Relatório da implementação dos Algoritmos K-Nearest Neighbors e Naïve de Bayes

Aprendizagem Automática Jailson Varela nº40699

26 de novembro de 2024

1 Introdução

Como descrito no enunciado do presente trabalho, proposto pela professora Teresa Gonçalves, o objetivo do mesmo é implementar as classes K Nearest Neighbors e Naïve Bayes e os seus métodos essenciais que lidam com os dados (atributos numéricos e classes nominais) para, deste modo, implementar modelos de acordo com os diferentes hiperparâmetros k, p e suave.

2 Metodologia

2.1 Objectivo

O objetivo do projeto passa por implementar 6 modelos KNN e modelos Naive Bayes com os diferentes hiperparâmetros KNN: k=1, 5, 9, p=1, 2 Naïve Bayes: suave=1e-9, 1e-5 para cada um dos dados em estudo iris.csv, rice.csv e entrega_antecipada.csv.

2.2 Descrição do algoritmo

No projeto foi implementado o tratamento dos ficheiros recebidos, os quais foram iris.csv, rice.csv(este foi necessário a alteração do nome da classe 'class' para um nome 'classerice' que não dê conflitos), wdbc.csv(este que foi necessário a remoção da coluna de "id"presente) e entrega_antecipada.csv que passa também por separar as classes dos atributos.

Depois de efetuar o tratamento de dados, passamos à implementação dos algoritmos KNN e Naive Bayes(NB), com os valores por omissão/default (k=3, p=2.0 e suave=1e-9). Guardamos os dados de KNN em:

- self.X_train com formato numpy.array;
- self.y_train com formato numpy.array;

e de NB em:

- self.X_train;
- self.y_train;
- self.classes;
- self.mean;
- self.var;
- self.priors.

contendo os valores como descrito pelo nome atributo. Testamos em simultâneo e não consegui concluir com a criação dos modelos desejados/propostos pela professora abaixo para realizar estudo/análise.

2.3 Conjunto de dados criados

O conjunto de dados criados foi com a intenção de estudar a "entrega antecipada dos trabalhos"em contexto acadêmico. Reunimos alguns exemplos com os atributos numéricos "ID", "Tempo Dedicado(horas)". "Tempo Dedicado Previsto(horas)"e "Horas Para Entregar"e a classe nominal denominada "class"com o domínio: Sim ou Não.

3 Análises dos modelos

Modelos a apresentar:

- •KNN: k=1, 5, 9, p=1, 2
- • Naïve Bayes: suave=1e-9, 1e-5.

3.1 Modelos K-Nearest Neighbors (KNN)

3.1.1 Conjunto de dados iris

Modelo KNN: k=1, p=1:

- • Desempenho sobre conjunto de treino : 1.0
- • Desempenho sobre conjunto de teste : 0.95

Modelo KNN: k=5, p=1:

- • Desempenho sobre conjunto de treino : 0.97
- • Desempenho sobre conjunto de teste : 0.95

Modelo KNN: k=9, p=1:

- • Desempenho sobre conjunto de treino : 0.97
- • Desempenho sobre conjunto de teste : 0.97

Modelo KNN: k=1, p=2:

- • Desempenho sobre conjunto de treino : 1.0
- • Desempenho sobre conjunto de teste : 0.95

KNN	k=1	k=5	k=9
conj. treino	1.0	0.97	0.97
conj. teste	0.95	0.95	0.97

Tabela 1: Desempenho para p=1 Conjunto de dados iris

KNN	k=1	k=5	k=9
conj. treino	1.0	0.97	0.97
conj. teste	0.95	0.95	0.97

Tabela 2: Desempenho para p=2 Conjunto de dados iris

Modelo KNN: k=5, p=2:

- • Desempenho sobre conjunto de treino : 0.97
- • Desempenho sobre conjunto de teste : 0.95

Modelo KNN: k=9, p=2:

- • Desempenho sobre conjunto de treino : 0.97
- • Desempenho sobre conjunto de teste : 0.97

3.1.2 Análise do Conjunto de dados iris

Nota-se que os modelos com...

- p = 1 e p = 2 são idênticos em termos desempenho nos conjuntos de treino e de teste consoante o número de vizinhos :
  - K = 9 apresenta ser um modelo mais generalista, ou seja, com melhor desempenho tendo maior número de vizinhos. Este modelo é o melhor destes 3 modelos pois apresenta melhor desempenho no conjunto de teste e não está demasiado ajustado aos dados de treino.
  - e nos casos em que k é menor:
    - k=5 o modelo tem menor desempenho sobre o conjunto de teste e é menos ajustado ao conjunto de treino;
    - k=1 o modelo apresenta evidências de sobre ajustamento, o seja, o modelo está demasiado ajustado ao conjunto de treino e com desempenho menor, então está menos generalista.

