Naïve Bayes
some problems for estimation

Akito Sakurai

Contents

- MAP and MLE
- "Frequency=0" problem
  – Parameter estimation of binomial distribution
  – Laplace correction
  – In R

MAP estimation

$$P(h \mid D) = \frac{P(D \mid h)\, P(h)}{P(D)}$$

When data is given, a hypothesis with highest posterior is the one to be selected.
Maximum a posteriori hypothesis h_{MAP}:

$$\begin{aligned} h_{MAP} &= \arg\max_{h \in H} P(h \mid D) \\ &= \arg\max_{h \in H} \frac{P(D \mid h)\, P(h)}{P(D)} \\ &= \arg\max_{h \in H} P(D \mid h)\, P(h) \end{aligned}$$

ML estimation

Maximum Likelihood hypothesis h_{ML}:

$$h_{ML} = \underset{h \in H}{\mathrm{argmax}} P(D|h)$$

which is equivalent to MAP with $P(h_i) = P(h_j)\ \forall i, j$

$$h_{MAP} = \underset{h \in H}{\mathrm{argmax}} P(D|h)P(h)$$

Contents

- MAP and MLE
- "Frequency=0" problem
  – Parameter estimation of binomial distribution
  – Laplace correction
  – In R

Laplace correction in a nutshell

Frequency=0 problem

- What will happen if, for a class and an attribute value, there is no occurrence? (e.g. for "Play=No" and "Outlook = Overcast")
  – Prob(Outlook=Overcast | Play=no) is equal to 0 !!
- Therefore its posterior probability = 0
  – Although all the other attributes say the sample is very likely, it is 0.
- A remedy:
  – Add 1 to the frequency of all combinations of class and attribute values (called Laplace correction);
  – Note that the denominators (see the righthand side) must be increased by k (# of class values).

P(Play=yes | E)
= P(Outlook=Sunny | Play=yes) *
P(Temp=Cool | Play=yes) *
P(Humidity=High | Play=yes) *
P(Windy=True | Play=yes) *
P(play=yes) / P(E)
= (2/9) * (3/9) * (3/9) * (3/9) *(9/14) / P(E)
= 0.0053 / P(E)
would be

# of values of 'Outlook'

= ((2+1)/(9+3)) * ((3+1)/(9+3)) *
((3+1)/(9+2)) * ((3+1)/(9+2)) *(9/14) /
P(E)
= 0.007 / P(E)

'Windy' のとりうる値の個数

Contents

- MAP and MLE
- "Frequency=0" problem
  - – Parameter estimation of binomial distribution
  - – Laplace correction
  - – In R

addendum

Estimation of parameters (binomial distribution)

- Toss a thumbtack in the air and it will land with the point up (H: head) or not up (T: tail).
- Set θ = *P* (*H*) which is unknown

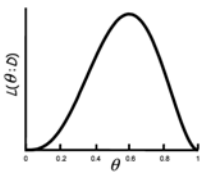
Estimation problem:
From the results of tossing D=x[1],x[2],…,x[M], we want to estimate the probability *P*(*H*)= θ and *P*(*T*)=1- θ .

Likelihood function

- How to measure goodness of estimation of θ?
  Use the likelihood of a hypothesis "assuming θ to be true, the data were generated by the hypothesis"

$$L(\theta : D) = P(D \mid \theta) = \prod_m P(x[m] \mid \theta)$$

- E.g., for a sequence H,T,T,H,H :



$$L(\theta : D) = \theta \cdot (1-\theta) \cdot (1-\theta) \cdot \theta \cdot \theta$$

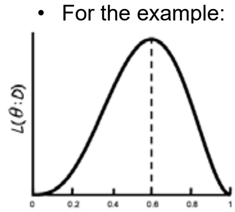
MLE: Maximum likelihood estimator

- The method of :

<u>Get a parameter value that maximizes its likelihood</u>

- For the example:

$$\hat{\theta} = \frac{N_H}{N_H + N_T}$$



We may claim that this is the best we can.

Derivation: starts from "is MLE best?"

- What if the number of observations is very large?
  - – Same principle is applied. But, there may be many unobserved attribute values. Should the estimate of probability be 0? Could we think that we happen to have no occurrences? …
- Suppose that the variable X take 20 values with equal probability and we have 30 samples. Then there may be many values unobserved.

The problem is sever

```
> set.seed(100)
> x <- sample(1:20, 30, replace=T)
> table(x)
x
 2  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18
 1  2  1  3  1  4  1  2  4  1  1  2  2  2  1  2
```


MAP solution

- We suppose that "parameters distribute."
  - – Bayesian way.
  - – Maximize, not P(D|θ) , but P(D|θ)P(θ)
  - – By using Bayes theorem:

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$$

$$P(\theta|N_H, N_T) = \frac{\theta^{N_H}(1-\theta)^{N_T}P(\theta)}{P(D)}$$

i.e., MAP rather than MLE

- – We will take $\hat{\theta}$ that maximize the above probability.
- – In this case, in general, $\theta = 0$ is not necessarily a solution even if $N_H = 0$.

Prior distribution?

