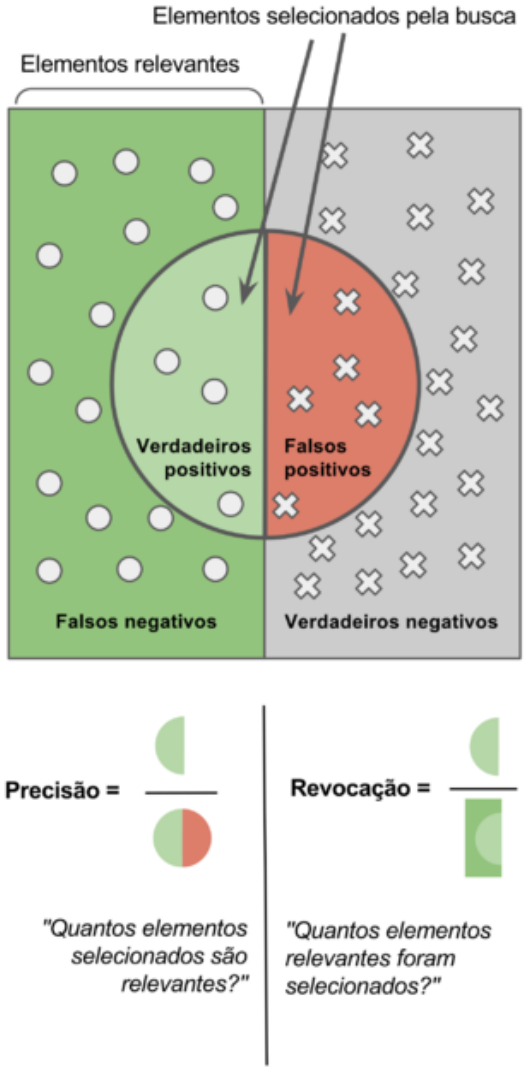


ONE-AGAINST-ALL

One-Against-All: (um contra todos) - Para cada uma das classes cria um modelo de predição binário - treina o modelo criado pra classe 1 contra todas as outras, depois a da classe 2 contra os demais e assim por diante.

Quando um valor de **teste** chega, ele aplica a cada um dos modelos e escolhe o melhor no caso de monorrótulo e os melhores em caso de multirrótulo.

Precision e Recall



Acurácia *- fração das decisões (positivo/negativo) que são corretas.*

Curva ROC X Curva Precision e Recall

Curva ROC - *Receiver Operating Characteristic* - Ela mostra o quão bem o modelo criado pode distinguir entre duas alternativas (classes balanceadas)

Curva *Precision-Recall* - Mais usada pra classes desbalanceadas

AUC - é que uma maneira de resumir a informação da curva ROC num único valor

TREINAMENTOS E TESTES

Combinação	Z5News **(M)**	Z5News **(O)**	Z5BrasilNews **(M)**	Z5BrasilNews **(O)**	Z12News **(M)**	Z12News **(O)**
RF+BoW	0,7818	0,7273	0,7651	0,6745	0,5671	0,4699
SVM(RBF)+FT (E2V-IDF)	0,7846	0,8309	0,7501	0,6888	0,5828	0,6595
SVM(RBF)+W2V (E2V-IDF)	0,7838	0,8047	0,7111	0,6104	0,5954	0,6614

M - Treinado e testado na mesma coleção
O - Treinado em uma coleção e testado em outras coleções

Observação
A Z5NewsBrasil foi testada apenas com outra fonte; As demais foram testadas com outras fonte, mas tem também notícias da própria fonte (ainda não vistas).

DIVERSOS

Lemmatization - Relacionado a forma da Escrita;

RMSE - mede a magnitude média do erro, ou seja, é a raiz quadrada da média das diferenças quadradas entre previsão e observação real;