T-61.3050 Machine Learning: Basic Principles

Bayesian Decision Theory

Kai Puolamäki

Laboratory of Computer and Information Science (CIS)
Department of Computer Science and Engineering
Helsinki University of Technology (TKK)

Autumn 2007

Outline

Dimensions of a Supervised Learner

Model

$$g(\mathbf{x} \mid \theta)$$

Loss Function

$$E(\theta \mid \mathcal{X}) = \frac{1}{N} \sum_{t=1}^{N} L\left(r^t, g(\mathbf{x}^t \mid \theta)\right).$$

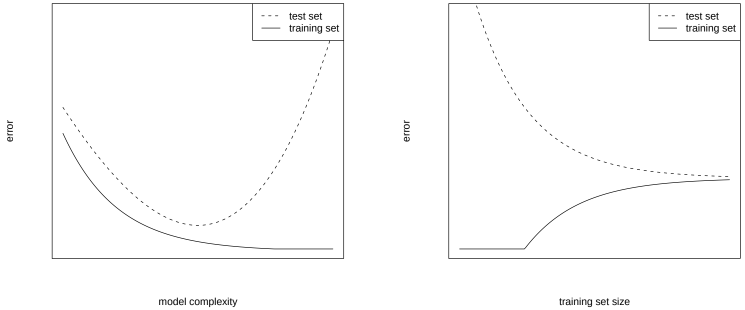
Optimization Procedure

$$\theta \leftarrow \arg\min_{\theta} E(\theta \mid \mathcal{X}).$$

Outline

Model Selection and Generalization

Schematic illustration of the empirical vs. generalization error



- empirical error = error on training set
- generalization error = error on test set
- We see empirical error, but want to minimize the error on new data

Validation

Question 1

What is the correct model complexity?

Question 2

What is the generalization error?

- To answer the Question 1 divide the data into training and validation sets. Choose model complexity that has the smallest error on the validation set.
- To answer the Question 2 divide the data into training and test sets. The generalization error is approximately the error on the test set.
- To answer both questions the data should be divided into training, validation and test sets.
- There are more efficient methods, such as cross-validation.

Model Selection and Generalization

- Learning is ill-posed problem: data is not sufficient to find unique/correct solution.
- Inductive bias is needed; we need assumptions about the hypothesis class (model family) $\mathcal{H}$.
- Generalization: how well model performs on new data.
- Overfitting: $\mathcal{H}$ more complex than C or f.
- Underfitting: $\mathcal{H}$ less complex than C or f.
- Triple trade-off (Diettrich 2003):
  - complexity of $\mathcal{H}$;
  - amount of training data; and
  - generalization error on new data.

Outline

Basic of Probability

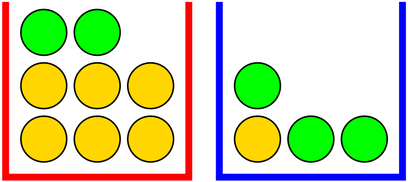
- You should know basics of probability (Mat-1.2600/2620 or Appendix A of Alpaydin (2004)).
- Probability can be interpreted as a frequency or degree of belief.
- Sample space S: the set of all possible outcomes.
- Event $E \subseteq S$: one possible set of outcomes.
- Probability measure P satisfies:
  - $P(S) = 1$.
  - $0 \leq P(E) \leq 1$ for all $E \subseteq S$.
  - $E \subseteq S \wedge F \subseteq S \wedge E \cap F = \emptyset \Rightarrow P(E \cup F) = P(E) + P(F)$.

