Bayesian Network

Bor-shen Lin

bslin@cs.ntust.edu.tw

http:www.cs.ntust.edu.tw/~bslin

Outlines

- Random Variables and Distribution
- Expectation
- Joint distribution and conditional probability
- Bayesian Theory
- Bayesian Classifier
- Bayesian Network

Type of Random Variables

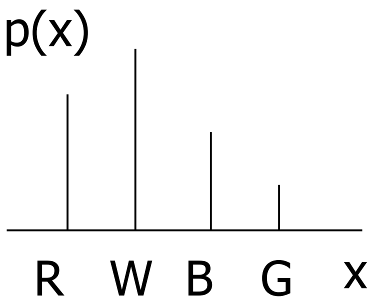
- Discrete variables
  - Coin/dice tossing
  - Lottery
  - Blood type
- Continuous variable
  - Temperature
  - Time of leaving home
  - Sound
  - Image

Discrete Random Variables

- A basket: 3 red(R) balls, 4 white(W) balls, 2 blue(B) balls, 1 green(G) ball.
- Random variable X: the color of a ball drawn randomly from the bag
- *Probability Weight Function* (p.w.f.)

  $p(X=x) = p(x)$

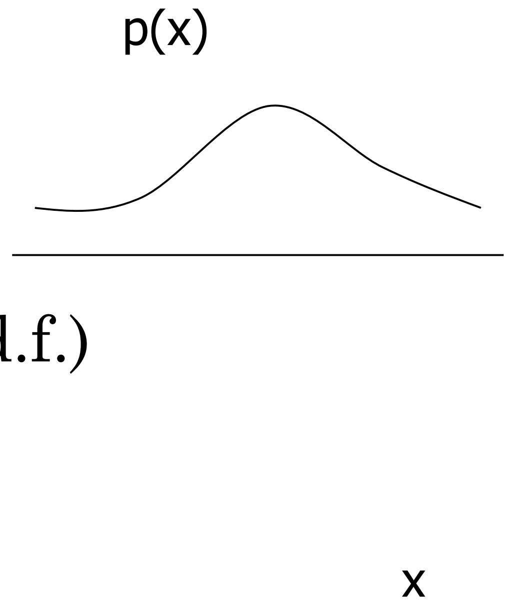
  - $p(X = R) = 3/10$, $p(X = W) = 4/10$, $p(X = B) = 2/10$, $p(X = G) = 1/10$
  - $\Sigma_x\, p(x) = 1.0$



Continuous Random Variables

- Example: temperature X
- *Probability Density Function* (p.d.f.)
  - p(X = x) = p(x)
  - $\int$p(x)dx = 1
- *Cumulated distribution function* (c.d.f.)

  $$\mathrm{P(x)} = \int_{-\infty}^{x} p(x)dx$$
- Example: height、weight、score



EXPECTATION

- Continuous Random Variable
  - $\mu = E(X) \equiv \int_{-\infty}^{\infty} xp(x)dx$
  - $var(X) = E\left((X-\mu)^2\right) \equiv \int_{-\infty}^{\infty}((x-\mu)^2 p(x)dx$
  - $\sigma \equiv \sqrt{var(X)}$
- Discrete Random Variable
  - $\mu = E(X) \equiv \sum_{-\infty}^{\infty} xp(x)$
  - $var(X) = E\left((X-\mu)^2\right) \equiv \sum_{-\infty}^{\infty}(x-\mu)^2 p(x)$
  - $\sigma \equiv \sqrt{var(X)}$
- Numerical only
  - might not be able to compute the expectations
  - Examples: blood type、job

Expectation

- For any $f(x)$, we may take the expectation
  - $E(f(X)) \equiv \int_{-\infty}^{\infty} f(x)p(x)dx$
- Entropy
  - Average information
  - $I(X) \equiv E\left(\log\left(\frac{1}{p(X)}\right)\right) = \int_{-\infty}^{\infty} \log(\frac{1}{p(x)})p(x)dx$

    $= -\int_{-\infty}^{\infty} \log(p(x))p(x)dx.$

Joint Distribution

- Joint distribution : $p(x,y)$
  - $\iint p(x, y)dxdy = 1.0$
  - $p(x) \equiv \int p(x, y)dy \rightarrow \int p(x)dx = 1$
  - $p(y) \equiv \int p(x, y)dx \rightarrow \int p(y)dy = 1$
- Conditional probability.
  - $p(x \mid y) \equiv p(x, y)/p(y)$
    - $p.\ d.\ f.$ *of* X *when* y *is given*
    - $\int p(x \mid y)dx = \int [p(x,y)/p(y)]dx = 1$
  - $p(x) \equiv \int p(x, y)dy = \int p(x \mid y)p(y)dy$
  - $p(y) \equiv \int p(x, y)dx$

