How a robot can use basic probability theory to classify things:

The basics of statitical classification theory

Georgia Tech

PROJECT: Text-classification: an application

We are going to classify test into topics. We assume a random model for generating documents.

- Given the topic the words are independent.
- We do not consider the order the words appear in sentence nor how often they appear. Only if they appear or not.

Text-classification: the word $\times$ document matrix

We consider the document $\times$ word matrix X which records which words appear in which document:

		dog	cat	car	home	run	be	nice
Document1		1	0	0	1	1	0	0
Document2		0	1	0	1	0	1	0
Document3		1	0	0	0	0	1	0
Document4		0	0	1	1	0	1	0
Document5		0	0	1	0	1	0	1
Document6		0	0	1	0	0	1	1

(12)

One topic for each document

we consider two topics:

$$T0 = \texttt{animal life}, T1 = \texttt{home and car}$$

Topic			dog	cat	car	home	run	be	nice
$T0$	Document1		1	0	0	1	1	0	0
$T0$	Document2		0	1	0	1	0	1	0
$T0$	Document3		1	0	0	0	0	1	0
$T1$	Document4		0	0	1	1	0	1	0
$T1$	Document5		0	0	1	0	1	0	1
$T1$	Document6		0	0	1	0	0	1	1

(13)

We assume independence of words given topic

First document :

$$\texttt{document 1} = (dog, run, home)$$

has probability under topic 0:

$$P(dog, run, home|\texttt{topic 0}) = P(dog|topic0) \cdot P(run|topic0) \cdot P(home|topic0)$$

and under topic 1:

$$P(dog, run, home|\texttt{topic 1}) = P(dog|topic1) \cdot P(run|topic1) \cdot P(home|topic1)$$

Conditional probabilites of words given topic

Let the conditional probabilities under topic 0 be:

$$\vec{p} = (P(dog|T0), P(cat|T0), P(car|T0), P(home|T0), P(run|T0), P(be|T0), P(nice|T0))$$

Let the conditional probabiities under topic 1 be:

$$\vec{q}(P(dog|T1), P(cat|T1), P(car|T1), P(home|T1), P(run|T1), P(be|T1), P(nice|T1))$$

Bayse classification

Given document (dog, run, cat) compare the probabilites and chose the one with higher probability as the topic. IF

$$\pi_0 \cdot P(dog|T0) \cdot P(run|T0) \cdot P(cat|T0) > \pi_1 \cdot P(dog|T1) \cdot P(run|T1) \cdot P(cat|T1)$$

we classifiy as topic 0. Here π_0 is probability of topic 0 and π_1 is probability of topic 1.

Example of estimation of condition probabilities

Topic			dog	cat	car	home	run	be	nice
$T0$	Document1		1	0	0	1	1	0	0
$T0$	Document2		0	1	0	1	0	1	0
$T0$	Document3		1	0	0	0	0	1	0
$T1$	Document4		0	0	1	1	0	1	0
$T1$	Document5		0	0	1	0	1	0	1
$T1$	Document6		0	0	1	0	0	1	1

(14)

Estimated conditional probabilites given T0:

$$\hat{\vec{p}} = (\frac{2}{3}, \frac{1}{3}, 0, \frac{2}{3}, \frac{1}{3}, \frac{2}{3}, 0)$$

Estiamted conditional probabilites given T1:

$$\hat{\vec{q}} = (0, 0, 1, \frac{1}{3}, \frac{1}{3}, \frac{2}{3}, \frac{2}{3})$$

Summary text classficiation naive bayse

- Estimate the coditional probabilities $\vec{p}$ and $\vec{q}$.
- Use the estimates $\hat{\vec{p}}$ and $\hat{\vec{q}}$ to estimate the probabilities of a new document which you which to classify.
- Chose the topic classficiation which has higher estimated probability.