Till now - Probabilistic models for data

Model in words

Dirichlet allocation
Regression
Clustering

ORIE 4742 - Info Theory and Bayesian ML

Bayesian Networks

April 9, 2020

Sid Banerjee, ORIE, Cornell

From: Ch. 8, PRML by Chris Bishop
probabilistic graphical models

graphical representation of complex probability distributions

types of graphical models

BayesNets: directed acyclic graphs
Markov random fields: undirected graphs
factor graphs: bipartite graphs

why are they useful?

• visualizing helps in design of probabilistic models
• complex inference/learning calculations → simpler graph operations
• gives insight into properties of model: conditional independence, causal relationships
BayesNets

directed acyclic graph (DAG) encoding conditional distributions

e.g. for r.v.s $A$, $B$, $C$, BN on left encodes:

\[
p(A, B, C) = p(C|A, B)p(A, B) \\
= p(c|A, B)p(B|A)p(A)
\]
BayesNets

\[ P(x_1, x_2, \ldots, x_7) = P(x_1) \cdot P(x_2) \cdot P(x_3 | x_1, x_2, x_3) \cdot P(x_4 | x_1, x_2, x_3) \cdot \]

\[ P(x_5 | x_1, x_3) \cdot P(x_6 | x_4) \cdot P(x_7 | x_5, x_4) \]

- Any DAG has a 'topological ordering' (i.e., numbering s.t. no edge from higher to lower number), use to generate prob expansion / factorization

- For any \( x \), \( \text{Pa}(x) \equiv \text{'parents' of } x \) \cdot \( P(\bigcap_{i=1}^{n} x_i) = \prod_{i=1}^{n} P(x_i | \text{Pa}(x_i)) \)
example: (Bayesian) regression

\[
\text{Input: } (x_1, t_1), (x_2, t_2), \ldots, (x_n, t_n) \quad t
\]

\[
\text{Tasks: } t_i = \sum_{j=1}^{m} w_{ij} f_j(x_i) + w_0
\]

\[
\begin{align*}
W_i & \sim N(\mu_i, \tau_i) \\
\varepsilon & \sim N(0, \tau_e)
\end{align*}
\]

Want to learn \((w_1, w_2, \ldots, w_m)\) from data

2) Given new point \(x_{n+1}\), infer \(t_{n+1}\)

\[
\begin{align*}
E_g &: f_1(x) = 1 \quad \text{(constant)} \\
f_2(x) &= x \quad \text{(linear regression)} \\
f_3(x) &= x^{k-1} \quad \text{(polynomial regression)} \\
f_4(x) &= e^{-\frac{(x-\mu)^2}{2}} \quad \text{(Gaussian basis function)}
\end{align*}
\]
regression: basic BayesNet

\[ \mathbf{W} = \{W_0, W_1, \ldots, W_m\} \]

Data

Random vector
regression: plate notation
regression: inputs and hyperparameters

Hyperparameters are represented as solid dots (true for any ‘deterministic variable’).

'Nuisance' parameter (do not want to learn).

Prior
\[ W_i \sim N(\mu_i, \sigma_i^2) \]
\[ \mathcal{W} \{\mu_1, \sigma_1, \ldots, \mu_m, \sigma_m\} \]

Model params

\[ \sigma^2 \]
\[ W_0 \sim N(0, \sigma^2) \]
regression: learning

\[ \sigma^2 \]

\[ t_n \]

\[ N \]

\[ x_n \]

\[ \alpha \]

\[ w \]

unshaded node = latent variable

shaded node = "observed variable"
regression: prediction

\[(x_i, y_i), \ldots, (x_m, y_m)\]

\[t_i = \sum_{j=1}^{m} w_j f_j(x_i) + w_0\]

\[|t_i - y_i|\]

If \(X_1, X_2, \ldots, X_n\) were random, 
and \(\hat{x} \in \mathbb{R}^d\)

\[\hat{t} \parallel t_i \mid w\]

\[\text{Note}\]

\[X_1, X_2, \ldots, X_n\]

\[\text{are not i.i.d.}\]
example: naive Bayes

Assumption: \( X^i_j \perp \perp X^i_j \) \( \forall i, j \)

