Machine Learning 2018 (521289S) Bayesian Networks

M.Sc. Antti Isosalo

Physiological Signal Analysis Team
Center for Machine Vision and Signal Analysis (CMVS)
University of Oulu

Discriminant Functions: Two Category Case

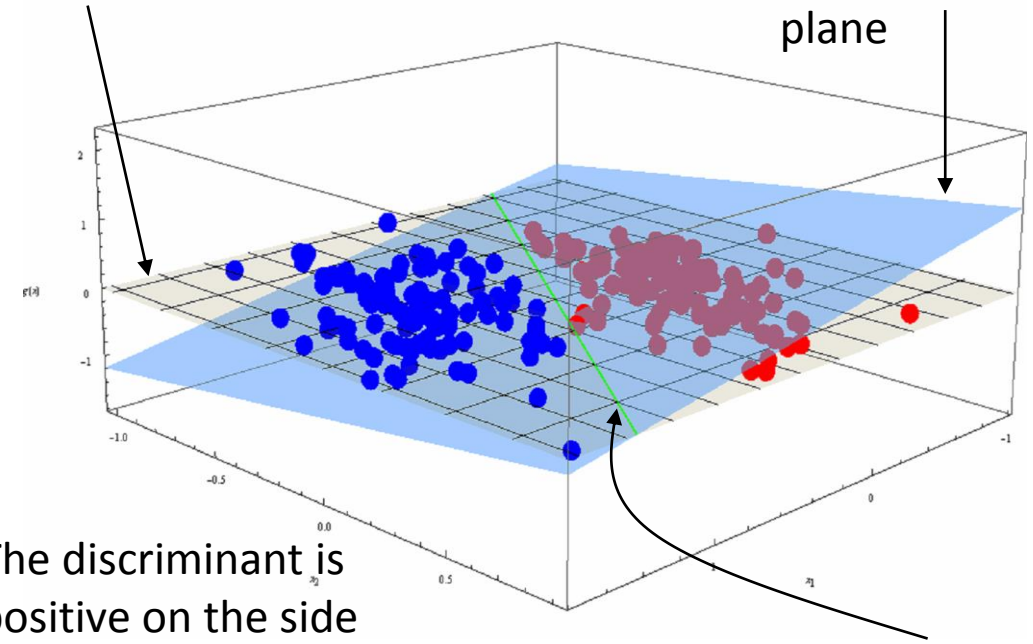
- The discriminant function is a mapping $g: \mathbb{R}^2 \to \mathbb{R}: \boldsymbol{x} = (x_1, x_2) \mapsto g(\boldsymbol{x})$, and the classifier that is based on the discriminant function is also a mapping

$$g: \mathbb{R}^2 \to \{1, 2\}: \boldsymbol{x} \mapsto \alpha(\boldsymbol{x}) = \begin{cases} 1 & g(\boldsymbol{x}) > 0 \\ 2 & \text{otherwise} \end{cases}.$$

- The discriminant function can be thought as a surface in $\mathbb{R}^3$ such that $x_3 = g\ (x_1, x_2)$, and that the feature space is embedded in the three dimensional space as the x_1x_2-plane.
- In general, such a plane is described by the equation $x_3 = g(x_1, x_2) = ax_1 + bx_2 + c$.
- The decision boundary is then the equipotential surface of points satisfying $g(\boldsymbol{x}) = 0$, in our case a line.

Feature space, 2 features have been measured

The discriminant function is a plane that intersects the feature space plane



The discriminant is positive on the side where all the red dots are and negative on the side where all the blue dots are.