- Beta distribution as a prior:
  $P_B(x;\alpha,\beta)=x^{\alpha-1}(1-x)^{\beta-1}/B(\alpha,\beta)$
- Since
  $\int \theta^{N_H}(1-\theta)^{N_T}P_B(\theta;\alpha,\beta)d\theta = \theta^{N_H+\alpha}(1-\theta)^{N_T+\beta}$,
  $\hat{\theta} = (N_H+\alpha)/((N_H+\alpha)+(N_T+\beta))$
  is obtained.
- Substituting 1's for α and β, we get Laplace correction.
  $\hat{\theta} = (N_H+1)/((N_H+1)+(N_T+1))$

Contents

- MAP and MLE
- "Frequency=0" problem
  - Parameter estimation of binomial distribution
  - Laplace correction
  - In R

Laplace correction

- What happens when $N_H = 0$?
- Laplace correction is an answer:
  - Suppose you had observed, before real observations, two tossing:
    one head and one tail landed.
  - then the estimation is:
    $\hat{\theta} = (N_H+1)/((N_H+1)+(N_T+1))$

Extension of Laplace correction

- It is obvious that we can extend Laplace correction as
  $\hat{\theta} = (N_H+\alpha)/((N_H+\alpha)+(N_T+\beta))$
- We can assign 1/2, 1, 2,.. to α and β and/or different values to them.
- There are no theories some of which is better than the others.

Before/after Laplace correction

A1=Outlook			A2=Temperature			A3=Humidity			A4=Windy			m=Play	
	Yes	No		Yes	No		Yes	No		Yes	No	Yes	No
Sunny	2	3	Hot	0	2	High	3	4	False	6	2	9	5
Overcast	4	0	Mild	6	2	Normal	6	1	True	3	3		
Rainy	3	2	Cool	3	1								
Sunny	2/9	3/5	Hot	0/9	2/5	High	3/9	4/5	False	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	Mild	6/9	2/5	Normal	6/9	1/5	True	3/9	3/5		
Rainy	3/9	2/5	Cool	3/9	1/5								

改変後の値です

A1=Outlook			A2=Temperature			A3=Humidity			A4=Windy			m=Play	
	Yes	No		Yes	No		Yes	No		Yes	No	Yes	No
Sunny	3	4	Hot	1	3	High	4	5	False	7	3	10	6
Overcast	5	1	Mild	7	3	Normal	7	2	True	4	4		
Rainy	4	3	Cool	4	2								
Sunny	3/12	4/8	Hot	1/12	3/8	High	4/11	5/7	False	7/11	3/7	9/14	5/14
Overcast	5/12	1/8	Mild	7/12	3/8	Normal	7/11	2/7	True	4/11	4/7		
Rainy	4/12	3/8	Cool	4/12	2/8								

Contents

- MAP and MLE
- "Frequency=0" problem
  - Parameter estimation of binomial distribution
  - Laplace correction
  - In R

Laplace correction: in R

```
m <- naiveBayes(xy[,-5], xy[,5],laplace=1)
```

```
> m <- naiveBayes(xy[,-5], xy[,5],laplace=1)
> table(xy[,5],predict(m,xy[,-5]))

      No Yes
  No   4   1
  Yes  0   9
> m

Naive Bayes Classifier for Discrete Predictors

Call:
naiveBayes.default(x = xy[, -5], y = xy[, 5], laplace = 1)
```

```
A-priori probabilities:
xy[, 5]
       No       Yes
0.3571429 0.6428571

Conditional probabilities:
       Outlook
xy[, 5]  Overcast     Rainy     Sunny
    No  0.1250000 0.3750000 0.5000000
    Yes 0.4166667 0.3333333 0.2500000

       Temp.
xy[, 5]      Cool       Hot      Mild
    No  0.2500000 0.3750000 0.3750000
    Yes 0.3333333 0.2500000 0.4166667

       Humidity
xy[, 5]      High    Normal
    No  0.7142857 0.2857143
    Yes 0.3636364 0.6363636

       Windy
xy[, 5]     False      True
    No  0.4285714 0.5714286
    Yes 0.6363636 0.3636364
```

Laplace correction: in R

Locate differences. Too easy to do.

```
> library(e1071)
> setwd("D:/R/Sample")
> xy<-read.csv("PlayTennis.csv", header=TRUE)
> m <- naiveBayes(xy[,-5], xy[,5])
> table(xy[,5],predict(m,xy[,-5]))

      No Yes
  No   4   1
  Yes  0   9
> m

Naive Bayes Classifier for Discrete Predictors

Call:
naiveBayes.default(x = xy[, -5], y = xy[, 5])
```

```
A-priori probabilities:
xy[, 5]
       No       Yes
0.3571429 0.6428571

Conditional probabilities:
       Outlook
xy[, 5]  Overcast     Rainy     Sunny
    No  0.0000000 0.4000000 0.6000000
    Yes 0.4444444 0.3333333 0.2222222

       Temp.
xy[, 5]      Cool       Hot      Mild
    No  0.2000000 0.4000000 0.4000000
    Yes 0.3333333 0.2222222 0.4444444

       Humidity
xy[, 5]      High    Normal
    No  0.8000000 0.2000000
    Yes 0.3333333 0.6666667

       Windy
xy[, 5]     False      True
    No  0.4000000 0.6000000
    Yes 0.6666667 0.3333333
```