Eq: \( (X^i_1, X^i_2, \ldots, X^i_d) \sim \text{Dir}(\alpha^i_c) \)

classes: \( c \in \{1, 2, \ldots, C\} \)

data: \( (y^i, x^i_1, x^i_2, \ldots, x^i_d) \)
dictionary: \( d \in \{1, 2, \ldots, D\} \)

\( i \in \{1, 2, \ldots, N\} \)
example: naive Bayes
• Use a given Bayes Net to answer is: $A \perp B \mid C$

$\cdot (A_1, A_2) \perp (B_1, B_2, B_3) \mid (C_1, C_2, C_3)$

• $P(A, B \mid C) \overset{?}{=} P(A \mid C) P(B \mid C)$
conditional independence

- TLDR - You can answer this given a Bayes Net
- d-separation (Pearl '88)
- 3 building blocks
- Question
conditional independence: splits

\[ P(a, b, c) = P(c) \cdot P(a|c) \cdot P(b|c) \]

Without conditioning

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

\[ \Rightarrow a \perp b \]
conditional independence: splits

\[ P(a, b | c) = \frac{P(a, b, c)}{P(c)} \]

\[ = P(c) \frac{P(a | c) P(b | c)}{P(c)} \]

\[ = P(a | c) . P(b | c) \Rightarrow a \perp b | c \]
conditional independence: chains

\[ P(a, b, c) = P(a) \cdot P(c|a) \cdot P(b|c) \]

\[ P(a, b) = \sum_c P(c|a) P(b|c) \neq P(a) \cdot P(b) \]

\[ \implies a \perp b \quad (\text{if } b = c, c = a) \]
conditional independence: chains

\[ P(a, b | c) = \frac{P(a) P(c | a) P(b | c)}{P(c)} \]

\[ = P(a | c) P(b | c) \]

\[ \Rightarrow a \perp b | c \quad \text{(Markov chain)} \]
conditional independence: joins

\[ P(a, b, c) = P(a) \cdot P(b) \cdot P(c | a, b) \]

\[ \Rightarrow P(a, b) = P(a) \cdot P(b) \cdot \sum_c P(c | a, b) \]

\[ \Rightarrow a \perp b \]

\[ a \perp b \]
conditional independence: joins

\[ P(a, b | c) = P(a) P(b) \underbrace{P(c | c, b)} P(c) \]
\[ \neq P(a | c) P(b | c) \]
\[ \Rightarrow a \perp b | c \]
‘explaining away’

\[
\begin{align*}
\text{Eg} - C &= 1 \{ \text{fever + cough} \} \\
A &= 1 \{ \text{there is a burglar} \} \\
B &= 1 \{ \text{there is a raccoon} \} \\
\pi[A=1] &= 0.9, \pi[B=1] = 0.9
\end{align*}
\]

\[
\begin{align*}
\pi[B=0 | C=0] &= \frac{\pi[C=c | B=0] \pi[B=0]}{\pi[C=0]} \\ \\
\pi[B=0 | C=0, A=0] &= \frac{\pi[C=c | B=0, A=0] \pi[B=0]}{\sum_{a,i} \pi[A=i] \pi[C=c | A=i, B=0]}
\end{align*}
\]
‘explaining away’

Diagram:

- Nodes: a, b, c
- Edges: a to b, b to c, c to b, c to a
d-separation

- \( A \perp\!
\perp B \mid C \) if \( C \) is not a join or a descendant of a join

- A path from \( A \rightarrow B \) is blocked by \( C \) if
  i) \( \rightarrow C \) or \( \leftarrow C \)
  ii) \( C \) is not \( \rightarrow \) or a descendant of \( \leftarrow \)

A, B are d-separated by \( C \) if every path \( A \rightarrow B \) is blocked by \( C \)
d-separation: i.i.d. data

- $X_i \perp\!\!\!\!\!\!\perp X_j \mid \mu$
- $X_i \not\perp\!\!\!\!\!\!\perp X_j$

$\mu$ blocks path

$\mu$ does not block path
d-separation: example

Q: Is
i) $A \perp B$?
ii) $A \perp B \mid C$?