3.1.3 Conjunto de dados rice

Modelo KNN: k=1, p=1:

- • Desempenho sobre conjunto de treino : 1.0
- • Desempenho sobre conjunto de teste : 0.88

Modelo KNN: k=5, p=1:

- • Desempenho sobre conjunto de treino : 0.91
- • Desempenho sobre conjunto de teste : 0.9

Modelo KNN: k=9, p=1:

- • Desempenho sobre conjunto de treino : 0.9
- • Desempenho sobre conjunto de teste : 0.9

Modelo KNN: k=1, p=2:

- • Desempenho sobre conjunto de treino : 1.0
- • Desempenho sobre conjunto de teste : 0.87

Modelo KNN: k=5, p=2:

- • Desempenho sobre conjunto de treino : 0.91
- • Desempenho sobre conjunto de teste : 0.9

Modelo KNN: k=9, p=2:

- • Desempenho sobre conjunto de treino : 0.89
- • Desempenho sobre conjunto de teste : 0.89

3.1.4 Análise do Conjunto de dados rice

Nota-se que os modelos com p = 1 e p = 2 tem ligeira semelhança em termos desempenho nos c

- p = 1 :
  - K = 5 apresenta ser um modelo mais generalista, ou seja, com melhor desempenho tendo maior número de vizinhos. Este modelo é o melhor destes 3 modelos pois apresenta melhor desempenho no conjunto de teste e não está demasiado ajustado aos dados de treino.

KNN	k=1	k=5	k=9
conj. treino	1.0	0.92	0.91
conj. teste	0.88	0.91	0.90

Tabela 3: Desempenho para p=1 Conjunto de dados rice

KNN	k=1	k=5	k=9
conj. treino	1.0	0.91	0.89
conj. teste	0.87	0.9	0.89

Tabela 4: Desempenho para p=2 Conjunto de dados rice

  - e nos casos em que k é menor:
    * k=9 o modelo tem menor desempenho sobre o conjunto de teste e é menos ajustado ao conjunto de treino;
    * k=1 o modelo apresenta evidencias de sobre-ajustamento, os seja, os modelo está demasiado ajustado ao conjunto de treino e com desempenho menor, está menos generalista.

- p = 2 :
  - K = 5 apresenta-se como um modelo mais generalista, ou seja, com melhor desempenho, tendo maior número de vizinhos. Este modelo é o melhor destes 3 modelos, pois apresenta melhor desempenho no conjunto de teste e não está demasiado ajustado aos dados de treino.
  - e nos casos em que k é menor:
    * k=9 o modelo tem menor desempenho sobre o conjunto de teste e é menos ajustado ao conjunto de treino;
    * k=1 o modelo apresenta evidencias de sobre-ajustamento, os seja, os modelo está demasiado ajustado ao conjunto de treino e com desempenho menor, então está menos generalista.

3.1.5 Conjunto de dados entrega_antecipada

Modelo KNN: k=1, p=1:

- • Desempenho sobre conjunto de treino : 1.0
- • Desempenho sobre conjunto de teste : 1.0

Modelo KNN: k=5, p=1:

- • Desempenho sobre conjunto de treino : 0.98

KNN	k=1	k=5	k=9
conj. treino	1.0	0.98	0.98
conj. teste	1.0	1.0	1.0

Tabela 5: Desempenho para p=1 Conjunto de dados entrega_antecipada

KNN	k=1	k=5	k=9
conj. treino	1.0	0.98	0.98
conj. teste	1.0	1.0	1.0

Tabela 6: Desempenho para p=2 Conjunto de dados entrega_antecipada

- • Desempenho sobre conjunto de teste : 1.0

Modelo KNN: k=9, p=1:

- • Desempenho sobre conjunto de treino : 0.98
- • Desempenho sobre conjunto de teste : 1.0

Modelo KNN: k=1, p=2:

- • Desempenho sobre conjunto de treino : 1.0
- • Desempenho sobre conjunto de teste : 1.0

Modelo KNN: k=5, p=2:

- • Desempenho sobre conjunto de treino : 0.98
- • Desempenho sobre conjunto de teste : 1.0

Modelo KNN: k=9, p=2:

- • Desempenho sobre conjunto de treino : 0.98
- • Desempenho sobre conjunto de teste : 1.0

3.1.6 Análise do Conjunto de dados entrega_antecipada

Nota-se que os modelos com...

- p = 1 e p = 2 são idênticos em termos desempenho nos conjuntos de treino e de teste consoante o número de vizinhos :

- K=9 e K=5 apresentam ser modelos mais generalistas, ou seja, com melhores desempenhos tendo maiores números de vizinhos. Estes modelos são os melhores destes 3 modelos pois apresentam melhores desempenhos no conjunto de teste e não estão demasiado ajustados aos dados de treino.
- e no caso em que k=1: o modelo apresenta evidencias de sobre-ajustamento, os seja, os modelo está demasiado ajustado ao conjunto de treino e com desempenho menor, então está menos generalista.

Concluímos que quando numero de vizinhos = 1 o modelo apresenta evidências de sobre-ajus
Também notamos que o conjunto de dados que criamos não apresenta grande variação de dados

3.2 Modelos Naïve Bayes (NB)

3.2.1 iris

Modelo NB: suave=1e-9:

- • Desempenho sobre conjunto de treino :
- • Desempenho sobre conjunto de teste :

Modelo NB: suave=1e-5:

- • Desempenho sobre conjunto de treino :
- • Desempenho sobre conjunto de teste :

3.2.2 rice

Modelo NB: suave=1e-9:

- • Desempenho sobre conjunto de treino :
- • Desempenho sobre conjunto de teste :

Modelo NB: suave=1e-5:

- • Desempenho sobre conjunto de treino :
- • Desempenho sobre conjunto de teste :

3.2.3 dadosadicionaiscriados

Modelo NB: suave=1e-9:

- • Desempenho sobre conjunto de treino :
- • Desempenho sobre conjunto de teste :

Modelo NB: suave=1e-5:

- • Desempenho sobre conjunto de treino :
- • Desempenho sobre conjunto de teste :