Chapter 8

김선
서울대학교 컴퓨터 공학부
생물정보 연구소

8

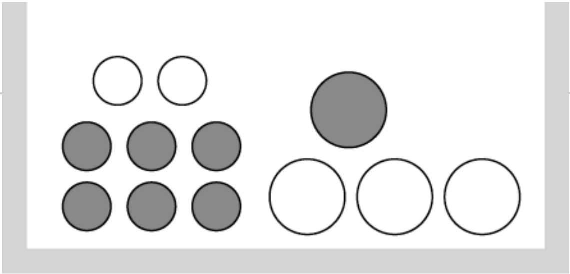
베이스 추정
(Bayesian Estimation)

김선
서울대학교 컴퓨터 공학부
생물정보 연구소

Bayes Theorem

- $P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$
  - $P(Y)$: Prior probability.
    - Probability of event before evidence is seen.
  - $P(Y|X)$: Posterior probability given X.
    - Probability of event after evidence is seen.
  - $P(X|Y)$: Likelihood.

Example – p.212



- There exist balls like above image. Assume that you randomly pick a single ball.
  - 1. Probability of black ball.
    - $P(Black) = \frac{7}{12}$
  - 2. Probability of black ball if we know the ball is large.
    - $P(Black|Large) = \frac{1}{4}$
  - 3. Probability of black and large ball.
    - $P(Black, Large) = \frac{1}{12}$
    - $P(Black, Large) = P(Black|Large)P(Large) = \frac{1}{4} * \frac{4}{12} = \frac{1}{12}$

Example2 – p.216

- There is a virus that generally infect 1% of people. Also, the accuracy of virus inspection is 95%. If the result of your inspection is positive, calculate the probability that you are indeed infected with virus.
  - From statements, we know
    - $P(V) = 0.01, P(\sim V) = 0.99$
    - $P(P|V) = P(N|\sim V) = 0.95$
    - $P(N|V) = P(P|\sim V) = 0.05$
  - With marginalize out
    - $P(P) = \sum_{\acute{V}\in\{V,\sim V\}} P(P,\acute{V}) = \sum_{\acute{V}\in\{V,\sim V\}} P(P|\acute{V})P(\acute{V}) = 0.95 * 0.01 + 0.05 * 0.99 = 0.059$
  - What we want to know is
    - $P(V|P) = \frac{P(P|V)P(V)}{P(P)} = \frac{0.95*0.01}{0.059} \cong 0.16$

Bayesian Estimation

- Directly calculate probability of parameters.
- Example p.219
  - Give N data points $\mathrm{t} = \{t_n\}_{n=1}^{N}$, find probability of parameter(μ) if the data is sampled from Gaussian distribution.
  - $N(t_n|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(t_n-\mu)^2}{2\sigma^2}}$
  - How to set prior probability $P(\mu)$?

Bayesian Estimation

- Conjugate Prior (켤레사전분포)
  - Prior probability distribution which makes prior and posterior to be same distribution family.
  - For example, Gaussian family is conjugate to Gaussian family. This is the reason why p.221 sets $P(\mu)$ as Gaussian distribution.
    - For more examples, see https://en.wikipedia.org/wiki/Conjugate_prior

Bayesian Estimation

Bayesian Estimation

- $P(\mu) = N(\mu|\mu_0, \sigma_0^2)$
- Then, $P(\mu|t) = N(\mu|\mu_N, \beta_N^{-1})$: another Gaussian distribution.
  - See p.224 for detail notations.
  - $P(\mu|t)$ means the probability of mean value for given data points.
- How to calculate probability of new data $P(x)$ after estimation?
  - => marginalize out
  - $P(x|t) = \int P(x|t,\mu)P(\mu|t)d\mu = \int P(x|\mu)P(\mu|t)d\mu = \int N(\mu|\mu_N, \beta_N^{-1})N(x|\mu,\sigma^2)d\mu = \underline{N(x|\mu_N, \beta^{-1} + \beta_N^{-1})}$

Previous data does not affect new data

Eq 8.49 of page 228

ML vs MAP

- MAP is used to estimate best parameter from Bayesian estimation.
  - Maximum likelihood
    - Maximize likelihood of model
    - $arg\max_\theta P(D|\theta)$
  - Maximum a posteriori
    - Maximize posterior probability
    - $arg\max_\theta P(\theta|D) = arg\max_\theta P(D|\theta)\, P(\theta)$

