

# Aula08

DATA SCIENCE IPT

TURMA 02

## Classificação/Regressão com K-Nearest-Neighbors

A ideia básica do KNN é classificar ou avaliar valores (como na regressão) com base nas amostras “mais próximas”. Para saber quais amostras são “mais próximas”, devemos utilizar alguma métrica. Norma Euclidiana é uma opção.

O “K” do nome refere-se à quantidade de “vizinhos” mais próximos que usaremos na “votação”. Número ímpar é bom para evitar empate. A escolha do K é essencial para boa precisão. O KNN pertence à classe dos lazy learners, já que não há treinamento. Também é não paramétrico, não faz hipótese sobre o modelo e chega nos parâmetros como na regressão linear o faz, por exemplo.

Vamos implementar o KNN para separar doentes e saudáveis com base no dataset novadoenca2.csv (partir do notebook knn.ipynb)

Atividade 1 :

Criar a função distância (Euclidiana) entre duas instâncias...usar produto interno (.dot) ou linalg.norm

Atividade 2 : Criar a função classifica, que recebe as k classificações mais próximas da instância

Atividade 3: Obter a classificação para uma instância qualquer

Tentar acurácia  $> 0.5$  no treinamento

Discussão: como criar uma função mais “sofisticada” eficiente de votação no knn

Vamos agora, partindo do notebook `knn-scikit.ipynb`, tentar acurácia  $>0.5$  em dataset `novadoenca2.csv`.

A extração de features de um texto é tarefa complexa e é uma das bases do NLP (Natural language Processing).

### Bag of Words Model

"Represent each document which the bag of words it contains"

d1 : Mary loves Movies, Cinema and Art

Class 1 : Arts

d2 : John went to the Football game

Class 2 : Sports

d3 : Robert went for the Movie Delicatessen

Class : Arts

	Mary	Loves	Movies	Cinema	Art	John	Went	to	the	Delicatessen	Robert	Football	Game	and	for
d1	1	1	1	1	1									1	
d2						1	1	1	1			1	1		
d3			1				1		1		1				1

É uma grande simplificação, mas pode gerar boas predições...há o problema de lidar com matrizes esparsas

As palavras "utilizadas" na vetorização(?) formam o vocabulário

Intelligent applications creates intelligent business processes

intelligent	applications	creates	business	processes
2	1	1	1	1



Com base no notebook bag of words.ipynb,  
Obtenha as representações (BoW) do dataset opinioes.csv

### Teorema de Thomas Bayes :

$$P(H | e) = \frac{P(e | H) P(H)}{P(e)}$$

Diagram illustrating the components of Bayes' Theorem:

- Posteriori**:  $P(H | e)$
- Likelihood**:  $P(e | H)$
- Priori**:  $P(H)$
- Evidence**:  $P(e)$

The joint probability  $P(H \cap e)$  is indicated above the likelihood and prior terms.

<b>Likelihood</b> How probable is the evidence given that our hypothesis is true?	<b>Prior</b> How probable was our hypothesis before observing the evidence?
$P(H   e) = \frac{P(e   H) P(H)}{P(e)}$	
<b>Posterior</b> How probable is our hypothesis given the observed evidence? (Not directly computable)	<b>Marginal</b> How probable is the new evidence under all possible hypotheses? $P(e) = \sum P(e   H_i) P(H_i)$

$P(H \cap e) = P(H) * P(e)$  para eventos independentes.  
Mostre!

