Learning Directed Probabilistic Models

Special Bonus Topic: Markov chains for sequence data!

Daniel Lowd
CIS 473/573

What is P(‘covfefe’)?

Donald J. Trump @realDonaldTrump

Despite the constant negative press covfefe

RETWEETS: 127,484  LIKES: 162,762

9:06 PM – 30 May 2017
What is $P(\text{Yrvjseubpihfcovswtvnjhfefesxnklmnoq})$?

Key Questions

1. **Representation**: How do we represent a probability distribution over words that aren’t in the dictionary?
2. **Learning**: How do we choose the parameters for that probability distribution?
3. **Inference**: How do we use this probability distribution to compute $P(\text{covfefe})$?
Representation: Markov Chain

• We can use a Markov chain to describe a probability distribution over sequences:

\[ P(X_1, ..., X_n) = P(X_1)P(X_2 | X_1) P(X_3 | X_2, X_1) ... P(X_n | X_{n-1}, ..., X_1) \]

• First-order Markov property: next state \( X_t \) only depends on previous state \( X_{t-1} \):

\[ P(X_{t+1} | X_t, ..., X_1) = P(X_{t+1} | X_t) \]

\[ = P(X_1)P(X_2 | X_1) P(X_3 | X_2) P(X_4 | X_3) ... P(X_n | X_{n-1}) \]
\[ = P(X_1) \prod_{t=2}^{n} P(X_t | X_{t-1}) \]

Markov Chain as a BN

• CPDs:
  • \( P(X_1) \): Distribution over starting state
  • \( P(X_t | X_{t-1}) \): Transition probabilities

• Q2: How do we estimate these CPDs?
• Q3: How do we compute \( P('covfefe') \)?
Basics of Statistical Estimation

Learning Probabilities:
Classical Approach

Simplest case: Flipping a thumbtack

Given iid data, estimate $\theta$ using an estimator with good properties: low bias, low variance, consistent (e.g., maximum likelihood estimate)
Maximum Likelihood Principle

Choose the parameters that maximize the probability of the observed data

Maximum Likelihood Estimation

\[ p(\text{heads} \mid \theta) = \theta \]
\[ p(\text{tails} \mid \theta) = (1 - \theta) \]
\[ p(\text{hhth...ttth} \mid \theta) = \theta^h (1 - \theta)^t \]

(Number of heads is binomial distribution)
Computing the ML Estimate

• Use log-likelihood
• Differentiate with respect to parameter(s)
• Equate to zero and solve
• Solution:

$$\theta = \frac{\# h}{\# h + \# t}$$

Sufficient Statistics

$$p(hhth...ttth \mid \theta) = \theta^{\# h} (1 - \theta)^{\# t}$$

(#h,#t) are sufficient statistics
Bayesian Estimation

True probability $\theta$ is unknown

Bayesian probability density for $\theta$

$\int_{\theta} p(\theta) d\theta$

Use of Bayes’ Theorem

$p(\theta | \text{heads}) = \frac{p(\theta)p(\text{heads} | \theta)}{\int p(\theta)p(\text{heads} | \theta) d\theta}$

$\propto p(\theta)p(\text{heads} | \theta)$
Example: Application to Observation of Single “Heads”

\[ p(θ) \times p(\text{heads}|θ) = θ \times \infty \]

prior  
likelihood  
posterior

Probability of Heads on Next Toss

\[
p(n+1 \text{th toss is } h \mid d) = \int p(X_{N+1} = h \mid θ) p(θ | d) dθ \\
= \int θ \ p(θ | d) \ dθ \\
= E_{p(θ | d)}(θ)
\]
MAP Estimation

• Approximation:
  • Instead of averaging over all parameter values
  • Consider only the most probable value
    (i.e., value with highest posterior probability)
• Usually a very good approximation, and much simpler
• MAP value ≠ Expected value
• MAP → ML for infinite data
  (as long as prior ≠ 0 everywhere)

Prior Distributions for θ

• Direct assessment
• Parametric distributions
  • Conjugate distributions
    (for convenience)
  • Mixtures of conjugate distributions
Conjugate Family of Distributions

**Beta distribution:**

\[ p(\theta) = \text{Beta}(\alpha_h, \alpha_t) \propto \theta^{\alpha_h - 1} (1 - \theta)^{\alpha_t - 1} \quad \alpha_h, \alpha_t > 0 \]

**Resulting posterior distribution:**

\[ p(\theta \mid h \text{ heads}, t \text{ tails}) \propto \theta^\#h + \alpha_h - 1 (1 - \theta)^\#t + \alpha_t - 1 \]

---

Estimates Compared

- **Prior prediction:**
  \[ E(\theta) = \frac{\alpha_h}{\alpha_h + \alpha_t} \]

- **Posterior prediction:**
  \[ E(\theta) = \frac{\#h + \alpha_h}{\#h + \alpha_h + \#t + \alpha_t} \]

- **MAP estimate:**
  \[ \theta = \frac{\#h + \alpha_h - 1}{\#h + \alpha_h - 1 + \#t + \alpha_t - 1} \]

- **ML estimate:**
  \[ \theta = \frac{\#h}{\#h + \#t} \]
Intuition

- The hyperparameters $\alpha_h$ and $\alpha_t$ can be thought of as imaginary counts from our prior experience, starting from "pure ignorance"
- Equivalent sample size = $\alpha_h + \alpha_t$
- The larger the equivalent sample size, the more confident we are about the true probability