On the decision boundary

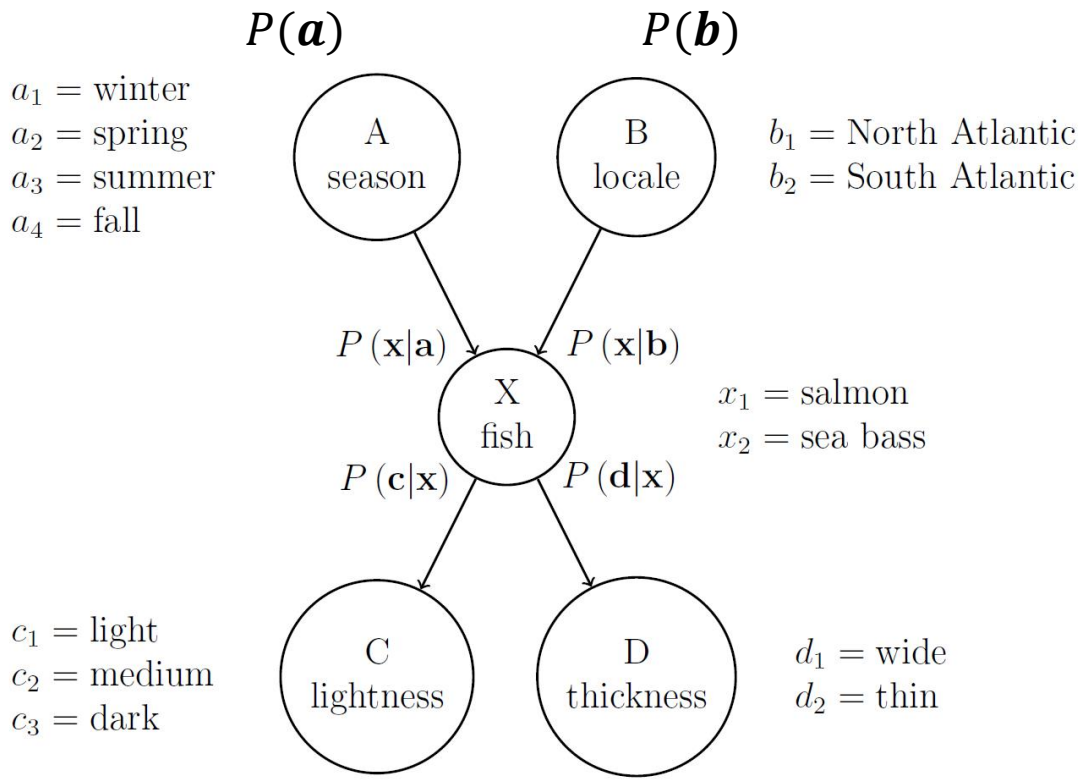
$$g(x_1, x_2) = 0$$
$$\Leftrightarrow ax_1 + bx_2 + c = 0$$

Why are Bayesian Networks Important?

- Expert knowledge is usually information about dependencies and, e.g., knowledge about what leads to a certain outcome and what not.
- When we have multiple amount of variables describing some phenomena, it becomes more and more difficult to determine the joint probability distribution from the data.
  - We would then need more samples to compensate the huge amount of variables.
  - See the Curse of Dimensionality.
- If we are able to restrict the dependencies between variables, then we have more simple estimates of the distribution to solve with less variables.
  - We are then able manage with less data.
- With networks it is possible to make indirect assumptions about certain variables by knowing some of the other variables.
- In this exercise we use Bayesian Networks in visualizing probabilistic models

