Fundamentals of Probability for AI

INF8225 - Lesson 1b

See the following selection of slides from:

Data Mining

Practical Machine Learning Tools and Techniques

Slides from Chapter 9 of *Data Mining*
by I. H. Witten, E. Frank, M. A. Hall and C.J. Pal

Random variables

- In probabilistic approaches to machine learning it is common to think of data as observations arising from an underlying probability model for *random variables*
- Given a discrete random variable A, $P(A)$ is a function that encodes the probabilities for each of the categories, classes or states that A may be in
- For a continuous random variable x, $p(x)$ is a function that assigns a probability density to all possible values of x
- In contrast, $P(A=a)$ is the single probability of observing the specific event $A=a$

Notation

- The $P(A=a)$ notation is often simplified to simply $P(a)$, but one must remember if a was defined as a random variable or as an observation
- Similarly for the observation that continuous random variable x has the value x_1 it is common to write this as $p(x_1)=\mathrm{p}(x=x_1)$, a simplification of the longer but clearer notation

The product rule

- The *product rule*, sometimes referred to as the "fundamental rule of probability," states that the joint probability of random variables *A* and *B* can be written

$$P(A,B) = P(A \mid B)P(B)$$

- The product rule also applies when *A* and *B* are groups or subsets of events or random variables.

The sum rule

- The *sum rule* states that given the joint probability of variables $X_1, X_2, \ldots, X_N$, the *marginal probability* for a given variable can be obtained by summing (or integrating) over all the other variables.
- For example, to obtain the marginal probability of X_1, sum over all the states of all the other variables:

$$P(X_1) = \sum_{x_2} \ldots \sum_{x_N} P(X_1, X_2 = x_2, \ldots, X_N = x_N)$$

Marginalization

- The previous notation can be simplified to

$$p(x_1) = \sum_{x_2} \ldots \sum_{x_N} P(x_1, x_2, \ldots, x_N)$$

- The sum rule generalizes to continuous random variables, ex. for $x_1, x_2, \ldots, x_N$ we have

$$p(x_1) = \int_{x_2} \ldots \int_{x_N} p(x_1, x_2, \ldots, x_N) dx_2 \ldots dx_N$$

- These procedures are known as *marginalization*
- They give us *marginal distributions* of the variables not included in the sums or integrals

Bayes' Rule

- Can be obtained by swapping A and B in the product rule and observing $P(B|A)P(A)=P(A|B)P(B)$ and therefore

$$P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A)}$$

- Suppose we have models for $P(A|B)$ and $P(B)$
  – We observe that $A=a$, and
  – we want to compute $P(B|A=a)$
  – $P(A=a|B)$ is referred to as the *likelihood*
  – $P(B)$ is the *prior* distribution of B
  – $P(B|A=a)$ is posterior distribution, obtained from:

$$P(A=a) = \sum_b P(A=a, B=b) = \sum_b P(A=a \mid B=b)P(B=b)$$

Maximum Likelihood

- Our goal is to estimate a set of parameters θ of a probabilistic model, given a set of *observations* $x_1, x_2, \ldots, x_n$.
- Maximum likelihood techniques assume that:
  1) the examples have no dependence on one another, the occurrence of one has no effect on the others, and
  2) each can be modeled in exactly the same way.
- These assumptions are often summarized by saying that events are *independent and identically distributed* (i.i.d.).

Maximum Likelihood

- The i.i.d. assumption corresponds to the use of a joint probability density function for all observations consisting of the product of the same probability model $p(x_i;\theta)$ applied to each observation independently.
- For n observations, this could be written as

$$p(x_1,x_2,\ldots,x_n;\theta) = p(x_1;\theta)p(x_2;\theta)\ldots p(x_n;\theta)$$

where each function $p(x_i;\theta)$ has the same θ

Maximum Likelihood

- The likelihood of our data can be written

$$\mathrm{L}(\theta; x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} p(x_i; \theta)$$