Conditional Mean

- $E(X \mid y) = \int x \cdot p(x \mid y) dx \neq E(X)$
  - Expectation of X given Y = y
  - $E(X \mid y)$ is a function of y (varies w.r.t. y)
- $E(X) = \int xp(x)dx = \iint xp(x,y)dxdy$

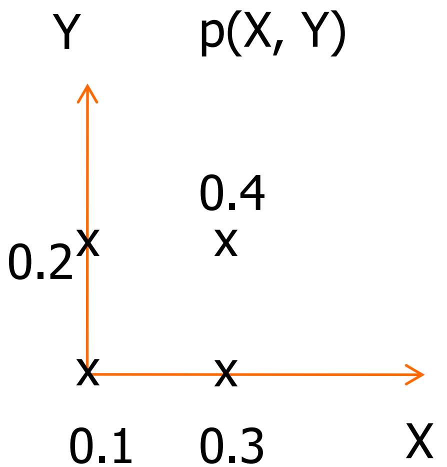
  $$= \iint [xp(x \mid y)dx]\, p(y)dy$$

  $$= \int E(X \mid y)\, p(y)dy$$
- $E(X)$ is not the function of y (the weighting average of $E(X \mid y)$ on all y's)

Statistical Independence

- $p(x, y) = p(x)\, p(y)$
  → X, Y are statistically independent
- $p(x|y) = \frac{p(x,y)}{p(y)} = p(x)$
  the probability weight of x is NOT influenced by y
- Example
  - Coin tossing result X : Head/Tail
  - Dice tossing result Y: point 1,2,…, 6
  - When X and Y are *statistically independent*
    $p(X = Head,\ Y = 4) = p(X = Head)p(Y = 4)$
    $= (1/2) * (1/6)$

Example of Statistically dependent

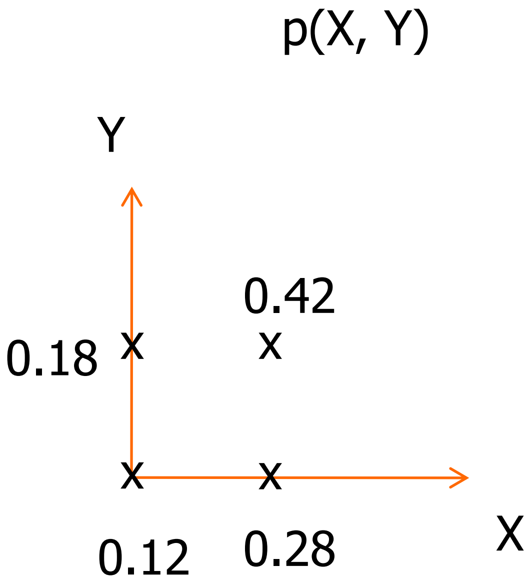


X, Y are statiscally dependent or independent?

$p(X = 0, Y=0) = 0.1$
$p(X = 0, Y = 1) = 0.2$
$p(X = 1, Y=0) = 0.3$
$p(X = 1, Y = 1) = 0.4$

1. $p(X=0)=0.1+0.2 = 0.3$
   $p(X=1) = 0.3+0.4=0.7$
2. $p(X=0 \mid Y = 0)=0.1/0.4=0.25$
   $p(X=1 \mid Y = 0)=0.3/0.4=0.75$
3. $p(X=0 \mid Y = 1)=0.2/0.6=0.333$

Example of Statistic Independence



X,Y are statistically independent

$p(X = 1) = 0.42 + 0.28 = 0.7$
$p(X = 0) = 0.18 + 0.12 = 0.3$
$p(X = 1|Y = 1) = \frac{p(X=1|,Y=1)}{p(Y=1)} =$
$\frac{0.42}{0.18+0.42} = 0.7$
$p(X = 0|Y = 1) = \frac{0.18}{0.18 + 0.42} = 0.3$

Bayesian Theory

- $p(y|x) = \frac{p(x,y)}{p(x)} = \frac{p(y)p(x|y)}{\sum_y p(y)p(x|y)}$
- Could be used for inference between observable phenomena(x) and unseen cause (y)
  - X: phenomena/events (e.g.: having fever, cough,…)
    - Could have multiple dimensions!
    - X could be continuous/discrete/composite!
  - Y: cause(e.g. having cold, inflammation, plague, …)
    - Classification/detection: Y is discrete variable
    - Estimation: Y is continuous variable

Correlation Coefficient

$$\rho_{xy} \equiv \frac{E[(X - E(X))(Y - E(Y))]}{[Var(X) \cdot Var(Y)]^{\frac{1}{2}}}$$

- If X and Y are statistically independent

  $\rightarrow \rho_{xy} = 0$

- $\rho_{xy} = 0$ does not guarantee statistical independence

Conditional Probability for Medical Diagnosis

- F: fever (symptom), C: having a cold (cause)
  - F → C or C → F does not hold in general
- $p(F | C) = 0.8$ having fever in case of having cold

  $p(F) = 0.001$ having fever

  $p(C) = 0.0001$ having cold

  $p(C | F) = p(C)p(F | C)/p(F) = 0.008$
- Though it is highly probable that one who has cold also have fever (0.8), it can be concluded that one must have fever because he/she caught cold (only 0.008)
- $p(C) = 0.0001 \rightarrow p(C | F) = 0.008$

  The event F increases the probability of having cold (from 0.0001 to 0.008), but the probability is not high.

