Machine Learning - Lecture 11

Introduction to Graphical Models

16.06.2010

Bastian Leibe
RWTH Aachen
http://www.mmp.rwth-aachen.de
leibe@umic.rwth-aachen.de

Many slides adapted from B. Schiele, S. Roth

Course Outline

• Fundamentals (2 weeks)
  - Bayes Decision Theory
  - Probability Density Estimation
• Discriminative Approaches (4 weeks)
  - Lin. Discriminants, SVMs, Boosting
• Generative Models (4 weeks)
  - Bayesian Networks
  - Markov Random Fields
  - Exact Inference
  - Approximate Inference
• Unifying Perspective (2 weeks)

Graphical Models - What and Why?

• It’s got nothing to do with graphics!

• Probabilistic graphical models
  - Marriage between probability theory and graph theory.
    - Formalize and visualize the structure of a probabilistic model through a graph.
    - Give insights into the structure of a probabilistic model.
    - Find efficient solutions using methods from graph theory.
  - Natural tool for dealing with uncertainty and complexity.
  - Becoming increasingly important for the design and analysis of machine learning algorithms.
  - Often seen as new and promising way to approach problems related to Artificial Intelligence.

Topics of This Lecture

• Graphical Models
  - Introduction

• Directed Graphical Models (Bayesian Networks)
  - Notation
  - Conditional probabilities
  - Computing the joint probability
  - Factorization
  - Conditional Independence
  - D-Separation
  - Explaining away

• Outlook: Inference in Graphical Models

Graphical Models

• There are two basic kinds of graphical models
  - Directed graphical models or Bayesian Networks
  - Undirected graphical models or Markov Random Fields

• Key components
  - Nodes
  - Edges
    - Directed or undirected

Directed graphical model
Undirected graphical model

Topics of This Lecture

• Graphical Models
  - Introduction

• Directed Graphical Models (Bayesian Networks)
  - Notation
  - Conditional probabilities
  - Computing the joint probability
  - Factorization
  - Conditional Independence
  - D-Separation
  - Explaining away

• Outlook: Inference in Graphical Models
Example: Wet Lawn

- Mr. Holmes leaves his house.
  - He sees that the lawn in front of his house is wet.
  - This can have several reasons: Either it rained, or Holmes forgot to shut the sprinkler off.
  - Without any further information, the probability of both events (rain, sprinkler) increases (knowing that the lawn is wet).

- Now Holmes looks at his neighbor’s lawn
  - The neighbor’s lawn is also wet.
  - This information increases the probability that it rained. And it lowers the probability for the sprinkler.

⇒ How can we encode such probabilistic relationships?

Directed Graphical Models

- or Bayesian networks
  - Are based on a directed graph.
  - The nodes correspond to the random variables.
  - The directed edges correspond to the (causal) dependencies among the variables.
    - The notion of a causal nature of the dependencies is somewhat hard to grasp.
    - We will typically ignore the notion of causality here.
  - The structure of the network qualitatively describes the dependencies of the random variables.

Example: Wet Lawn

- Directed graphical model / Bayesian network:
  - Rain can cause both lawsns to be wet.
  - “Rain can cause both lawsns to be wet.”
  - “Holmes’ lawn may be wet due to his neighbor’s lawn may not.”

Directed Graphical Models

- Nodes or random variables
  - We usually know the range of the random variables.
  - The value of a variable may be known or unknown.
  - If they are known (observed), we usually shade the node:
    - unknown
    - known

- Examples of variable nodes
  - Binary events: Rain (yes / no), sprinkler (yes / no)
  - Discrete variables: Ball is red, green, blue, ...
  - Continuous variables: Age of a person, ...

Directed Graphical Models

- Most often, we are interested in quantitative statements
  - I.e. the probabilities (or densities) of the variables.
    - Example: What is the probability that it rained? ...
      - These probabilities change if we have:
        - more knowledge,
        - less knowledge, or
        - different knowledge
        about the other variables in the network.