Rules of Probability

- Interpret E, F as random variables getting values of e, f (coin tossing example: E can get a value of $e \in \{\text{heads}, \text{tails}\}$, F can get a value of coin landing in $f \in \{\text{table}, \text{floor}\}$).
- $P(E, F) = P(F, E)$: probability of both E and F happening.
- $P(E) = \sum_F P(E, F)$ (sum rule, marginalization)
- $P(E, F) = P(F \mid E)P(E)$ (product rule, conditional probability)
- Consequence: $P(F \mid E) = P(E \mid F)P(F)/P(E)$ (Bayes' formula)
- We say E and F are independent if $P(E, F) = P(E)P(F)$ (for all e and f).
- We say E and F are conditionally independent given G if $P(E, F \mid G) = P(E \mid G)P(F \mid G)$, or equivalently $P(E \mid F, G) = P(E \mid G)$.

Fruits in Boxes

- $P(B = r, F = a) = n_{RA}/n = 1/6$.
- $P(B = r) = \sum_{x \in \{a,o\}} P(B = r, F = x) = n_{RA}/n + n_{RO}/n = n_R/n = 2/3$.
- $P(F = o \mid B = r) = n_{RO}/n_R = 3/4$.
- $P(B = r \mid F = o) = P(F = o \mid B = r)P(B = r)/P(F = o) = \frac{3}{4} \times \frac{8}{12} \times \frac{12}{7} = \frac{6}{7}$.



	apples	oranges	Σ
red box	$n_{RA} = 2$	$n_{RO} = 6$	$n_R = 8$
blue box	$n_{BA} = 3$	$n_{BO} = 1$	$n_B = 4$
Σ	$n_A = 5$	$n_O = 7$	$n = 12$

Table: Count of fruits in two boxes.

Fruits in Boxes

- B and F are random variables which can take two values (r or b; a or o, respectively).
- We computed probabilities of events of drawing one fruit in random such that the probability of drawing each fruit is $1/12$, independent of the box or type.
- We viewed the probabilities as frequencies.
- When all prior information (e.g., counts of the fruits in the boxes) is not known the probabilities turn into degrees of belief (it may be still easier to think them as frequencies, though).

Estimating Probability

- In real life, estimating the probabilities of various events from a sample is difficult.
- For the purposes of today, we mostly assume that someone gives us the probabilities.
- Today we can estimate the probabilities with sample frequencies.
  - Example: Someone is tossing a 0–1 coin that gives $X = 1$ with probability $P(X = 1) = p$ and $X = 0$ with probability $P(X = 0) = 1 - p$ (Bernoulli distribution). We notice he got n_1 ones and n_0 zeroes in a sample of $N = n_1 + n_0$ tosses. Based on this sample, we can estimate p with $\hat{p} = n_1 / N$.

Outline

Using Probabilities Classification

Coin Tossing

- Someone is tossing a 0–1 coin that gives $X = 1$ (HEADS) with probability $P(X = 1) = p$ and $X = 0$ (TAILS) with probability $P(X = 0) = 1 - p$ (Bernoulli distribution).
- Task: make a classifier for the next toss.
- Prediction: Choose $X = 1$ (HEADS) if $p \geq 1/2$, $X = 0$ (TAILS) otherwise.

Using Probabilities in Classification

Credit Scoring

- Task: classify a customer HIGH RISK ($C = 1$) or LOW RISK ($C = 0$) based on her income (x_1) and savings (x_2).
- Assume $P(C \mid x_1, x_2)$ is known.

Prediction:

$$\text{choose} \begin{cases} C = 1 & \text{if} \quad P(C = 1 \mid x_1, x_2) \geq \frac{1}{2}, \\ C = 0 & \text{otherwise.} \end{cases}$$

or equivalently

$$\text{choose} \begin{cases} C = 1 & \text{if} \quad P(C = 1 \mid x_1, x_2) \geq P(C = 0 \mid x_1, x_2), \\ C = 0 & \text{otherwise.} \end{cases}$$

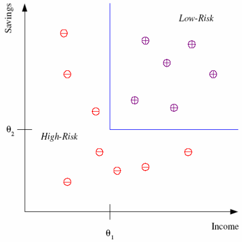


Figure 1.1 of Alpaydin (2004).