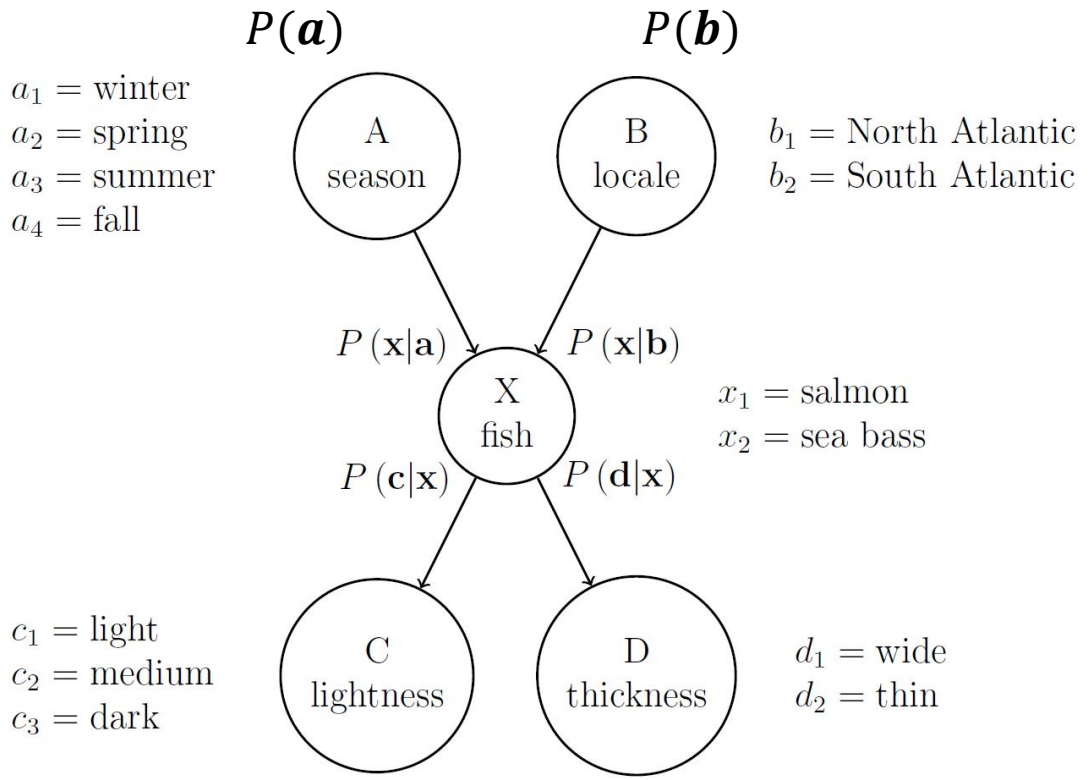
Bayesian Networks

- Bayesian Networks are directed acyclic graphs
  - Acyclic here means not forming part of a directed cycle
- A belief network consists of nodes (labeled with uppercase letters) and their associated discrete states (in lowercase)
  - Each node represents a random variable (or group of them)
- Thus node A has states $\{a_1, a_2, \dots\}$, which collectively are denoted simply $\boldsymbol{a}$; node B has states $\{b_1, b_2, \dots\}$, denoted $\boldsymbol{b}$, and so on



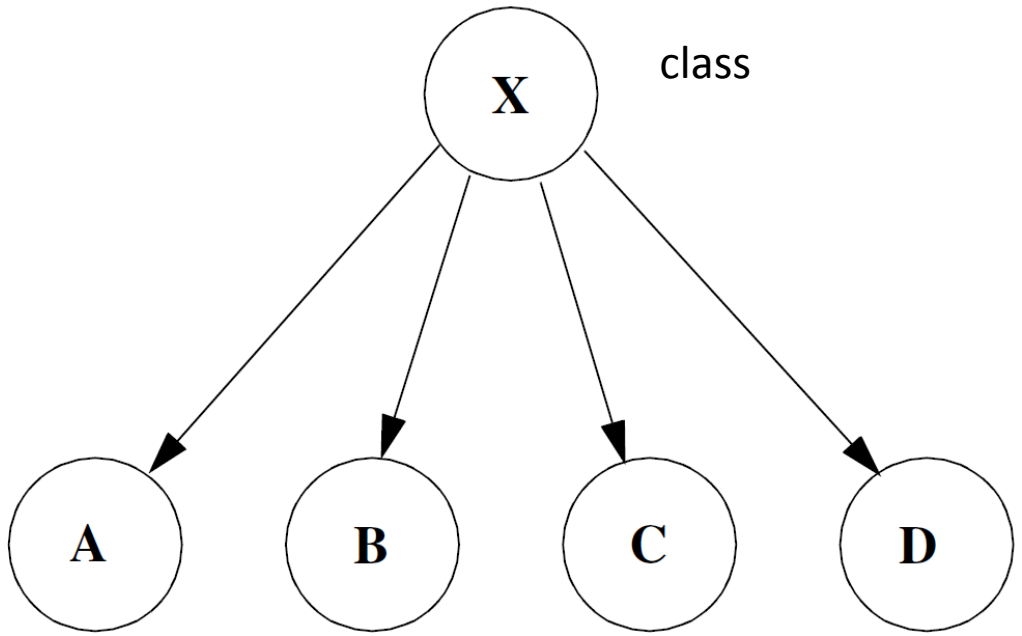
Bayesian Networks (cont.)

- The links between nodes represent direct causal influence indicated by arrows
  - For example the link from B to X represents the direct influence of B upon X
  - In this network, the variables at B may influence those at D, but only indirectly through their effect on X
- Conditional probabilities are denoted $P(\boldsymbol{x}|\boldsymbol{a})$, $P(\boldsymbol{x}|\boldsymbol{b})$, $P(\boldsymbol{c}|\boldsymbol{x})$ and $P(\boldsymbol{d}|\boldsymbol{x})$
- Simple probabilities are denoted for example $P(\boldsymbol{a})$ and $P(\boldsymbol{b})$
- In the exercise you are asked to simplify conditional independence assumptions, such as $P(\boldsymbol{c}|\boldsymbol{x},\boldsymbol{a})$



Naïve Bayes Network

- If we don't know anything about statistical dependencies describing the problem, we can use Naïve Bayes network
- Here variables are assumed conditionally independent and the net becomes particularly simple
- In the figure the root node X is, e.g., a class variable that we use to *identify the pattern* which is represented by the other variables A, B, C and D
  - We here make the assumption, that the value of a particular variable (feature) is independent of the value of any other feature, given the class variable
- Decision is based on calculating the a posterior probabilities
  - α is a scaling parameter
- Naïve Bayes Network has been successful for example in many applications



$$P(\boldsymbol{x}|\boldsymbol{a},\boldsymbol{b},\boldsymbol{c},\boldsymbol{d}) = \alpha \mathrm{P}(\boldsymbol{x})P(\boldsymbol{a}|\boldsymbol{x})P(\boldsymbol{b}|\boldsymbol{x})P(\boldsymbol{c}|\boldsymbol{x})P(\boldsymbol{d}|\boldsymbol{x})$$