Directed Graphical Models

- Simplest case:
  - $\alpha \rightarrow b$

  - This model encodes
    - The value of $b$ depends on the value of $\alpha$.
    - This dependency is expressed through the conditional probability:
      - $p(b|\alpha)$
    - Knowledge about $\alpha$ is expressed through the prior probability:
      - $p(\alpha)$
    - The whole graphical model describes the joint probability of $\alpha$ and $b$:
      - $p(\alpha, b) = p(b|\alpha)p(\alpha)$
**Directed Graphical Models**

- If we have such a representation, we can derive all other interesting probabilities from the joint.
  - E.g. marginalization
    \[
    p(a) = \sum_b p(a, b) = \sum_b p(b|a)p(a)
    \]
    \[
    p(b) = \sum_a p(a, b) = \sum_a p(a|b)p(a)
    \]
  - With the marginals, we can also compute other conditional probabilities:
    \[
    p(a|b) = \frac{p(a, b)}{p(b)}
    \]

**Example**

- Evaluating the Bayesian network...
  - We start with the simple product rule:
    \[
    p(a, b, c) = p(a|b, c)p(b|c)p(c)
    \]
  - This means that we can rewrite the joint probability of the variables as
    \[
    p(S, R, W) = p(C)p(S|C)p(R|C)p(W|S, R)
    \]
  - But the Bayesian network tells us that
    \[
    p(C, S, R, W) = p(C)p(S|C)p(R|C)p(W|S, R)
    \]
  - I.e. rain is independent of sprinkler (given the cloudyness).
  - Wet grass is independent of the cloudyness (given the state of the sprinkler and the rain).

**Directed Graphical Models**

- Chains of nodes:
  - As before, we can compute
    \[
    p(a, b) = p(b|a)p(a)
    \]
  - But we can also compute the joint distribution of all three variables:
    \[
    p(a, b, c) = p(c|a, b)p(a, b) = p(c|b)p(b)\]
  - We can read off from the graphical representation that variable \( c \) does not depend on \( a \), if \( b \) is known.
  - How? What does this mean?

**Directed Graphical Models**

- A general directed graphical model (Bayesian network) consists of
  - A set of variables: \( U = \{x_1, \ldots, x_n\} \)
  - A set of directed edges between the variable nodes.
  - The variables and the directed edges define an acyclic graph.
  - Acyclic means that there is no directed cycle in the graph.
  - For each variable \( x_i \) with parent nodes \( p_{i} \), in the graph, we require knowledge of a conditional probability:
    \[
    p(x_i|\{x_j|j \in p_{i}\})
    \]
Directed Graphical Models

- **Given**
  - Variables: \( U = \{x_1, \ldots, x_n\} \)
  - Directed acyclic graph: \( G = (V, E) \)
    - \( V \): nodes = variables, \( E \): directed edges
  
  We can express / compute the joint probability as
  \[
p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i | \{x_j | j \in \text{pa}_i\})
  \]
  We can express the joint as a product of all the conditional distributions from the parent-child relations in the graph.
  We obtain a factorized representation of the joint.

**Exercise:** Computing the joint probability

\[
p(x_1, \ldots, x_7) = ?
\]
Directed Graphical Models

- Exercise: Computing the joint probability
  \[ p(x_1, \ldots, x_n) = p(x_1)p(x_2|x_1)p(x_3|x_1, x_2, x_3) \]
  \[ p(x_5|x_1, x_3)p(x_4|x_2)p(x_1|x_4, x_3) \]

General factorization
\[ p(x) = \prod_{i=1}^{k} p(x_i|x_{pa_i}) \]

We can directly read off the factorization of the joint from the network structure!

Example: Classifier Learning

- Bayesian classifier learning
  - Given \( N \) training examples \( x = \{x_1, \ldots, x_N\} \) with target values \( t \)
  - We want to optimize the classifier \( y \) with parameters \( w \).
  - We can express the joint probability of \( t \) and \( w \):
    \[ p(t, w) = p(t|w) \prod_{i=1}^{k} p(w_i|x_{pa_i}) \]
    - Corresponding Bayesian network:
      Short notation:
      \[ \sim "Plate" \]
      (short notation for \( N \) copies)

Conditional Independence

- \( p(x_0, x_1, x_2, x_3) = p(x_3|x_0, x_1, x_2)p(x_2|x_0, x_1)p(x_1|x_0)p(x_0) \)

- Now, we can make a simplifying assumption
  - Only the previous word is what matters, i.e. given the previous word we can forget about every word before the previous one.
  - E.g. \( p(x_4|x_3, x_2, x_1) = p(x_4|x_3) \) or \( p(x_2|x_3, x_1) = p(x_2|x_3) \)
  - Such assumptions are called conditional independence assumptions.

\( \sim \) It’s the edges that are missing in the graph that are important! They encode the simplifying assumptions we make.