Bayes' Rule

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}},$$

or

$$P(C \mid \mathbf{x}) = \frac{P(\mathbf{x} \mid C) \times P(C)}{P(\mathbf{x})}.$$

- The likelihood $P(\mathbf{x} \mid C = 1)$ is the probability that a HIGH RISK customer ($C = 1$) has the associated observed value $\mathbf{x}$. (This is usually easy to compute.)
- The prior probability $P(C = 1)$ is the probability of observing $C = 1$ (before $\mathbf{x}$ is known).
- The evidence $P(\mathbf{x})$ is the marginal probability that an observation $\mathbf{x}$ is seen, regardless of the value of C. (This is usually difficult to compute directly.)

Bayes' Rule

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}},$$

or

$$P(C \mid \mathbf{x}) = \frac{P(\mathbf{x} \mid C) \times P(C)}{P(\mathbf{x})}.$$

Using the sum and product rules we obtain:

- $P(C = 0) + P(C = 1) = 1$.
- $P(C = 0 \mid \mathbf{x}) + P(C = 1 \mid \mathbf{x}) = 1$.
- $P(\mathbf{x}) = P(\mathbf{x} \mid C = 1)P(C = 1) + P(\mathbf{x} \mid C = 0)P(C = 0)$.

Bayes' Rule

Classification to K classes

$$P(C_i \mid \mathbf{x}) = \frac{P(\mathbf{x} \mid C_i)P(C_i)}{P(\mathbf{x})} = \frac{P(\mathbf{x} \mid C_i)P(C_i)}{\sum_{k=1}^{K} P(\mathbf{x} \mid C_k)P(C_k)}$$

- $P(C_k) \geq 0$ and $\sum_{k=1}^{K} P(C_k) = 1$.
- Naive Bayes Classifier: choose C_k where $k = \arg\max_k P(C_k \mid \mathbf{x})$.
- A customer is associated with vector $\mathbf{x}$ such that $P(\mathbf{x} \mid C = 1) = 0.002$ and $P(\mathbf{x} \mid C = 0) = 0.001$.
- 20% of the customers are HIGH RISK $(C = 1)$, we therefore set the prior probabilities to $P(C = 1) = 0.2$ and $P(C = 0) = 0.8$.
- Inserting in equation we obtain $P(C = 1 \mid \mathbf{x}) = 0.33$ and $P(C = 0 \mid \mathbf{x}) = 0.67$, we therefore classify the customer as LOW RISK $(C = 0)$.

Bayes' Rule

Classification to K classes

$$P(C_i \mid \mathbf{x}) = \frac{P(\mathbf{x} \mid C_i)P(C_i)}{P(\mathbf{x})} = \frac{P(\mathbf{x} \mid C_i)P(C_i)}{\sum_{k=1}^{K} P(\mathbf{x} \mid C_k)P(C_k)}$$

- $P(C_k) \geq 0$ and $\sum_{k=1}^{K} P(C_k) = 1$.
- Naive Bayes Classifier: choose C_k where $k = \arg\max_k P(C_k \mid \mathbf{x})$.
- A customer is associated with vector $\mathbf{x}$ such that $P(\mathbf{x} \mid C = 1) = 0.002$ and $P(\mathbf{x} \mid C = 0) = 0.001$.
- 20% of the customers are HIGH RISK ($C = 1$), we therefore set the prior probabilities to $P(C = 1) = 0.2$ and $P(C = 0) = 0.8$.
- Inserting in equation we obtain $P(C = 1 \mid \mathbf{x}) = 0.33$ and $P(C = 0 \mid \mathbf{x}) = 0.67$, we therefore classify the customer as LOW RISK ($C = 0$).

Outline

Risks and Losses

- Often, the cost of errors differs. For example, a wrong decision to grant credit may be much more costly than a wrong decision not to grant credit.
- Decision theory: how to make optimal decisions, given all available information.
- At each time, you can choose one action α_i.
- Action α_i causes loss λ_{ik} when the state is C_k.