- The data is fixed, but we can adjust θ so as to *maximize the likelihood* or *log-likelihood*

$$\theta_{ML} = \arg\max_{\theta} \sum_{i=1}^{n} \log p(x_i; \theta)$$

- We use the the log-likelihood as it is more numerically stable

Maximum a posteriori (MAP) parameter estimation

- If we treat our parameters as random variables we can compute the posterior

$$p(\theta \mid x_1, x_2, \ldots, x_n) = \frac{p(x_1, x_2, \ldots, x_n \mid \theta) p(\theta)}{p(x_1, x_2, \ldots, x_n)}$$

- We have used | or the "given" notation in place of ; to emphasize that θ is random, but
- Conditioned on a point estimate for the posterior we have a conditionally i.i.d. model
- MAP parameter estimation seeks

$$\theta_{MAP} = \arg\max_{\theta} \left[\sum_{i=1}^{n} \log p(x_i; \theta) + p(\theta) \right]$$

The chain rule of probability

- Results from applying the product rule recursively between a single variable and the rest of the variables
- The *chain rule* states that the joint probability of *n* attributes $A_{i=1...m}$ can be decomposed into the following product:

$$P(A_1, A_2, \ldots, A_n) = P(A_1) \prod_{i=1}^{n-1} P(A_{i+1} | A_i, A_{i-1}, \ldots, A_1)$$

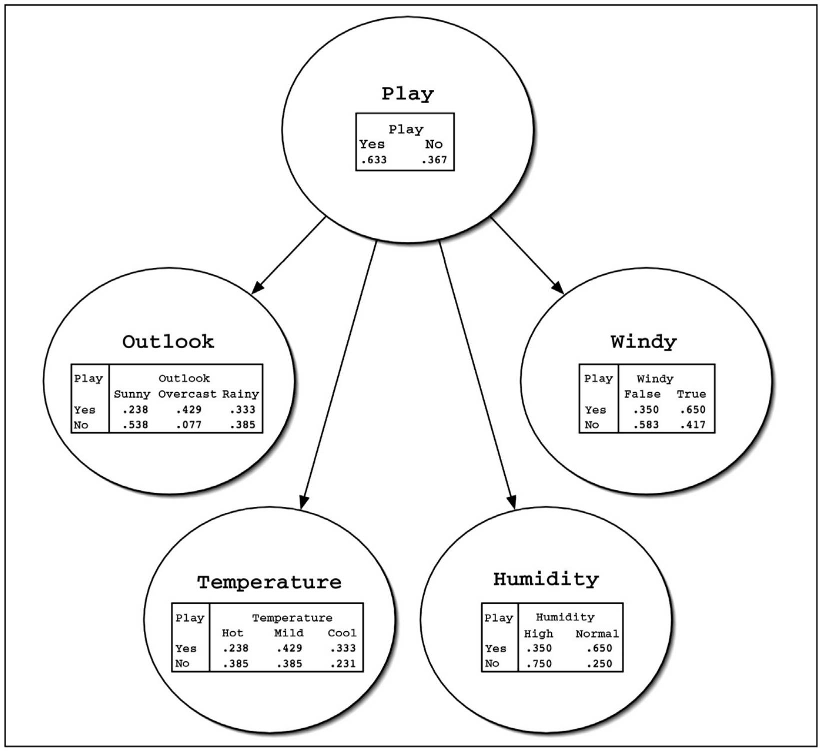
Bayesian networks

- The chain rule holds for any order for the A_is
- A Bayesian network is an acyclic graph,
- Therefore its nodes can be given an ordering where ancestors of node A_i have indices < i
- Thus a Bayesian network can be written