Factorized Representation

- Reduction of complexity
  - Joint probability of \( n \) binary variables requires us to represent values by brute force
    \[ O(2^n) \] terms
  - The factorized form obtained from the graphical model only requires
    \[ O(n \cdot 2^k) \] terms
    - \( k \): maximum number of parents of a node.

Conditional Independence

- Suppose we have a joint density with 4 variables.
  \[ p(x_0, x_1, x_2, x_3) \]

- For example, 4 subsequent words in a sentence:
  \( x_0 = "Machine", \ x_1 = "learning", \ x_2 = "is", \ x_3 = "fun" \)

- The product rule tells us that we can rewrite the joint density:
  \[ p(x_0, x_1, x_2, x_3) = p(x_0)p(x_1|x_0, x_2)p(x_2|x_0, x_1)p(x_3|x_0, x_1) \]

Conditional Independence

- The notion of conditional independence means that
  - Given a certain variable, other variables become independent.
  - More concretely here:
    \[ p(x_3|x_0, x_1, x_2) = p(x_3|x_0) \]
    - This means that \( x_3 \) is conditionally independent from \( x_0 \) and \( x_1 \) given \( x_2 \).
      \[ p(x_2|x_0, x_1) = p(x_2|x_1) \]
      - This means that \( x_2 \) is conditionally independent from \( x_0 \) given \( x_1 \).
    - Why is this?
      \[ p(x_0, x_2|x_1) = p(x_0|x_2)p(x_2|x_1) = p(x_0)p(x_2|x_1) \]
Conditional Independence - Notation

- **X** is conditionally independent of **Y** given **V**
  
  Equivalence: \( X \perp Y | V \iff p(X, Y | V) = p(X | V)p(Y | V) \)
  
  Also:
  
  \( X \perp Y | V \iff p(X, Y | V) = p(X | V)p(Y | V) \)
  
  Special case: Marginal independence
  
  \( X \perp Y \iff X \perp Y | \emptyset \iff p(X, Y) = p(X)p(Y) \)
  
  Often, we are interested in conditional independence between sets of variables:
  
  \( X \perp Y | V \iff \{ X \perp Y, \forall X \in X \text{ and } \forall Y \in Y \} \)

Conditional Independence

- Directed graphical models are not only useful...
  
  - Because the joint probability is factorized into a product of simpler conditional distributions.
  
  - But also, because we can read off the conditional independence of variables.
  
- Let’s discuss this in more detail...

First Case: “Tail-to-tail”

- Divergent model
  
  - Are \( a \) and \( b \) independent?
  
  - Marginalize out \( c \):
    
    \[ p(a, b) = \sum_c p(a, b, c) = \sum_c p(a|c)p(b|c)p(c) \]
    
    - In general, this is not equal to \( p(a)p(b) \).
      
      \( \Rightarrow \) The variables are not independent.

First Case: “Tail-to-tail”

- What about now?
  
  - Are \( a \) and \( b \) independent?
  
  - Marginalize out \( c \):
    
    \[ p(a, b) = \sum_c p(a, b, c) = \sum_c p(a|c)p(b|c)p(c) = p(a)p(b) \]
    
    \( \Rightarrow \) If there is no undirected connection between two variables, then they are independent.

First Case: Divergent (“Tail-to-Tail”)

- Let’s return to the original graph, but now assume that we observe the value of \( c \):
  
  - The conditional probability is given by:
    
    \[ p(a, b | c) = \frac{p(a, b, c)}{p(c)} = \frac{p(a|c)p(b|c)p(c)}{p(c)} = p(a|c)p(b|c) \]
    
    \( \Rightarrow \) If \( c \) becomes known, the variables \( a \) and \( b \) become conditionally independent.

Second Case: Chain (“Head-to-Tail”)

- Let us consider a slightly different graphical model:
  
  Chain graph
  
  - Are \( a \) and \( b \) independent? No!
    
    \[ p(a, b) = \sum_c p(a, b, c) = \sum_c p(b|c)p(c|a)p(a) = p(b|a)p(a) \]
    
    - If \( c \) becomes known, are \( a \) and \( b \) conditionally independent? Yes!
      
      \[ p(a, b | c) = \frac{p(a, b, c)}{p(c)} = \frac{p(a|c)p(b|c)p(c)}{p(c)} = p(a|c)p(b|c) \]
Third Case: Convergent ("Head-to-Head")

- Let’s look at a final case: Convergent graph

\[ p(a, b) = \sum_c p(a, b, c) = \sum_c p(c|a, b)p(a) = p(a)p(b) \]

- This is very different from the previous cases.
- Even though \(a\) and \(b\) are connected, they are independent.

