Aprendizagem Automática

Licenciatura em Engenharia Informática

Trabalho Prático 2023/2024 (v1.1)

-- KNN e Naïve Bayes --

1. Objetivo

Implementar os algoritmos KNN e Naïve de Bayes para problemas de classificação com atributos nominais ambiente scikit-learn.

2. Descrição do trabalho

Pretende-se implementar os algoritmos KNN e Naïve de Bayes para problemas de **classificação** que **permitam a integração com o ambiente scikit-learn**. As classes a desenvolver devem permitir criar o modelo, aplicar o modelo e calcular o desempenho. A criação dos modelos dependem da escolha de:

- KNN: nº de vizinhos (óbvio) e métrica de Minkowski;
- Naïve de Bayes: estimador de probabilidades.

3. Implementação

As classes a implementar, **KNeigborsClassUE** e **NBayesClassUE**, deverão ser o mais compatível possível com o ambiente scikit-learn, ou seja, os parâmetros de entrada e saída dos métodos deverão permitir a substituição destes por outros algoritmos de classificação implementados no scikit-learn. Assim:

- **inicialização** do objeto: definição dos parâmetros do algoritmo
- método **fit(X,y)**: constroi (e guarda na estrutura de dados adequada) o modelo a partir do conjunto fornecido:
  - X: array com forma (nexemplos, natributos). Dados de treino;
  - y: array com forma (nexemplos). Etiquetas;
- método **predict(X)**: aplica o modelo, e devolve as etiquetas previstas para o conjunto fornecido:
  - X: array com forma (nexemplos, natributos). Dados de teste;
- método **score(X,y)**: prevê o valor associado a cada exemplo do conjunto e devolve a **exatidão**:
  - X: array com forma (nexemplos, natributos). Dados (treino ou teste);
  - y: array com forma (nexemplos). Etiquetas (treino ou teste);

3.1. KNeigborsClassUE

Para este algoritmo, assuma que os atributos são **numéricos**. Os parâmetros para a inicialização do objeto são:

- **k**: int; valor por omissão é 3
  Nº de vizinhos mais próximos a considerar.
- **p**: float, valor por omissão é 2.0

Potência para a métrica Minkowski; quando p=1 é equivalente à distância de Manhattan (l1); com p=2 temos a distância Euclidiana (l2).

Métrica de Minkowski

$$D(X,Y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{\frac{1}{p}}$$

3.2. NBayesClassUE

Para este algoritmo, assuma que os atributos são **nominais**. Considere a situação em que, no conjunto de teste, poderão existir valores dos atributos que não ocorrem nos dados de treino; proponha uma solução, e implemente-a.

Os parâmetros para a inicialização do objeto são:

- **alpha**: float (valor positivo), valor por omissão é 1
  Parâmetro aditivo de suavização para o cálculo da estimativa de probabilidades (Lidstone); quando alpha=1 é equivalente ao estimador de Laplace; com alpha=0 não há suavização.

Estimador Lidstone

$$P(x) = \frac{n_x + \alpha}{total + \alpha * nvals}$$

4. Dados

Estão disponíveis no moodle vários conjuntos de dados para testar os algoritmos. Assuma que o input é um **ficheiro csv** onde:

- a 1ª linha do ficheiro identifica os atributos
- o atributo a prever está na última coluna

Recomenda-se a consulta e análise dos ficheiros para verificar a pertinência de cada atributo para o problema.

A utilização da biblioteca **Pandas** (pandas.pydata.org) permite a leitura do ficheiro csv para uma estrutura de dados compatível com o sklearn. A função **read_csv()** lê o ficheiro para um **DataFrame**, uma estrutura de dados semelhante ao Excel (onde as linhas e colunas são etiquetadas), que permite identificar colunas (e linhas) a partir do nome. Num DataFrame também é possível identificar colunas (e linhas) através de índices.

O ficheiro **read_split_write.py** exemplifica a utilização de DataFrame (usando identificadores dos atributos e índices).

5. Condições gerais

O trabalho deverá ser efetuado em grupos de 2 ou 3 alunos e será apresentado em dia e horário a combinar. Deve ser submetido no moodle através de um ficheiro **.tar.gz** ou **.zip**. O conteúdo deve incluir o

código fonte do trabalho, adequadamente comentado, e um relatório em formato **PDF**. O relatório deve incluir a:

- explicação da implementação escolhida, nomeadamente as **estruturas de dados** que permitem guardar o modelo;
- Para cada conjunto, **fazer uma análise crítica** do desempenho, sobre os sub-conjuntos de treino e teste, dos modelos criados com os seguintes parâmetros:
    - KNN: K={1, 3, 5, 9}, p={1, 2}
    - Naïve Bayes: alpha={0, 1, 3, 5}

O trabalho deve ser submetido através moodle até à data indicada, e nome do ficheiro deverá ser os números dos alunos por ordem crescente (e.g. "33333_44444_55555.zip").