DMML, 4 Feb 2019

Supervised learning

Decision Tree

Regression

Logistic Regression

Class Association Rules

Probabilistic prediction

Bayes rule for conditional probabilities

$$P(A \mid B) = \frac{P(A \wedge B)}{P(B)}$$

$$P(A|B) = \frac{P(A \wedge B)}{P(B)}$$

$$P(B|A) = \frac{P(B \wedge A)}{P(A)} = \frac{P(A \wedge B)}{P(A)}$$

$$P(B|A) = \frac{P(A|B) \cdot P(B)}{P(A)}$$

$A_1, A_2, \ldots, A_k \quad , \quad C$

By simple counting from given training data

$$P(A_1=a_1, A_2=a_2, \ldots, A_k=a_k \mid C=c)$$

When we see $(a_1', a_2', \ldots, a_k')$,

what is the category value for C?

Compare $P(C=c_1 \mid a_1', \ldots, a_k')$

$P(C=c_2 \mid a_1', \ldots, a_k')$

Use Bayes' rule to calculate

A	B	C
m	b	t
m	s	t
g	q	t
h	s	t
g	q	t
g	q	f
g	s	f
h	b	f
h	q	f
m	b	f

$$P(X|Y) = \frac{P(Y|X)\,P(X)}{P(Y)}$$

$$P(C=t) = \frac{5}{10}$$

$$P(A=m, B=b) = \frac{2}{10}$$

$$P(A=m, B=b \mid C=t) = \frac{1}{5}$$

$$P(A=m, B=b \mid C=f) = \frac{1}{5}$$

What is this capturing ?

We assume a probabilistic model for <u>generating</u> the data

Unless you can describe a generative model, don't claim to do probabilistic analysis

A	B	C
m	b	t
m	s	t
g	q	t
h	s	t
g	q	t
g	q	f
g	s	f
h	b	f
h	q	f
m	b	f

Here, a plausible model is:

1. Randomly chose $C = t/f$ with $P(C)$
2. Given C, choose attributes using $P(A, B \mid C)$

Given this generative model, reverse engineer the probabilities from the data

— Parameter estimation

Digression

Given 7 heads in 10 tosses, we estimate P(heads) as $\frac{7}{10}$ Why?

Suppose P(heads) is some unknown value p

Probability of observed outcome : $\alpha \times p^7 (1-p)^3$

What value of p maximizes this probability

$= 7/10$

Maximum Likelihood Estimate — MLE

Observation $\longrightarrow$ Parameter Θ

$\theta \longrightarrow$ likelihood $L(\theta)$ of observation

$$\arg\max_{\theta} L(\theta)$$

Back to Bayesian classification

A	B	C
m	b ✓	t
m	s	t
g ✓	q	t
h	s	t
g ✓	q	t
g ✓	q	f
g ✓	s	f
h	b ✓	f
h	q	f
m	b ✓	f

A=g occurs 4 times

B=b occurs 3 times

$\{A=g, B=b\}$?

$P(Y)=0$

$$P(C=t \mid A=g, B=b) = \frac{P(g,b \mid t) \cdot P(t)}{P(g,b)}$$

$$P(C=f \mid A=g, B=b) = \frac{P(g,b \mid f) \cdot P(f)}{P(g,b)}$$

Naïve Bayes assumption

Attributes are independent

$$P(A \land B) = P(A) \cdot P(B)$$

$$P(A, B \mid C) = P(A \mid C) \cdot P(B \mid C)$$ Conditional independ.

Estimate $P(A \mid C)$, $P(B \mid C)$ ... from data

A	B	C
m	b	t
m	s	t
g	q	t
h	s	t
g	q	t
g	q	f
g	s	f
h	b	f
h	q	f
m	b	f

Can we estimate a class for $(A=g,\ B=b)$

$P(g|t) = \frac{2}{5}$ $\quad P(b|t) = \frac{1}{5}$

$P(g|f) = \frac{2}{5}$ $\quad P(b|f) = \frac{2}{5}$

$$P(t|g,b) = \frac{P(g|t)P(b|t)P(t)^2}{\cancel{P(b)P(g)}}$$

$P(g) = \frac{4}{10}$ $\quad P(b) = \frac{3}{10}$

$P(t) = P(f) = \frac{5}{10}$

$$P(f|g,b) = \frac{P(g|f)P(b|f)P(f)^2}{\cancel{P(b)P(g)}}$$

$\Rightarrow 2 \cdot P(t|g,b)$

$\therefore$ answer is f.

No theoretical justification for naive Bayes assumption — works well "in practice"

Very easy to implement — build a model, one pass over data updating counts

One very popular and successful use case is spam filters for emails

Application Text classification

Spam, or topic classification

Document Model?

Vocabulary V of "known" words

Simplest model – set of words

Boolean model – each $w \in V$ is either present or absent

Generative model

1. Choose a topic $t_i \in T = \{t_1, \ldots, t_m\}$
   - Equally likely? Historical ratios?
   - $P(t_i)$
2. Given t_i, for each $w \in V$, include w with probability $P(w \mid t_i)$
   - Like tossing one coin per word

Given a training set of labelled documents

N documents

N_i documents of topic t_i, for each i

$$P(t_i) = \frac{N_i}{N}$$

For each topic t_i, for each $w \in V$

Among N_i documents, W_i documents contain the word w

$$P(w \mid t_i) = \frac{W_i}{N_i}$$

Given a document d — some subset of V

$$P(T = t_j \mid d)$$

$d = d_1 \ldots d_M$ to a 0-1 vector, $|V| = M$

$$\frac{\prod_{i=1}^{M} P(d_i \mid t_j) \cdot P(t_j)}{\prod_{i=1}^{M} P(d_i)}$$

$\leftarrow$ same for all t_j

Back to zero counts

Full Bayes $\longrightarrow$ Naive Bayes

May still have rare attribute values.

Suppose 'zebra' is not in training data

$P(\text{'zebra'} \mid t_i) = 0$ for all t_i

$P(t_i \mid \text{-- 'zebra'}) = \text{--} \, P(\text{'zebra'} \mid t_i) \ldots$

All numerators are 0 !

"Smoothing" — Laplace

n_v = # of times we see v

n = total sample

$$P(v) = \frac{n_v}{n}$$

Suppose values are $\{v_1, \ldots, v_m\}$

$$\frac{n_v}{n} \Rightarrow \frac{n_v + 1}{n + m}$$

Variations

$$\frac{n_v + \lambda}{n + \lambda m}$$

$\lambda = \frac{1}{n}$ is used

Tomorrow – a slightly more sophisticated document model