Inference with Conditional Probability

- $p(F) = 0.001$ (F: one has fever)
  $p(C) = 0.0001$ (C: have a cold)
  $p(P) = 0.000000001$ (P: plague)
  $p(F|C) = 0.8$ (having fever when having a cold)
  $p(F|P) = 0.99$ (have fever in case getting plague)
  $p(F|P) > p(F|C)$? (ML detection)
- With F (one has fever) → guess C or P? (disease)
  $p(C|F) / p(P|F) = p(C)(F|C) / (p(P)p(F|P))$
  $= (0.0001 * 0.8) / (0.000000001*0.99) = 84210$
  → $p(C|F) >> p(P|F)$ (C is more likely)
  → it is more reasonable to guess C!
- Without F: $p(C)/p(P) = 100000$

Simple Bayesian Learning (1)

- Suppose we have a set of hypotheses $H_1 \ldots H_n$.
- For each H_i, $p(H_i|E) = \frac{p(H_i)p(E|H_i)}{p(E)}$
  - $p(H_i|E)$ is used to represent the probability that some hypothesis, H, is true, given evidence E.
  - Hence, given a piece of evidence, a learner can determine which is the most likely explanation by finding the hypothesis that has the highest posterior probability.
- Maximum a Posterior (MAP) detection

  $i^* = argmax_i\{p(H_i|E)\}$

Simple Bayesian Concept Learning (2)

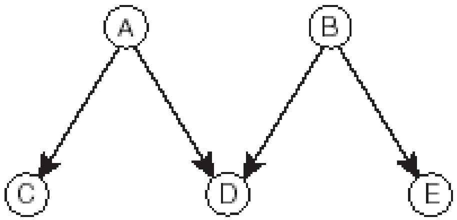
- Since P(E) is independent of H_i it will have the same value for each hypothesis.
- Hence, it can be ignored, and we can find the hypothesis with the highest value of:

$$P(E|H_i) \cdot P(H_i)$$

- We can simplify this further if all the hypotheses are equally likely, in which case we simply seek the hypothesis with the highest value of $P(E|H_i)$.
- This is the likelihood of E given H_i.
- Maximum Likelihood (ML) detection

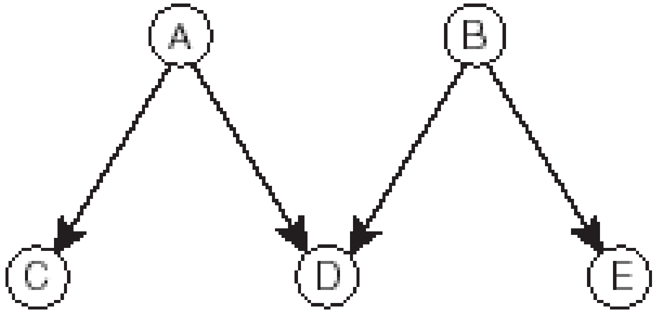
Bayesian Belief Networks (1)

- A belief network assumes some statistical dependencies between a set of variables.
  - Such assumptions can simplify the computations of complicated conditional probabilities
- Two variables A and B are statistically independent if the likelihood that A will occur has nothing to do with whether B occurs.
- Example : C and D are dependent on A; D and E are dependent on B.
- The Bayesian belief network has the conditional probabilities associated with each link.



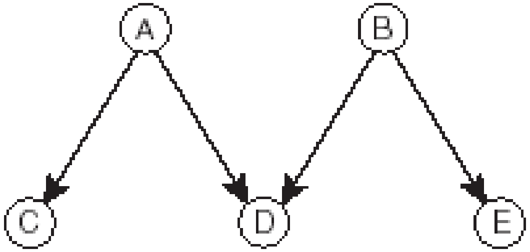
Bayesian Belief Networks (2)

- A complete set of probabilities for this BBN
  - $P(A) = 0.1$
  - $P(B) = 0.7$
  - $P(C|A) = 0.2$
  - $P(C|\neg A) = 0.4$
  - $P(D|A \land B) = 0.5$
  - $P(D|A \land \neg B) = 0.4$
  - $P(D|\neg A \land B) = 0.2$
  - $P(D|\neg A \land \neg B) = 0.0001$
  - $P(E|B) = 0.2$
  - $P(E|\neg B) = 0.1$