나이브 베이즈 분류
(Naïve Bayesian)

Bayes Classifier

Problem setting

- Data: $D = \{(X, Y)^n\}_{n=1}^N$
- Input features: $\mathrm{X} = (x_1, .., x_k)$
- Output: Y ∈ {1, …, L}
- Hypothesis: $\underset{Y}{\operatorname{argmax}} P(Y|X) = \widehat{Y}$

Bayes Model

- $\hat{Y} = \underset{Y}{\operatorname{argmax}} P(Y|X)$
- $P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} = \frac{P(X|Y)P(Y)}{\sum P(X|Y)P(Y)}$
- $P(X|Y) = P(x_1, .., x_k|Y)$

Bayes Classifier

Problem setting

- Data: $D = \{(X,Y)^n\}_{n=1}^N$
- Input features: $\mathrm{X} = (x_1, .., x_k)$
- Output: Y ∈ {1, …, L}
- Hypothesis: $\underset{Y}{\operatorname{argmax}}\, P(Y|X) = \hat{Y}$

Bayes Model

- $\hat{Y} = \underset{Y}{\operatorname{argmax}}\, P(Y|X)$
- $P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} = \frac{P(X|Y)P(Y)}{\sum P(X|Y)P(Y)}$
- $P(X|Y) = P(x_1, .., x_k|Y)$
  - Problems
    - In sufficient data as it should consider all possible combinations of x.
    - Lack of space and time.

Naïve Bayes

Problem setting

- Data: $D = \{(X, Y)^n\}_{n=1}^N$
- Input features: $X = (x_1, .., x_k)$
- Output: $Y \in \{1, \dots, L\}$
- Hypothesis: $\underset{Y}{\operatorname{argmax}} P(Y|X) = \widehat{Y}$

Naïve Bayes Model

- $\hat{Y} = \underset{Y}{\operatorname{argmax}} P(Y|X)$
- $P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} = \frac{P(X|Y)P(Y)}{\sum P(X|Y)P(Y)}$
- $P(X|Y) = P(x_1, .., x_k|Y) =$ $\prod_i P(x_i|Y)$
  - Assuming that all features are independent.

Naïve Bayes

- When attributes are discrete values.
  - Calculate all $P(x_i|Y)$ and $P(Y)$ by counting.
  - To prevent 0 probability, use pseudo-count if data is insufficient.
- Exercise) Play tennis

Naïve Bayes

Example

PlayTennis: training examples

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

- Learning Phase

Outlook	Play=*Yes*	Play=*No*
Sunny	2/9	3/5
Overcast	4/9	0/5
Rain	3/9	2/5

Temperature	Play=*Yes*	Play=*No*
Hot	2/9	2/5
Mild	4/9	2/5
Cool	3/9	1/5

Humidity	Play=*Yes*	Play=*No*
High	3/9	4/5
Normal	6/9	1/5

Wind	Play=*Yes*	Play=*No*
Strong	3/9	3/5
Weak	6/9	2/5

$P(\text{Play}=Yes) = 9/14$

$P(\text{Play}=No) = 5/14$

Slide from: https://fenix.tecnico.ulisboa.pt/downloadFile/563568428721191/Naive-Bayes.pdf

Exercise) Play tennis

- For new weather $x = (Sunny, Cool, High, Strong)$, which is more probable (playing or not)?
- Answer is in next slide…

Naïve Bayes

- Exercise) Play tennis
  - For new weather $x = (Sunny, Cool, High, Strong)$, which is more probable (playing or not)?
  - $P(Yes|x) = \frac{P(x|Yes)P(Yes)}{p(x)} = \frac{P(Sunny|Yes)P(Cool|Yes)P(High|Yes)P(Strong|Yes)P(Yes)}{p(x)} = \frac{0.0053}{p(x)}$
  - $P(No|x) = \frac{P(x|No)P(No)}{p(x)} = \frac{P(Sunny|No)P(Cool|No)P(High|No)P(Strong|No)P(No)}{p(x)} = \frac{0.0206}{p(x)}$
  - $\operatorname*{argmax}_Y P(Y|X) = No$, since $P(No|x)$ is bigger than $P(Yes|x)$.