λ	$C=0$	$C=1$
$\alpha_0 = \text{grant credit}$	EUR 0	EUR 1000
$\alpha_1 = \text{don't grant credit}$	EUR 100	EUR 0

- Expected risk: $R(\alpha_i \mid \mathbf{x}) = E[\lambda_{ik}] = \sum_{k=1}^{K} \lambda_{ik} P(C_k \mid \mathbf{x})$.
- Choose α_i where $i = \arg\min_i R(\alpha_i \mid \mathbf{x})$.

Risks and Losses

0/1 loss

- 0/1 loss:

$$\lambda_{ik} = \begin{cases} 0 & i = k \\ 1 & i \neq k \end{cases}$$

$$\begin{aligned} R(\alpha_i \mid \mathbf{x}) &= \sum_{k=1}^{K} \lambda_{ik} P(C_k \mid \mathbf{x}) \\ &= \sum_{k \neq i} P(C_k \mid \mathbf{x}) \\ &= 1 - P(C_i \mid \mathbf{x}). \end{aligned}$$

For minimum risk, choose the most probable class.

Risks and Losses

0/1 loss with reject

- Assume mis-classification has a cost of 1 (0/1 loss).
- Assume (almost) certain classification (e.g., by a human expert) has a cost of λ.
- Define additional action REJECT α_{K+1} and loss by

$$\lambda_{ik} = \begin{cases} 0 & i = k \\ \lambda & i = K+1 \\ 1 & \text{otherwise} \end{cases} .$$

- $R(\alpha_{K+1} \mid \mathbf{x}) = \sum_{k=1}^{K} \lambda P(C_k \mid \mathbf{x}) = \lambda$.
- $R(\alpha_i \mid \mathbf{x}) = \sum_{k \neq i} P(C_k \mid \mathbf{x}) = 1 - P(C_i \mid \mathbf{x})$.

$$\text{Choose} \begin{cases} C_k & \text{if} \quad k = \arg\max_k P(C_k \mid \mathbf{x}) \text{ and } P(C_k \mid \mathbf{x}) \geq 1 - \lambda \\ \text{reject} & \text{otherwise} \end{cases}$$

Discriminant Functions

- Discriminant function: choose α_i where $i = \arg\max_k g_k(\mathbf{x})$, where

$$g_k(\mathbf{x}) = \begin{cases} -R(\alpha_k \mid \mathbf{x}) \\ P(C_k \mid \mathbf{x}) \\ p(\mathbf{x} \mid C_k)P(C_k) \end{cases}$$

- K decision regions $\mathcal{R}_1, \ldots, \mathcal{R}_K$:

$$\mathcal{R}_i = \left\{ \mathbf{x} \mid i = \arg\max_k g_k(\mathbf{x}) \right\} .$$

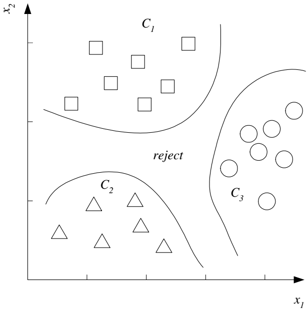


Figure 3.1 of Alpaydin (2004).

Discriminant Functions

K=2 classes

- Dichtotomizer ($K = 2$) vs. Polychotomizer ($K > 2$)
- $g(\mathbf{x}) = g_1(\mathbf{x}) - g_2(\mathbf{x})$: choose C_1 if $g(\mathbf{x}) \geq 0$, C_2 otherwise.
- Log odds:

$$g(\mathbf{x}) = \log \frac{P(C_1 \mid \mathbf{x})}{P(C_2 \mid \mathbf{x})}.$$

Utility Theory

- In utility theory, one usually tries to maximize expected utility (instead of minimize risk).
- Utility of α_i when state is k: U_{ik}

$$EU(\alpha_i \mid \mathbf{x}) = E\left[U_{ik}\right] = \sum_k U_{ik} P(C_k \mid \mathbf{x}).$$

- Choose α_i where $i = \arg\max_i EU(\alpha_i \mid \mathbf{x})$.
- (Choosing $U_{ik} = \delta_{ik} \log P(C_k \mid \mathbf{x})$ makes utility equal to information and leads to probabilistic modeling.)

