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 Networks (aka Bayesian Networks or Probabilistic Directed Acyclic Graphical Models): 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:

$$P_r(y_1, \dots, y_n) = \prod_n P_r(y_i \mid Parents(y_i))$$

  - In belief networks, we define the Parents of a variable to be the variable's immediate predecessors in the network.
  - 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?
  - Facilitates visualizing the information provided by data.

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:

- Naïve Bayes classifiers are classifiers that represent a special case of the belief networks, but with stronger independence assumptions. For our classifier to be a naive Bayes classifier, we make the naïve assumption that every attribute variable is conditionally independent of every other attribute variable.
- For the classification variable V, we would like to find the most probable target value V_{map}, given the values for attributes $(a_1, a_2, \dots, a_n)$. We can write the expression for V_{map} and then use Bayes theorem to manipulate the expression as follows:

$$V_{map} = argmax_{v_j \in V}\, P_r(v_j \mid a_1, a_2, \dots, a_n)$$

$$V_{map} = argmax_{v_j \in V} \frac{P_r(a_1, a_2, \dots, a_n \mid v_j)P_r(v_j)}{P_r(a_1, a_2, \dots, a_n)}$$

$$V_{map} = argmax_{v_j \in V}\, P_r(a_1, a_2, \dots, a_n \mid v_j)P_r(v_j)$$

- Using the chain rule, and the naïve conditional assumption:

$$P_r(a_1, a_2, \dots, a_n \mid v_j) = P_r(a_1, a_2, \dots, a_n, v_j)P_r(a_2 \mid a_3, \dots, a_n, v_j) \dots P_r(a_n \mid v_j)$$
$$= P_r(a_1 \mid v_j)P_r(a_2 \mid v_j) \dots P_r(a_n \mid v_j)$$

- Substituting into V_{map}:

$$V_{map} = argmax_{v_j \in V}\, P_r(v_j) \prod_i (a_i \mid v_j)$$

- Why Naïve Bayes is useful?
  - Inference is cheap: Each of the terms to be estimated is a one dimensional probability, which can be estimated with a smaller data set than the joint probability.
  - Few parameters: The total number of terms to be estimated is the number of attributes n multiplied by the number of distinct values that v can take.
  - We can estimate the parameters with labeled data.
  - Connects inference and classification.
  - Empirically successful and can handle missing attribtes.
- Disadvantages:
  - Because of the strong conditional independence assumption placed on the attributes in the model, Naïve Bayes doesn't model the inner relationships between attributes.