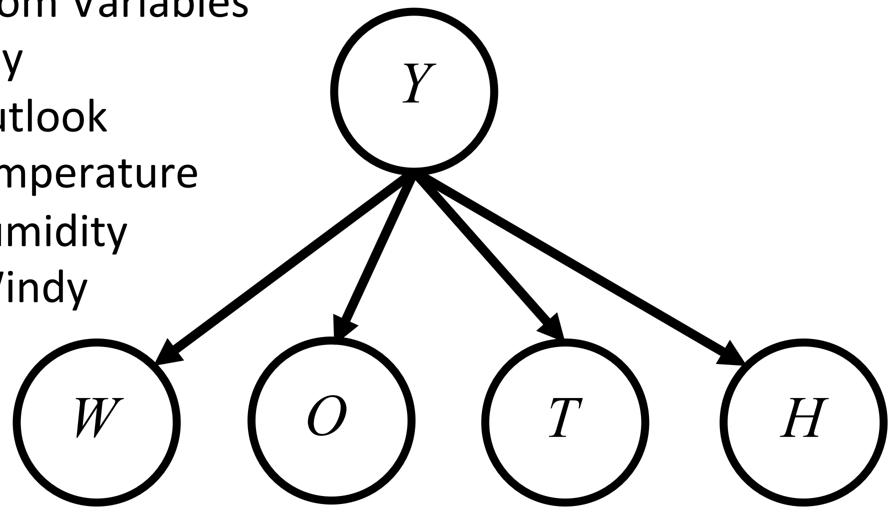
$$P(A_1, A_2, \ldots, A_n) = \prod_{i=1}^{n} P(A_i \mid \text{Parents}(A_i))$$

- When a variable has no parents, we use the unconditional probability of that variable

Bayesian network #1 for the weather data



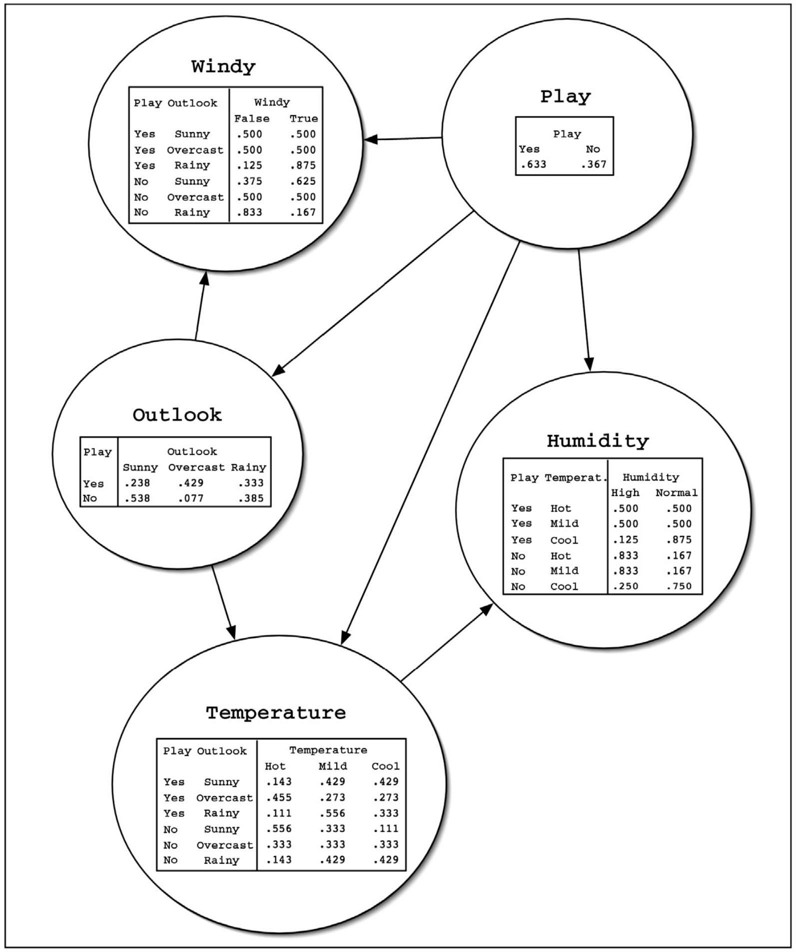
Random Variables
Y: Play
O: Outlook
T: Temperature
H: Humidity
W: Windy



