SNS COLLEGE OF TECHNOLOGY

Coimbatore-35
An Autonomous Institution
Accredited by NBA – AICTE and Accredited by NAAC – UGC with 'A++' Grade
Approved by AICTE, New Delhi & Affiliated to Anna University, Chennai

DEPARTMENT OF INFORMATION TECHNOLOGY

19CSE303 – ARTIFICIAL INTELLIGENCE

III YEAR IV SEM

UNIT V – **LEARNING**

TOPIC : Statistical Learning

Statistical Approaches

- Statistical Learning
- Naïve Bayes
- Instance-based Learning
- Neural Networks

Example: Candy Bags

- Candy comes in two flavors: cherry (☺) and lime (☹)
- Candy is wrapped, can’t tell which flavor until opened
- There are 5 kinds of bags of candy:
  - H_1= all cherry (P(c|h1) = 1, P(l|h1) = 0)
  - H_2= 75% cherry, 25% lime
  - H_3= 50% cherry, 50% lime
  - H_4= 25% cherry, 75% lime
  - H_5= 100% lime
- Given a new bag of candy, predict H
- Observations: D_1, D_2 , D_3, …

Bayesian Learning

- Calculate the probability of each hypothesis, given the data, and make prediction weighted by this probability (i.e. use all the hypothesis, not just the single best)

$$P(h_i \mid d) = \frac{P(d|h_i)P(h_i)}{P(d)} = \alpha P(d \mid h_i)P(h_i)$$

- Now, if we want to predict some unknown quantity X

$$P(X \mid d) = \sum_i P(X \mid h_i)P(h_i \mid d)$$

Bayesian Learning cont.

- **Calculating P(h|d)**

$$P(h_i \mid d) \propto \underset{\substack{\uparrow \\ \textbf{likelihood}}}{P(d \mid h_i)}\,\underset{\substack{\uparrow \\ \textbf{prior}}}{P(h_i)}$$

- **Assume the observations are i.i.d.—independent and identically distributed**

$$P(d \mid h_i) = \prod_j P(d_j \mid h_i)$$

Example:

- Hypothesis Prior over h_1, ..., h_5 is {0.1,0.2,0.4,0.2,0.1}
- Data:

- Q1: After seeing d_1, what is $P(h_i|d_1)$?
- Q2: After seeing d_1, what is P(d_2= 🟢 |d_1)?

Making Statistical Inferences

- Bayesian –
  - predictions made using all hypothesis, weighted by their probabilities

- MAP – maximum a posteriori
  - uses the single most probable hypothesis to make prediction
  - often much easier than Bayesian; as we get more and more data, closer to Bayesian optimal

- ML – maximum likelihood
  - assume uniform prior over H
  - when

Naïve Bayes (20.2)

Naïve Bayes

- aka Idiot Bayes
- particularly simple BN
- makes overly strong independence assumptions
- but works surprisingly well in practice…

Bayesian Diagnosis

- suppose we want to make a diagnosis D and there are n possible mutually exclusive diagnosis d_1, …, d_n
- suppose there are m boolean symptoms, E_1, …, E_m

$$P(d_i \mid e_1, \ldots, e_m) = \frac{P(d_i)P(e_1, \ldots, e_m \mid d_i)}{P(e_1, \ldots, e_m)}$$

how do we make diagnosis?

we need: $P(d_i)$ $P(e_1, \ldots, e_m \mid d_i)$

Naïve Bayes Assumption

- Assume each piece of evidence (symptom) is independent give the diagnosis
- then

$$P(e_1,\ldots,e_m \mid d_i) = \prod_{k=1}^{m} P(e_k \mid d_i)$$

what is the structure of the corresponding BN?

Naïve Bayes Example

- possible diagnosis: Allergy, Cold and OK
- possible symptoms: Sneeze, Cough and Fever

	Well	Cold	Allergy
P(d)	0.9	0.05	0.05
P(sneeze|d)	0.1	0.9	0.9
P(cough|d)	0.1	0.8	0.7
P(fever|d)	0.01	0.7	0.4

my symptoms are: sneeze & cough, what is the diagnosis?

Learning the Probabilities

- aka parameter estimation
- we need
  - $P(d_i)$ – prior
  - $P(e_k | d_i)$ – conditional probability
- use training data to estimate

Maximum Likelihood Estimate (MLE)

- use frequencies in training set to estimate:

$$p(d_i) = \frac{n_i}{N}$$

$$p(e_k \mid d_i) = \frac{n_{ik}}{n_i}$$

where n_x is shorthand for the counts of events in training set

Example:

what is: P(Allergy)?

P(Sneeze| Allergy)?

P(Cough| Allergy)?

D	Sneeze	Cough	Fever
Allergy	yes	no	no
Well	yes	no	no
Allergy	yes	no	yes
Allergy	yes	no	no
Cold	yes	yes	yes
Allergy	yes	no	no
Well	no	no	no
Well	no	no	no
Allergy	no	no	no
Allergy	yes	no	no

Laplace Estimate (smoothing)

- use smoothing to eliminate zeros:

$$p(d_i) = \frac{n_i + 1}{N + n}$$

many other smoothing schemes...

$$p(e_k \mid d_i) = \frac{n_{ik} + 1}{n_i + 2}$$

where n is number of possible values for d
and e is assumed to have 2 possible values

Comments

- Generally works well despite blanket assumption of independence
- Experiments show competitive with decision trees on some well known test sets (UCI)
- handles noisy data

Learning more complex Bayesian networks

- Two subproblems:
- learning structure: combinatorial search over space of networks
- learning parameters values: easy if all of the variables are observed in the training set; harder if there are ‘hidden variables’

THANK YOU