Utility Theory

Value of information

- Utility of using $\mathbf{x}$ only is $EU(\mathbf{x}) = \max_i EU(\alpha_i \mid \mathbf{x})$.
- Utility of using $\mathbf{x}$ and new feature z is $EU(\mathbf{x}, z) = \max_i EU(\alpha_i \mid \mathbf{x}, z)$.
- z is useful if $EU(\mathbf{x}, z) > EU(\mathbf{x})$.
- You should probably measure z if the expected gain in utility, $EU(\mathbf{x}, z) - EU(\mathbf{x})$ exceeds the measurement costs.

Decision Theory in Court

- Classification problem GUILTY vs. NOT GUILTY.
- Typically, DNA evidence has small match probabilities. How should it be combined with other evidence?
- Sentencing innocent should have a higher loss.
- R v. Denis John Adams.

Instructions to the Jury?

Suppose the match probability is 1 in 20 million. That means that in Britain (population about 60 million) there will be on average about 2 or 3 people, and certainly no more than 6 or 7, whose DNA matches that found at the crime scene, in addition to the accused. Now your job, as a member of the jury, is to decide on the basis of the other evidence, whether or not you are satisfied that it is the person on trial who is guilty, rather than one of the few other people with matching DNA. We don't know anything about the other matching people. They are likely to be distributed all across the country and may have been nowhere near the crime scene at the time of the crime. Others may be ruled out as being the wrong sex or the wrong age group.

Outline

Graphical Models

- Graphical models are diagrammatic representations of probability distributions.
- Advantages:
  - The structure is more apparent in graphical representation.
  - Properties of the model, such as conditional independence, are easy to see.
  - Complex computations are reduced to graphical manipulations.
- Variations:
  - Bayesian networks (belief networks, probabilistic networks) [today]
  - Markov random fields
  - Factor graphs
- Applications:
  - Construction of probabilistic models
  - Biological networks (see T-61.6070 Modeling of biological networks)
  - ...

Bayesian Networks
Motivation

- How to efficiently represent joint probability distributions such as $P(\mathit{Sky}, \mathit{AirTemp}, \ldots, \mathit{Forecast}, \mathit{EnjoySport})$ (useful in computing Aldo's sport preferences $P(\mathit{EnjoySport} \mid \mathit{Sky}, \ldots, \mathit{Forecast})$)

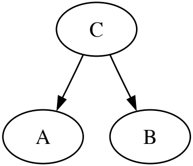
	$\mathbf{x}^t$						$r(\mathbf{x}^t)$
t	*Sky*	*AirTemp*	*Humidity*	*Wind*	*Water*	*Forecast*	*EnjoySport*
1	Sunny	Warm	Normal	Strong	Warm	Same	1
2	Sunny	Warm	High	Strong	Warm	Same	1
3	Rainy	Cold	High	Strong	Warm	Change	0
4	Sunny	Warm	High	Strong	Cool	Change	1

Table: Aldo's observed sport experiences in different weather conditions.

Bayesian Networks

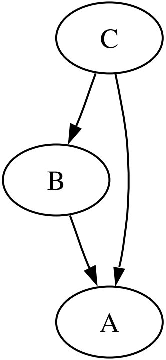
Examples

Example 1:



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

Example 2:



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

Bayesian Networks

Bayesian network is a directed acyclic graph (DAG) that describes a joint distribution over the vertices $X_1,\ldots,X_d$ such that

$$P(X_1,\ldots,X_d) = \prod_{i=1}^{d} P(X_i \mid \mathrm{parents}(X_i)),$$

where $\mathrm{parents}(X_i)$ are the set of vertices from which there is an edge to X_i.

