DMML, 4 Feb 2020

Classification: Class Association Rules, Decision Trees, logistic regression

Probabilistic Approach

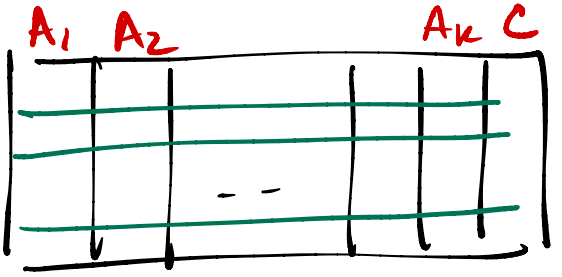
Attributes $a_1, a_2, .., a_k$, predict a class $c$

$A_1$ $A_2$ $A_k$ $\quad c \in C$

$$P(C = c \mid A_1 = a_1 \wedge A_2 = a_2 \wedge .. \wedge A_k = a_k)$$

Among all $c \in C$, choose the one with highest prob.

Training data - table




Use frequencies to estimate probabilities

Fix a class $c \in C$ — subset of rows

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

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

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

$$P(C=c \mid A_1=a_1, \ldots, A_k=a_k) = \frac{P(A_1=a_1, \ldots, A_k=a_k \mid C=c) \cdot P(C=c)}{P(A_1=a_1, \ldots, A_k=a_k)}$$

$P(C=c)$ = fraction of rows with $C=c$

$P(A_1=a_1, \ldots, A_k=a_k)$ = fraction of rows with this combination of value

## Generative models

Pick $c \in C$ with probability $P(C=c)$

Given $c \in C$, choose $a_1, \ldots, a_k$ with $P(A_i=a_i \mid C=c)$

# Estimating parameters of generative model

Why we interpret frequency ratio as a probability?

6 Heads in 10 tosses $\longrightarrow$ $P(\text{Head}) = 0.6$

Why not 0.55?

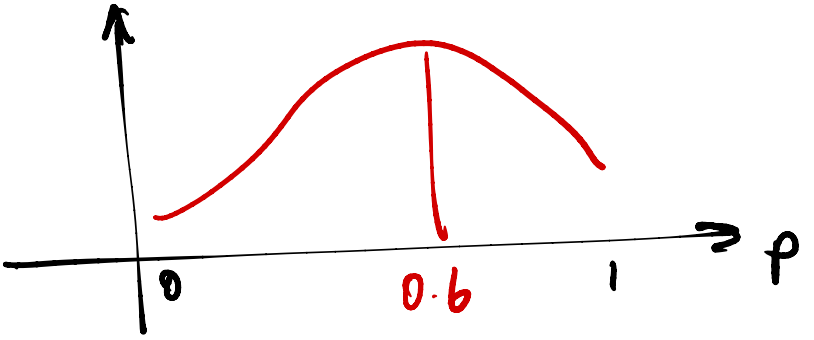
Observation $\longrightarrow$ Parameter Estimate

probability of observation

$[6H, 4T] \longrightarrow P(\text{Heads}) \in [0, 1]$

$P(\text{Heads}) = p \rightsquigarrow \alpha \cdot p^6 (1-p)^4$ is the probability of seeing 6H, 4T

**LIKELIHOOD**




Maximum Likelihood Estimator

MLE

| 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 |

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

$C=f$

$P(C=t) = P(C=f) = \frac{1}{2}$

$P(A=g, B=q) = \frac{3}{10}$

$P(A=g, B=q \mid C=t) = \frac{2}{5}$

$C=f \quad = \frac{1}{5}$

If $A=g, B=q$, what is $C$?

$P(C=t \mid A=g, B=q) = P(A=g, B=q \mid C=t) \cdot P(C=t) / P(g,q)$

$C=f$ $\qquad \frac{2}{5}$ $\qquad \frac{1}{2}$

$\searrow 1/5 \qquad | C=f \quad P(C=f) / P(g,q)$

| 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 |

$P(C=t \mid A=g, B=b)$ ?

$0 = P(A=g, B=b)$

Simplifying assumption = attributes are independent

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

Naive Bayes Classifier

Unjustifiable in theory

$P(A=g, B=b) = P(A=a) \cdot P(B=b)$

$P(A=g, B=b \mid C=t) = P(A=a \mid C=t) \cdot P(b \mid t)$

## Typical Example

Text classification <= Topic assignment
                    <= Junk Mail

Simplifying assumption

- Document is just a collection of words
  - Order is unimportant
- Document is just a set of words

Assume a vocabulary $V = (w_1, w_2, \ldots, w_N)$ of relevant words

Each document is a subset of V

0-1 vector of length N

| $W_1$ | $W_2$ | ---- | $W_N$ |
|---|---|---|---|
| 0 | 1 | 0 | 1 1 -- 1 |

N - attributes (boolean) + a category

| | $W_1$ | $W_2$ | — — — | $W_N$ | C |
|---|---|---|---|---|---|
| $d_1$ | 0 | 1 | | 0 | J |
| $d_2$ | 1 | 0 | — — — | 1 | NJ |

$P(C=c)$ , $P(W_1|C) \cdot P(W_2|C) \cdots P(W_N|C)$

represents $P(W_1, W_2 \cdots, W_N|C)$

Generating model

Choose $c$ with $P(C=c)$

For each $w \in V$, include $w$ with $P(w|c)$

---

Independence solves the problem of missing combinations of attributes

Still could have some data values that do not appear

If $A_i = v$ never appears in our table

$P(v|c)$, $P(v)$ etc are 0

$$P(c|v_1, v_2) = \frac{P(v_1|c) \cdot P(v_2|c) \cdot P(c)}{P(v_1)\,P(v_2)}$$

($P(v_1|c) = 0$)

$P(v_1) = 0$?

No meaningful prediction is possible

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

$n_v$ ← # of items where $v$ occurs

$n$ ← total # of items

If $n_v = 0$, we have a problem

Solution due to Laplace

$V$ takes values $v_1, v_2, \ldots, v_m$

Observations $o_1, o_2, \ldots, o_n + v_1, v_2, \ldots, v_m$

$$P(v_i) = \frac{n_{v_i} + 1}{n + m}$$

Laplace Smoothing

Can generalize this to $\frac{n_v + \lambda}{n + \lambda m}$

$\lambda = 1$ : Laplace

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

---

A more general document model

Count number of occurrences of each word

Length of document comes into play

"Bag" of words model – Bag = Multiset = Set with multiplicities