Universidade Federal do Paraná
Departamento de Informática
INFO 7004 – Aprendizagem de Máquina

Relatório - Trabalho 1 - kNN

Aluno: Dimmy Karson Soares Magalhães

Especificação do trabalho

Considere a base IMDB disponível na página da disciplina, a qual é composta de duas classes e contém 100000 registros.
Atividades:

- Extrair uma representação da sua escolha (Bag of Words, Word Embedding, etc).
- Implemente o classificador kNN (a saída deve deve ser a taxa de reconhecimento e a matriz de confusão)
- Avalie diferentes valores de k e métricas de distância na base de validação. Verifique o desempenho das suas escolhas nas bases de teste.

Para implementar a solução para o problema foi usada a linguagem python com as bibliotecas Word2Vec, Numpy e Scipy. O código encontra-se disponível no endereço do GitHub: https://github.com/dimmykarson/ml/tree/master/trab1. Para realizar a configuração necessária para replicar os experimentos é necessário que as bibliotecas sejam instaladas através do comando “pip install -r requirements.txt”.

Para executar o script deve-se rodar o seguinte comando:

python knn.py “train.csv” “test.csv” k

Onde “train.csv” consiste no arquivo de treinamento do algoritmo, “test.csv” e *k* um valor inteiro maior do que zero para determinar o tamanho de vizinhos a serem analisados pelo kNN.

Do desenvolvimento

Para o desenvolvimento, foi usado Word2Vec para criar um vetor representativo de cada palavra de cada comentário da base. Em seguida é criado um vetor médio de todas as palavras do comentários, para assim criar uma representação do comentário como um todo. Essa representação conta com 300 elementos discricionários.

```
model = gensim.models.Word2Vec(sentences, min_count=1, size=300)
```

Para a estrutura de dados, foi construída uma árvore em que cada nó era expandido de acordo com a mediana de uma dada característica aleatória, sendo que cada nó folha possuiria no máximo 1000 vetores de características.

Para o cálculo das distância entre os vetores, foi utilizado a biblioteca Scipy, e foram adicionadas aos testes as distâncias:

- Euclidiana
- Manhattan
- Cosine

Por padrão o algoritmo irá realizar o cálculo de distâncias usando o método euclideano.

Dos Testes

Para realização dos testes um grupo de experimentos foi montado no qual eram avaliados os valor de K e a tipo de medida de distância utilizada. Os valores de K variaram, para cada medida de distância, em 3, 5, 10, 20 e 30. Cada uma sendo testada em toda a base de validação, no repositório denominada como “validation.csv”.

Dos resultados

Os seguintes resultados foram encontrados:

Teste 1. Tamanho do teste: 25000
Treinamento... k=3 e distância: euclidean
Precisão: 60,252%
Matriz de confusão:

	P	N
P	10172	2328
N	7609	4891

Teste 2. Tamanho do teste: 25000
Treinamento... k=5, distância: euclidean
Precisão: 60,816%
Matriz de confusão:

	P	N
P	5364	7136
N	2660	9840

Teste 3. Tamanho do teste: 25000
Treinamento... k=10, distância: euclidean
Precisão: 59,32%
Matriz de confusão:

	P	N
P	3466	9034
N	1136	11364

Teste 4. Tamanho do teste: 25000
Treinamento... k=20, distância: euclidean
Precisão: 67,84%
Matriz de confusão

	P	N
P	7862	4638
N	3402	9098

Teste 5. Tamanho do teste: 25000
Treinamento... k=30, distância: euclidean
Precisão: 59,02%
Matriz de confusão

	P	N
P	3206	9294
N	951	11549

Teste 6. Tamanho do teste: 25000
Treinamento... k=3, distância: manhattan
Precisão: 65,5%
Matriz de confusão

	P	N
P	8948	3552

N	5073	7427

Teste 7. Tamanho do teste: 25000
Treinamento... k=5, distância: manhattan
Precisão: 57,39%
Matriz de confusão

	P	N
P	2896	9604
N	1047	11453

Teste 8. Tamanho do teste: 25000
Treinamento... k=10, distância: manhattan
Precisão: 68,00%
Matriz de confusão

	P	N
P	5716	6784
N	1214	11284

Teste 9. Tamanho do teste: 25000
Treinamento... k=20, distância: manhattan
Precisão: 72,12%
Matriz de confusão

	P	N
P	7134	5366
N	1603	10897

Teste 10. Tamanho do teste: 25000
Treinamento... k=30, distância: manhattan
Precisão: 52,15%
Matriz de confusão

	P	N

P	3722	8778
N	3183	9317

Teste 11. Tamanho do teste: 25000
Treinamento... k=3, distância: cosine
Precisão: 29,18%
Matriz de confusão

	P	N
P	4879	7621
N	10022	2418

Teste 12. Tamanho do teste: 25000
Treinamento... k=5, distância: cosine
Precisão: 57,21%
Matriz de confusão

	P	N
P	6554	5496
N	4750	7750

Teste 13. Tamanho do teste: 25000
Treinamento... k=10, distância: cosine
Precisão: 67,04%
Matriz de confusão

	P	N
P	9273	3227
N	5013	7487

Teste 14. Tamanho do teste: 25000
Treinamento... k=20, distância: cosine
Precisão: 75,15%
Matriz de confusão

	P	N
P	10018	2482
N	3729	8771

Teste 15. Tamanho do teste: 25000
Treinamento... k=30, distância: cosine
Precisão: 59,61%
Matriz de confusão

	P	N
P	7481	5019
N	5077	7423

Segue o gráfico comparativo entre os métodos.