K-Nearest Neighbours & RSTUDIO

Rodrigo Augusto Igawa

Universidade Estadual de Londrina

igawa@uel.br

17 de novembro de 2015

Sumário

Relembrando a última aula

KNN

- Amplamente usado
- Aprendizado de máquina supervisionado
- Modelo simples
- Buscar os vizinhos mais próximos pode ser custoso

Na aula de hoje

Na aula de hoje

- Como usar o KNN no R
- KNN e sua performance em alguns datasets

KNN no R

Package *class*

O pacote class disponibiliza alguns algorítimos para classificação. Um deles é o KNN

- train $\leftarrow$ conjunto de teste (sem a classe)
- test $\leftarrow$ conjunto de treino (sem a classe)
- cl $\leftarrow$ a coluna das classes do conjunto de treino
- k $\leftarrow$ número de vizinho a serem considerados
- prob $\leftarrow$ distribuição probabilística dos resultados

Detalhes de implementação do KNN no R

Limitações

- A distância a ser usada é exclusivamente a euclidiana
- Não existe possibilidade de ponderar atributos

Atividades

Exercícios

Efetuar testes nos datasets descritos nos próximos slides.
Para cada conjunto de dados computar a matriz de confusão usando KNN e Random Forest.

O Dataset de iris

Descrição

Conjunto de dados com 150 instâncias (75 para teste, 75 para treino).
Objetivo: Separar o conjunto de dados em três classes distintas.
Atributos: 4 atributos numéricos
Exemplos: Petal.length{1, 6.9}, Petal.width{0.1, 2.5}

Atividade - Dataset de iris

Objetivo

Dividir o conjunto de dados em duas partes de mesmo tamanho e, em seguida, construir a matriz de confusão usando o 5NN e Random Forest como classificadores.

O Dataset de sites de *phishing*

Descrição

Conjunto de dados com 11578 instâncias (11054 para treino, 524 para teste).

Objetivo: Deseja-se classificar se um site é usado para a prática de phishing. Cada exemplar do conjunto de dados possui seu rótulo na última coluna.

Atributos: Todos numéricos, variando entre -1, 0 e 1.

Exemplos: Shortining_Service{1, -1}, Having_Sub_Domain{-1, 0, 1}, Domain_registeration_length{-1, 1}

Atividade - Dataset de sites de phishing

Objetivo

Dividir o conjunto de dados em duas partes de mesmo tamanho e, em seguida, construir a matriz de confusão usando o 5NN como classificador.

O Dataset de mãos de poker

Descrição

Conjunto de dados com 1025010 instâncias (25009 para treino, todo o resto para teste).
Objetivo: Deseja-se classificar se a mão do jogador possui: nada, um par, dois pares, ...
Atributos: Todos numéricos. Alguns variando de 1 a 13 representando o número da carta e de 1 a 4 para os símbolos.

Atividade - Dataset de mão de poker

Objetivo

Dividir o conjunto de dados em duas partes do mesmo tamanho e, em seguida, construir a matriz de confusão usando 5NN e Random Forest como classificadores.

Considerações finais

Resumo da aula

- KNN possui dificuldades para trabalhar com grandes quantidades de dados
- KNN assim como outros classificadores não apresenta boa performance para problemas com mais de duas classes.
- Ambos os itens anteriores são justificados na acurácia obtida em experimentos

Dúvidas

Muito Obrigado
Rodrigo A. Igawa