The graphs express the factorization below:

$$P(Y,O,.T,H,W) = P(W \mid Y)P(O \mid Y)P(T \mid Y)P(H \mid Y)P(Y)$$

Bayesian network #2 for the weather data



Random Variables

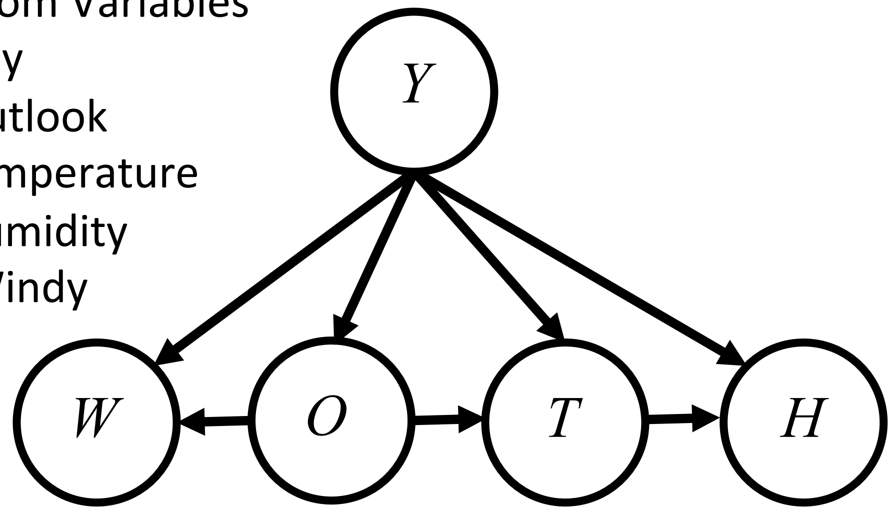
Y: Play

O: Outlook

T: Temperature

H: Humidity

W: Windy



The graphs express the factorization below:

$$P(Y,O,.T,H,W) = P(W \mid O,Y)P(O \mid Y)P(T \mid O,Y)P(H \mid T,Y)P(Y)$$

Estimating Bayesian network parameters

- The log-likelihood of a Bayesian network with *V* variables and *N* examples of complete variable assignments to the network is

$$\sum_{i=1}^{N} \log P(\{\tilde{A}_1, \tilde{A}_2, \ldots, \tilde{A}_V\}_i) = \sum_{i-1}^{N} \sum_{v=1}^{V} \log P(\tilde{A}_{v,i} \,\middle|\, \text{Parents}(\tilde{A}_{v,i}); \Theta_v)$$

  where the parameters of each conditional or unconditional distribution are given by Θ_v

- We use the $\tilde{A}_{v,i}$ notation to indicate the *i*th observation of variable *v*

Estimating probabilities in Bayesian networks

- The estimation problem *decouples* into separate estimation problems for each conditional or unconditional probability
- Unconditional probabilities can be written as

$$P(A = a) = \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}(\tilde{A}_i = a)$$

where $\mathbf{1}(\tilde{A}_i = a)$ is an indicator function returning 1 when the i^{th} observed value for $A_i = a$ and 0 otherwise

Estimating conditional distributions

- Estimating conditional distributions in Bayesian networks is equally easy and amounts to simply counting configurations and dividing, ex.

$$P(B=b \mid A=a) = \frac{P(B=b, A=a)}{P(A=a)} = \frac{\sum_{i=1}^{N} \mathbf{1}(\tilde{A}_i = a, \tilde{B}_i = b)}{\sum_{i=1}^{N} \mathbf{1}(\tilde{A}_i = a)}.$$

- Zero counts cause problems and this motivates the use of Bayesian priors

Fundamentals Elements of Probability and Statistics for AI

See A.2 of:

Appendix A : Theoretical Foundations

of

Data Mining

Practical Machine Learning Tools and Techniques

Slides for Chapter 9 of *Data Mining*

by I. H. Witten, E. Frank, M. A. Hall and C.J. Pal

Fundamentals of Linear Algebra for AI

See A.1 of :

Appendix A : Theoretical Foundations

of

Data Mining

Practical Machine Learning Tools and Techniques

Slides for Chapter 9 of *Data Mining*

by I. H. Witten, E. Frank, M. A. Hall and C.J. Pal