Aula08

DATA SCIENCE IPT

TURMA 02



KNN K-Nearest Neighbors

Conteúdo

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.

KNN K-Nearest Neighbors

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

KNN K-Nearest Neighbors + scikit

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

Representação de textos

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

Bag of Words

Conteúdo

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: Art

	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		T	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

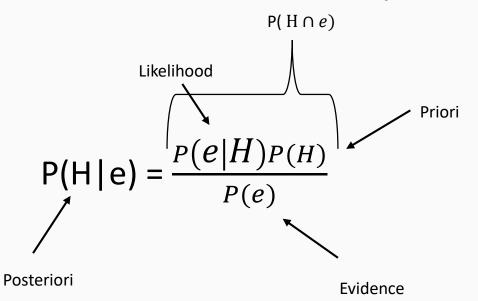
Bag of Words

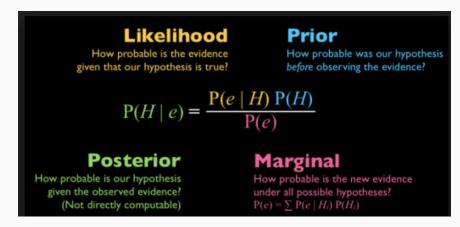
Com base no notebook bag of words.ipynb, Obtenha as representações (BoW) do dataset opinioes.csv



Conteúdo

Teorema de Thomas Bayes :





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

Do Teorema de Thomas Bayes ao Naïve Bayes :

$$P(A | B1,B2,...,Bn) = \frac{P(B1,B2,B,...Bn|A)P(A)}{P(B1,B2,...Bn)}$$

Se houver independência entre os eventos B1,B2...Bn (daí o "naive"..ingênuo):

$$P(Bi \mid A,B1,B2,Bi-1,Bi+1,..Bn)=P(Bi \mid A)$$

Assim,
$$P(A|B1,B2,...,Bn) = \frac{p(B1|A)p(B2|A)...p(Bn|A)P(A)}{P(B1,B2,...Bn)}$$

Estimando a Classe com Naive Bayes

Se tivermos que estimar a classe mais provável entre A1,A2...

$$P(Ai | B1,B2,...,Bn) = \frac{p(B1|Ai)p(B2|Ai)...p(Bn|Ai)P(Ai)}{P(B1,B2,...Bn)}$$

A Classe Ai mais provável é a que obtém o máximo produto : $p(B1|Ai)p(B2|Ai)\dots p(Bn|Ai)P(Ai)$...já que todas P(Ai) são divididas por P(B1,B2..Bn)

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 chuya...

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

P(imp|Ir)P(Perto|Ir)P(Barato|Ir)P(Não Chuva|Ir)P(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

Conteúdo

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

Suponha um conjunto de 10 textos. A palavra "mano" aparece em 8 delas. Seu IDF (inverse document frequecy) 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, TFxIDF é 2x0.22=0.44

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

TF*IDF

Conteúdo

Vamos obter (na mão) o tf*idf para a palavras serviço 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

Desafio: Modelo para polaridade

Conteúdo

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