SL10. Bayesian Inference

Conditional Independence:

- Conditional Independence: x is conditionally independent of y given z if the probability distribution governing x is independent of the value of y given the value of z:

$$\forall x, y, z \quad P_r(X = x \mid Y = y, Z = z) = P_r(X = x \mid Z = z)$$

More compactly:

$$P_r(X \mid Y, Z) = P_r(X \mid Z)$$

  - This means that x is conditionally independent of y given z.
  - This comes originally from normal Independence and Chain Rule:

$$P_r(X, Y) = P_r(X).P_r(Y)$$
$$P_r(X, Y) = P_r(X \mid Y).P_r(Y)$$
$$P_r(X \mid Y) = P_r(X)$$

Belief Networks:

- Belief (aka Bayesian or Graphical) Networks: A representation for probabilistic quantitates over complex spaces. It's a graphical representation of the conditional independence relationships between all the variables in a joint distribution, with nodes corresponding to the variables and edges corresponding to the dependencies.
  - Computations grow exponentially with adding more edges (dependencies).
  - A dependency relationship between two variables doesn't mean a causal relationship.
  - A Belief Network must has a topological order. We can't have cyclic relationships (two-way dependencies).
  - The Joint Probability of the graph is equal to the product of the probabilities of the variables in the graph.

$$P_r(A, B, C, D) = P_r(A).P_r(B).P_r(C).P_r(D)$$

  - Calculating independent probabilities from the graph is called sampling.

Sampling:

- Calculating independent probabilities of variables in a distribution from the graph is called sampling.
- Why sampling from a distribution is useful?
  - Simulation of a complex process.
  - Approximate inference: What might happen given some conditions?

Inferencing Rules:

- Marginalization:

$$P_r(x) = \sum_y P_r(x, y)$$

- Chain Rule:

$$P_r(x, y) = P_r(x \mid y).P_r(y)$$

- Bayes Rule:

$$P_r(y \mid x) = \frac{P_r(x \mid y)P_r(y)}{P_r(x)}$$

Naïve Bayes:

- Given a values v and attributes $a_1, a_2, \ldots, a_n$:

$$P_r(v \mid a_1, a_2, \ldots, a_n) = \frac{(\prod_i P_r(a_i \mid v)) \,.\, P_r(v)}{Z}$$

  - In a classification problem:

$$MAP\ class = \ argmax\ ((\prod_i P_r(a_i \mid v)) \,.\, P_r(v))$$

- Why Naïve Bayes is useful?
  - Inference is cheap.
  - Few parameters.
  - We can estimate the parameters with labeled data.
  - Connects inference and classification.
  - Empirically successful.
  - Handles missing attributes.
- Disadvantages:
  - Doesn't model the inner relationships between attributes.