Do Teorema de Thomas Bayes ao Naïve Bayes :

$$P(A | B_1, B_2, \dots, B_n) = \frac{P(B_1, B_2, \dots, B_n | A)P(A)}{P(B_1, B_2, \dots, B_n)}.$$

Se houver independência entre os eventos  $B_1, B_2, \dots, B_n$  (daí o “naive”..ingênuo):

$$P(B_i | A, B_1, B_2, B_{i-1}, B_{i+1}, \dots, B_n) = P(B_i | A)$$

$$\text{Assim, } P(A | B_1, B_2, \dots, B_n) = \frac{p(B_1|A)p(B_2|A)\dots p(B_n|A)P(A)}{P(B_1, B_2, \dots, B_n)}.$$

Estimando a Classe com Naive Bayes

Se tivermos que estimar a classe mais provável entre  $A_1, A_2, \dots$

$$p(A_i | B_1, B_2, \dots, B_n) = \frac{p(B_1 | A_i) p(B_2 | A_i) \dots p(B_n | A_i) P(A_i)}{P(B_1, B_2, \dots, B_n)}.$$

A Classe  $A_i$  mais provável é a que obtém o máximo produto :

$p(B_1 | A_i) p(B_2 | A_i) \dots p(B_n | A_i) P(A_i)$  ..já que todas  $P(A_i)$  são divididas por  $P(B_1, B_2, \dots, B_n)$

# Naïve Bayes

## Conteúdo

**Exemplo :** a sua ida a um jogo do seu time no estádio depende da distância, preço do ingresso, importância do jogo e do clima. Qual é a maior probabilidade ir ou não ir para o caso de importância alta, perto, preço baixo e não chuva...

Amostra	Importante	Perto	Barato	Chuva	Ir
1	0	1	1	0	1
2	0	0	0	0	0
3	1	1	0	1	0
4	1	1	0	0	1
5	0	0	1	0	0
6	1	1	1	1	1

$$P(\text{imp} | \text{Não ir})P(\text{Perto} | \text{Não ir})P(\text{Barato} | \text{Não ir})P(\text{Não Chuva} | \text{Não ir})P(\text{Não ir}) = \\ 1/3 * 1/3 * 1/3 * 2/3 * 3/6 = 0,0123$$

$$P(\text{imp} | \text{Ir})P(\text{Perto} | \text{Ir})P(\text{Barato} | \text{Ir})P(\text{Não Chuva} | \text{Ir})P(\text{ir}) = \\ 2/3 * 3/3 * 2/3 * 2/3 * 3/6 = 0.148$$

Ir : 92% Não ir : 8%

Partindo de Naive Bayes.ipynb e do dataset futebol.txt, chegue no mesmo resultado do slide anterior... Use scikit com Bernoulli

Amostra	Importante	Perto	Barato	Chuva	Ir
1	0	1	1	0	1
2	0	0	0	0	0
3	1	1	0	1	0
4	1	1	0	0	1
5	0	0	1	0	0
6	1	1	1	1	1

Continue o notebook bag of words, crie um modelo multinomial de Naive Bayes e faça previsões (com scikit)

Suponha um conjunto de 10 textos. A palavra “mano” aparece em 8 delas. Seu IDF (inverse document frequency) é :

$IDF = \log 10/8 = 0.22$  ...assim, uma palavra que aparece em muitos dos textos, não é boa “classificadora”..seu IDF é baixo.

TF é “term-frequency”, a frequência do termo em um texto (quantidade).

Por exemplo, na frase : Mano do céu, a polícia tá aí, mano!...

TF de mano é 2 (mano aparece 2 vezes no texto)

IDF de mano (depende de todos os textos) é 0.22

Assim, no texto,  $TF \times IDF$  é  $2 \times 0.22 = 0.44$

Podemos usar o TFIDF como feature de cada palavra e não sua frequência.



Vamos obter (na mão) o tf\*idf para as palavras do primeiro texto do dataset opinioes.csv ...no primeiro comentário

Continue Bag of words.ipynb

Depois, com scikit, processar o dataset e obter a matriz toda de opiniões com tfidf

Partindo de um pedaço (10k amostras) com rating (stars) de estabelecimentos do Yelp, vamos criar um modelo para predição de polaridade (até 3 estrelas negativo, 4 ou 5 positivo). Partir de `yelp-nlp.ipynb`.

Camaradas, vamos lutar por alta acurácia nos testes!



Cursos com Alta Performance de  
Aprendizado

© 2019 – Linked Education