- Example 1: $P(A,B,C) = P(A \mid C)P(B \mid C)P(C)$.
- Product rule: $P(A,B,C) = P(A,B \mid C)P(C) = P(A \mid B,C)P(B \mid C)P(C)$.
- Generally: $P(X_1,\ldots,X_d) = P(X_d \mid X_1,\ldots,X_{d-1})\ldots P(X_2 \mid X_1)P(X_1)$.
- Example 2: All joint distributions $P(X_1,\ldots,X_d)$ can be represented by a graph with $d(d-1)/2$ edges.

Outline

Causes and Bayes' Rule

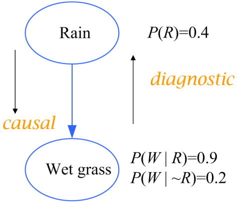
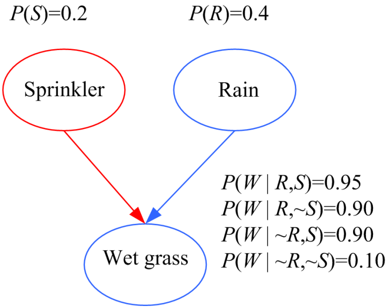


Figure 3.2 of Alpaydin (2004).
$P(W, R) = P(W \mid R)P(R)$

Diagnostic inference: Knowing that grass is wet, what is the probability that rain is the cause?

$$
\begin{aligned}
P(R \mid W) &= \frac{P(W \mid R)P(R)}{P(W)} \\
&= \frac{P(W \mid R)P(R)}{P(W \mid R)P(R) + P(W \mid \sim R)P(\sim R)} \\
&= \frac{0.9 \times 0.4}{0.9 \times 0.4 + 0.2 \times 0.6} = 0.75
\end{aligned}
$$

Causal vs. Diagnostic Inference



Causal inference: If the sprinkler is on, what is the probability that the grass is wet?

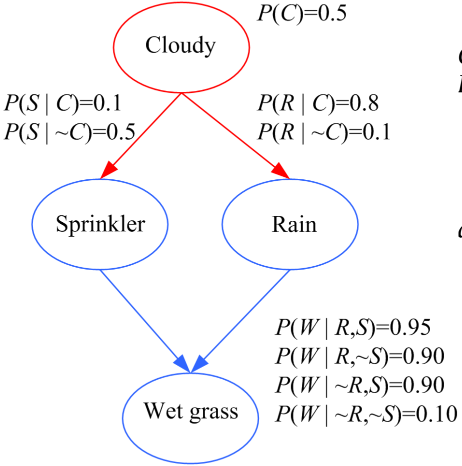
$$P(W|S) = P(W|R,S)\,P(R|S) + P(W|{\sim}R,S)\,P({\sim}R|S)$$
$$= P(W|R,S)\,P(R) + P(W|{\sim}R,S)\,P({\sim}R)$$
$$= 0.95\ 0.4 + 0.9\ 0.6 = 0.92$$

Diagnostic inference: If the grass is wet, what is the probability that the sprinkler is on? $P(S|W) = 0.35 > 0.2\ P(S)$

$P(S|R,W) = 0.21$ *Explaining away: Knowing that it has rained decreases the probability that the sprinkler is on.*

Alpaydin (2004) Ch 3 / slides

Bayesian Network: Causes



Causal inference:

$$P(W|C) = P(W|R,S)\,P(R,S|C) + P(W|\sim R,S)\,P(\sim R,S|C) + P(W|R,\sim S)\,P(R,\sim S|C) + P(W|\sim R,\sim S)\,P(\sim R,\sim S|C)$$