Ans: $A \not\perp B$
$A \perp B \mid C$
Q: Is $A \perp B \mid F$?

Ans: $A \perp B \mid F$
d-separation: model parameters

\[(\alpha^c_1, \alpha^c_2 - \alpha^c_d) \perp \perp (\alpha^c_i, \alpha^c_j - d^c) \mid \{y^i, x^i_1, \ldots, x^i_m\}\]
example: naive Bayes
example: naive Bayes
d-separation: model parameters
the Markov blanket
Markov random fields

- Bayes Nets encode ‘local conditioning’
  \[
  \prod_{i} P(x_i | Pa(x_i))
  \]
- Don’t directly capture global conditional independence
cliques and potentials

- Maximal cliques - set of nodes which form a clique and are not subsets of a larger 'selected' clique
- Clique cover - collection of (maximal) cliques st. every edge is in a clique

\[ P(x_1, x_2, x_3, x_4) = \frac{1}{Z} \cdot \prod_{\text{cliques } C} \Psi_c(x_i: i \in C) \]
conditional independence and Markov blanket in MRF

\[ C = \{ \{ x_1, x_2 \}, \{ x_1, x_4 \}, \{ x_2, x_3, x_4 \} \} \]

\[ P(x_1, x_2, x_3, x_4) = \frac{1}{\psi_{12}(x_1, x_2) \psi_{14}(x_1, x_4)} \psi_{234}(x_2, x_3, x_4) \]

- Conditional indep \( \iff \) separation

Q: \( X_3 \perp \!\!\!\!\!\!\!\!\!\perp X_1 \mid X_4, X_2 \) ?

Yes as \( (X_2, X_4) \) separate \( X_1 \) and \( X_3 \)
\[ P(x) = p(x_1) \prod p(x_i | x_{i-1}) \]

Choose \( \psi_{12}(x_1, x_2) = p(x_1) p(x_2 | x_1) \)
\( \psi_{23}(x_2, x_3) = p(x_3 | x_2) \)
\[ \vdots \]
\( \psi_{N-1,N}(x_{N-1}, x_N) = p(x_N | x_{N-1}) \)
$P(h) = P(x_1)P(x_2)P(x_3)P(x_4|x_1,x_2,x_3)$

$C = \{\{x_1\}, \{x_2\}, \{x_3\}, \{x_1, x_2, x_3, x_4\}\}$

$\psi_1(x_1) = p(x_1), \ldots, \psi_{1234}(x_1,x_2,x_3,x_4)$

$\mathcal{C} = \{x_1x_2x_3\} \{x_1x_2x_4\} \{x_2x_3x_4\}$

$\frac{1}{2} \psi_{123}(x_1x_2x_3) \psi_{124}(x_1x_2x_4) \psi_{234}(x_2x_3x_4)$
converting BayesNets to MRFs

- 'moralization'

- Add edges between unconnected parents of each child node
d-separation and moralization

- Is $A \perp B | C$?
  - Convert Bayes Net of ‘ancestors of C’ into MRF (via moralization)
  - Check for conditional independence
Markov chain

\[ E \]

\[ X_1 \xrightarrow{\text{indep}} X_2 \xrightarrow{\phantom{\text{indep}}} X_3 \xrightarrow{\text{indep}} X_4 \]

\[ X_i \perp X_j \mid X_k \quad \text{if} \quad i < k < j \quad \text{or} \quad i > k > j \]

\[ E \]

\[ A, B \perp C, D \mid F, G \]

\[ C \perp D \mid F, G \]

\[ A \not\perp B \mid F, G \]

\[ E \perp C, D \mid F, G \]