Summary: Conditional Independence

- Three cases
  - Divergent ("Tail-to-Tail")
    - Conditional independence when \(c\) is observed.
  - Chain ("Head-to-Tail")
    - Conditional independence when \(c\) is observed.
  - Convergent ("Head-to-Head")
    - Conditional independence when neither \(c\), nor any of its descendants are observed.

D-Separation

- Definition
  - Let \(A\), \(B\), and \(C\) be non-intersecting subsets of nodes in a directed graph.
  - A path from \(A\) to \(B\) is blocked if it contains a node such that either
    - The arrows on the path meet either head-to-tail or tail-to-tail at the node, and the node is in the set \(C\), or
    - The arrows meet head-to-head at the node, and neither the node, nor any of its descendants, are in the set \(C\).
  - If all paths from \(A\) to \(B\) are blocked, \(A\) is said to be d-separated from \(B\) by \(C\).

  \[ A \perp B \mid C \]

  Read: “\(A\) is conditionally independent of \(B\) given \(C\)”.

Exercise: What is the relationship between \(a\) and \(b\)?

Explaining Away

- Let’s look at Holmes' example again:

  Observation "Holmes’ lawn is wet" increases the probability of both "Rain" and "Sprinkler".
Explaining Away

- Let’s look at Holmes’ example again:

- Observation "Holmes’ lawn is wet" increases the probability of both "Rain" and "Sprinkler".
- Also observing "Neighbor’s lawn is wet" decreases the probability for "Sprinkler".
- The "Sprinkler" is explained away.

Topics of This Lecture

- Graphical Models
  - Introduction
- Directed Graphical Models (Bayesian Networks)
  - Notation
  - Conditional probabilities
  - Computing the joint probability
  - Factorization
  - Conditional independence
  - Determination
  - Explaining away
- Outlook: Inference in Graphical Models
  - Efficiency considerations

Inference in Graphical Models

- We know
  \[ p(A, B, C, D, E) = p(A)p(B)p(C|A, B)p(D|B, C)p(E|C, D) \]
- More efficient method for \( p(AC = c) \):
  \[ p(A, C = c) = \sum_{B, D, E} p(A)p(B)p(C = c|A, B)p(D|B, C = c)p(E|C = c, D) \]

  - Rest stays the same:
    Total: 4+2+2 = 8 operations

  \[ \text{Could'n't we have got this result easier?} \]

Inference in Graphical Models

- Computing \( p(AC = c) \)
  - We know
    \[ p(A, B, C, D, E) = p(A)p(B)p(C|A, B)p(D|B, C)p(E|C, D) \]
    \[ p(A)p(B)p(C = c)\]
  - Naïve approach:
    \[ p(A, C = c) = \sum_{B, D, E} p(A, B, C = c, D, E) \] 16 operations
    \[ p(C = c) = \sum_{A, B} p(A, C = c) \] 2 operations
    \[ p(A|C = c) = \frac{p(A, C = c)}{p(C = c)} \] 2 operations

  Total: 16+2+2 = 20 operations

Inference in Graphical Models

- Consider the network structure
  - Using what we know about factorization and conditional independence...
  - Factorization properties:
    - There is no directed path from \( D \) or \( E \) to either \( A \) or \( C \).
    - We do not need to consider \( D \) and \( E \).
  - Conditional independence properties:
    - \( C \) opens the path from \( A \) to \( B \) ("head-to-head").
    - \( A \) is conditionally dependent on \( B \) given \( C \).
    - When querying for \( p(AC = c) \), we only need to take into account \( A, B, \) and \( C = c \):
      \[ p(A, C = c) = \sum_{B} p(A)p(B)p(C = c|A, B) \]
Summary

- Graphical models
  - Marriage between probability theory and graph theory.
  - Give insights into the structure of a probabilistic model.
  - Direct dependencies between variables.
  - Conditional independence
  - Allow for efficient factorization of the joint.
  - Factorization can be read off directly from the graph.
  - Capability to explain away hypotheses by new evidence.

- Next week
  - Undirected graphical models (Markov Random Fields)
  - Efficient methods for performing exact inference.

References and Further Reading

- A thorough introduction to Graphical Models in general and Bayesian Networks in particular can be found in Chapter 8 of Bishop's book.

Christopher M. Bishop
Pattern Recognition and Machine Learning
Springer, 2006