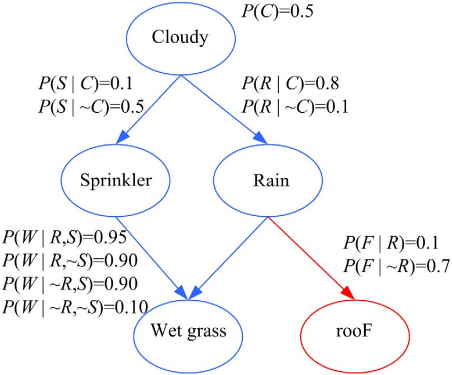
and use the fact that

$$P(R,S|C) = P(R|C)\,P(S|C)$$

Diagnostic: $P(C|W) = ?$

Alpaydin (2004) Ch 3 / slides

Bayesian Networks: Local Structure



$$P(F \mid C) = ?$$

$$P(C,S,R,W,F) = P(C)P(S \mid C)P(R \mid C)P(W \mid S,R)P(F \mid R)$$

$$P(X_1,\ldots X_d) = \prod_{i=1}^{d} P(X_i \mid \text{parents}(X_i))$$

Alpaydin (2004) Ch 3 / slides

Bayesian Networks: Inference

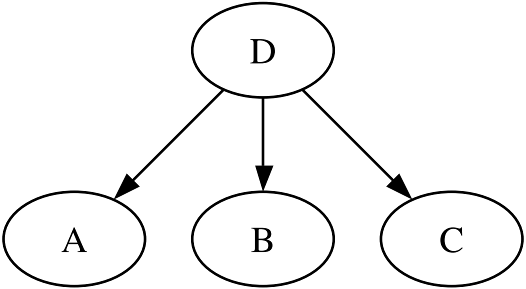
- $P(C, S, R, W, F) = P(F \mid R)P(W \mid R, S)P(R \mid C)P(S \mid C)P(C)$.
- $P(C, F) = \sum_S \sum_R \sum_W P(C, S, R, W, F)$.
- $P(F \mid C) = P(C, F)/P(C)$.
- More generally: To do inference in Bayesian networks one has to marginalize over variables.
- For example: $P(X_1) = \sum_{X_2} \cdots \sum_{X_d} P(X_1, \ldots, X_d)$.
- If we have Boolean arguments the sum has $O(2^{d-1})$ terms. This is inefficient!
- Generally, marginalization is a NP-hard problem.
- If Bayesian Network is a tree: Sum-Product Algorithm
- If Bayesian Network is "close" to a tree: Junction Tree Algorithm
- Otherwise: approximate methods (variational approximation, MCMC etc.)

Sum-Product Algorithm

- Idea: sum of products is difficult to compute. Product of sums is easy to compute, if sums have been re-arranged smartly.
- Example: disconnected Bayesian network with d vertices, computing $P(X_1)$.
  - sum of products: $P(X_1) = \sum_{X_2} \ldots \sum_{X_d} P(X_1) \ldots P(X_d)$.
  - product of sums:
    $P(X_1) = P(X_1) \left(\sum_{X_2} P(X_2)\right) \ldots \left(\sum_{X_d} P(X_d)\right) = P(X_1)$.
- Sum-Product Algorithm works if the Bayesian Network is directed tree.
- For details, see e.g., Bishop (2006).

Sum-Product Algorithm

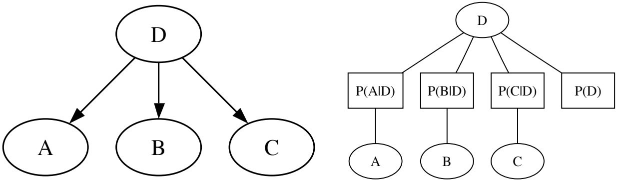
Example



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

Task: compute $\tilde{P}(D) = \sum_A \sum_B \sum_C P(A,B,C,D)$.