Bayesian Belief Networks (3)

- The joint probability
  - $P(A,B,C,D,E) = P(A) \cdot P(B|A) \cdot P(C|A,B) \cdot P(D|A,B,C) \cdot P(E|A,B,C,D)$
- With the assumption of independences, the computation of $P(A,B,C,D,E)$ can be simplified.
  - $P(C|A,B) = P(C|A)$,
  - $P(D|A,B,C) = P(D|A,B)$
  - $P(E|A,B,C,D) = P(E|B)$
  - $P(A,B,C,D,E) = P(A) \cdot P(B) \cdot P(C|A) \cdot P(D|A,B) \cdot P(E|B)$



Example of BN

- C: you go to college
- S: you will study
- P: you will party
- E: you will be successful in exams
- F: you will have fun
- P(E | F, ¬P, S, C) ?

P(C)
0.2

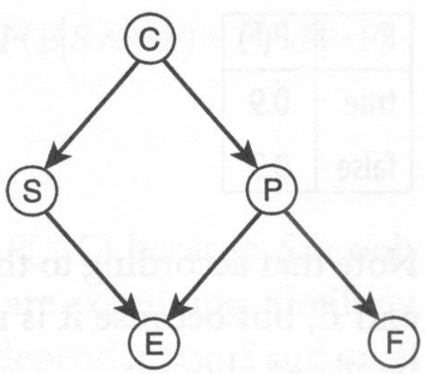
p(C)

C	P(S)
true	0.8
false	0.2

p(S|C)

C	P(P)
true	0.6
false	0.5

p(P|C)



P	P(F)
true	0.9
false	0.7

p(F|P)

S	P	P(E)
true	true	0.6
true	false	0.9
false	true	0.1
false	false	0.2

p(E|S,P)

Naïve Bayes Classifier (1)

- A vector of data is classified.
  - $P(c_i|x_1, x_2, \cdots, x_n)$
  - The classification with the highest posterior probability is chosen.
  - The hypothesis which has the highest posterior probability is the maximum a posteriori, or MAP hypothesis.
  - In this case, we are looking for the MAP classification.
- Bayes' theorem is used to find the posterior probability
  - $P(c_i|x_1, x_2, \cdots, x_n) = \frac{P(c_i)P(x_1, x_2, \cdots, x_n|c_i)}{P(x_1, x_2, \cdots, x_n)}.$

The Naïve Bayes Classifier (2)

- Since $P(x_1, \ldots, x_n)$ is independent of c_i, we can eliminate it, and simply aim to find the classification c_i, for which the following is maximized:

  $P(c_i)P(x_1, x_2, \cdots, x_n|c_i)$.
- We now assume that all the attributes $x_1, \ldots, x_n$ are independent , so $P(c_i, x_1, x_2, \cdots, x_n)$ can be rewritten as:

  $P(c_i)\prod_{j=1}^{n} P(x_j|c_i)$.
- The classification for which this is highest is chosen to classify the data.

The Naïve Bayes Classifier (3)

- $i^* = \text{argmax } P(C_i \mid x, y, z)$

  $= \text{argmax } P(C_i, x, y, z)$
- $C_1 = A$, $C_2 = B$, $C_3 = C$
- For $x=2$, $y=3$, $z = 4$, compare
  - $P(A, x=2, y=3, z=4)$
  - $P(B, x=2, y=3, z=4)$
  - $P(C, x=2, y=3, z=4)$

x	y	z	Classification
2	3	2	A
4	1	4	B
1	3	2	A
2	4	3	A
4	2	4	B
2	1	3	C
1	2	4	A
2	3	3	B
2	2	4	A
3	3	3	C
3	2	1	A
1	2	1	B
2	1	4	A
4	3	4	C
2	2	4	A

Example

Training data

	Buy A	Buy B	Buy C	Buy D
Customer 1	Yes	Yes	No	Yes
Customer 2	Yes	No	Yes	No
…	…	…	…	…

compare

$$P\left(D \middle| A, B, \bar{C}\right)$$

$$P\left(\bar{D} \middle| A, B, \bar{C}\right)$$

Recommend those products with high ratios

Applications of Bayesian Network

- Bayesian Network
  - A set of random variables with an assumption of statistical dependency
  - Detection and Estimation (ML, MAP)
- Bayesian Classifier and Probabilistic Reasoning
- Markov Process
- N-Gram Language Model (e.g. trigram $p(W_i|W_{i-2}, W_{i-2})$)
- Probabilistic Latent Semantic Indexing (PLSI)
- Gaussian Mixture Model
  - Continuous variables dependent of a state variable
- Hidden Markov Model
  - Hidden state sequence as Markov variables
  - Observation variables dependent of hidden state

References

- Artificial Intelligence Illuminated, Ben Coppin