Sum-Product Algorithm
Example



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

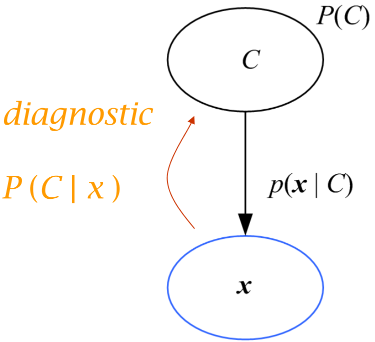
- Factor graph is composed of vertices (ellipses) and factors (squares), describing the factors of the joint probability.
- The Sum-Product Algorithm re-arranges the product (check!):

$$\begin{aligned}\tilde{P}(D) &= \left(\sum_A P(A \mid D)\right)\left(\sum_B P(B \mid D)\right)\left(\sum_C P(C \mid D)\right)P(D) \\ &= \sum_A \sum_B \sum_C P(A,B,C,D). \end{aligned} \quad (1)$$

Observations

- Bayesian network forms a partial order of the vertices. To find (one) total ordering of vertices: remove a vertex with no outgoing edges (zero out-degree) from the network and output the vertex. Iterate until the network is empty. (This way you can also check that the network is DAG.)
- If all variables are Boolean, storing a full Bayesian network of d vertices — or full joint distribution — as a look-up table takes $O(2^d)$ bytes.
- If the highest number of incoming edges (in-degree) is k, then storing a Bayesian network of d vertices as a look-up table takes $O(d2^{k+1})$ bytes.
- When computing marginals, disconnected parts of the network do not contribute.
- We can marginalize over unknown (hidden) variables.

Bayesian Network: Classification

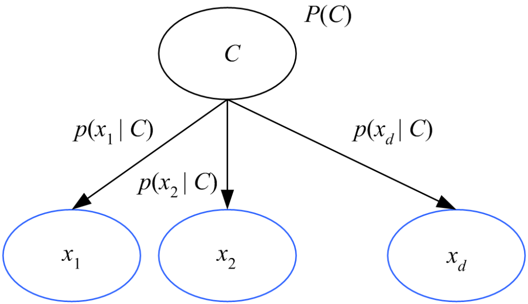


Bayes' rule inverts the arc:

$$P(C \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid C)P(C)}{p(\mathbf{x})}$$

Alpaydin (2004) Ch 3 / slides

Naive Bayes' Classifier



Given C, x_j are independent:

$$p(\mathbf{x}|C) = p(x_1|C)\ p(x_2|C)\ \dots\ p(x_d|C)$$

Alpaydin (2004) Ch 3 / slides

Outline

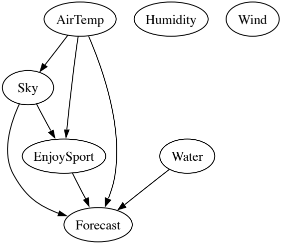
Finding a Network

- Often, the network structure is given by an expert.
- In probabilistic modeling, the network structure defines the structure of the model.
- Finding an optimal Bayesian network structure is NP-hard (given some complexity criterion, described in later lectures).

Finding a Network

- Full Bayesian network of d vertices and $d(d-1)/2$ edges describes the training set fully and the test set probably poorly.
- As before, in finding the network structure, we must control the complexity so that the the model generalizes.
- Usually one must resort to approximate solutions to find the network structure (e.g., DEAL package in R).
- A feasible exact algorithm exists for up to $d = 32$ variables, with a running time of $o(d^2 2^{d-2})$.
- See Silander et al. (2006) A Simple Optimal Approach for Finding the Globally Optimal Bayesian Network Structure. In Proc 22nd UAI. (pdf)

Finding a Network



Network found by Bene at `http://b-course.hiit.fi/bene`

t	*Sky*	*AirTemp*	*Humidity*	*Wind*	*Water*	*Forecast*	*EnjoySport*
1	Sunny	Warm	Normal	Strong	Warm	Same	1
2	Sunny	Warm	High	Strong	Warm	Same	1
3	Rainy	Cold	High	Strong	Warm	Change	0
4	Sunny	Warm	High	Strong	Cool	Change	1

Conclusion

- Next lecture on 2 October: Parametric Methods, Alpaydin (2004) Ch 4.
- Problem session on 28 September: last week's (2/2007) and this week's problem sheets (3/2007).