Beta Distributions

- $\text{Beta}(3, 2)$
- $\text{Beta}(0.5, 0.5)$
- $\text{Beta}(1, 1)$
- $\text{Beta}(19, 39)$
Assessment of a Beta Distribution

Method 1: Equivalent sample
- assess $\alpha_h$ and $\alpha_t$
- assess $\alpha_h+\alpha_t$ and $\alpha_h/(\alpha_h+\alpha_t)$

Method 2: Imagined future samples

$p(\text{heads}) = 0.2$ and $p(\text{heads} \mid 3 \text{ heads}) = 0.5 \Rightarrow \alpha_h = 1, \alpha_t = 4$

check: $0.2 = \frac{1}{1+4}$, $0.5 = \frac{1+3}{1+3+4}$

Generalization to $m$ Outcomes (Multinomial Distribution)

Dirichlet distribution:

$$p(\theta_1, \ldots, \theta_m) = \text{Dirichlet}(\alpha_1, \ldots, \alpha_m) \propto \prod_{i=1}^{m} \theta_i^{\alpha_i-1}$$

$$\sum_{i=1}^{m} \theta_i = 1 \quad \alpha_i > 0$$

Properties:

$$E(\theta_i) = \frac{\alpha_i}{\sum_{i=1}^{m} \alpha_i}$$

$$p(\theta \mid N_1, \ldots, N_m) \propto \prod_{i=1}^{m} \theta_i^{\alpha_i+N_i-1}$$
Other Distributions

Likelihoods from the exponential family
- Binomial
- Multinomial
- Poisson
- Gamma
- Normal

Learning Bayesian Networks From Data
**Dimensions of Learning**

<table>
<thead>
<tr>
<th>Model</th>
<th>Bayes net</th>
<th>Markov net</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data</td>
<td>Complete</td>
<td>Incomplete</td>
</tr>
<tr>
<td>Structure</td>
<td>Known</td>
<td>Unknown</td>
</tr>
<tr>
<td>Objective</td>
<td>Generative</td>
<td>Discriminative</td>
</tr>
</tbody>
</table>

**Learning Bayes nets from data**

<table>
<thead>
<tr>
<th>( X_1 )</th>
<th>( X_2 )</th>
<th>( X_3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>true</td>
<td>1</td>
<td>0.7</td>
</tr>
<tr>
<td>false</td>
<td>5</td>
<td>-1.6</td>
</tr>
<tr>
<td>false</td>
<td>3</td>
<td>5.9</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>true</td>
<td>2</td>
<td>6.3</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

+ prior/expert information

Bayes-net learner

Bayes net(s)
From thumbtacks to Bayes nets

Thumbtack problem can be viewed as learning the probability for a very simple BN:

\[ X \text{ heads/tails} \]

\[ \Theta \]

\[ X_1 \text{ toss 1} \]
\[ X_2 \text{ toss 2} \]
\[ \ldots \]
\[ X_N \text{ toss } N \]

The next simplest Bayes net

\[ X \text{ heads/tails} \]

\[ Y \text{ heads/tails} \]

\[ \text{heads} \]
\[ \text{tails} \]

\[ “\text{heads}” \]
\[ “\text{tails}” \]
The next simplest Bayes net

The next simplest Bayes net

"parameter independence"
The next simplest Bayes net

The next simplest Bayes net

A bit more difficult...

Three probabilities to learn:

- $\theta_{X=\text{heads}}$
- $\theta_{Y=\text{heads}|X=\text{heads}}$
- $\theta_{Y=\text{heads}|X=\text{tails}}$
A bit more difficult...

\[ X \quad \Theta_X \quad \Theta_{Y|X=\text{heads}} \quad \Theta_{Y|X=\text{tails}} \quad Y \]

\[ X_1, X_2, \ldots \]

\[ Y_1, Y_2, \ldots \]

\[ \text{heads} \quad \text{tails} \quad \text{heads} \quad \text{tails} \]

\[ \text{case 1} \quad \text{case 2} \]
A bit more difficult...

heads/tails $X$ heads/tails $Y$

$\Theta_X$ $\Theta_{Y|X=\text{heads}}$ $\Theta_{Y|X=\text{tails}}$

$X_1 \rightarrow Y_1$

$X_2 \rightarrow Y_2$

case 1

case 2

3 separate thumbtack-like problems

A bit more difficult...

heads/tails $X$ heads/tails $Y$

$\Theta_X$ $\Theta_{Y|X=\text{heads}}$ $\Theta_{Y|X=\text{tails}}$

$X_1 \rightarrow Y_1$

$X_2 \rightarrow Y_2$

case 1

case 2
In general ...

Learning probabilities in a Bayes net is straightforward if
• Complete data
• Local distributions from the exponential family (binomial, Poisson, gamma, ...)
• Parameter independence
• Conjugate priors

Back to $P(\text{`covfefe'})$...
Markov Chain as a BN

- CPDs:
  - $P(X_1)$: Distribution over starting state
  - $P(X_t | X_{t-1})$: Transition probabilities

- Q2: How do we estimate these CPDs?
- Q3: How do we compute $P(\text{`covfefe'})$?

Estimating $P(X_1)$

- Suppose we have a list of words, sampled from the distribution we want to estimate:
  - “ant”
  - “bear”
  - “cat”
  - “dog”
  - “elk”
  - “emu”
  - “goat”

- What is the probability that a randomly chosen word from this list starts with ‘a’, ‘b’, ‘z’?
- What is the probability that a randomly chosen word from this distribution starts with ‘a’, ‘b’, ‘z’?
Estimating $P(X_t | X_{t-1})$

- Suppose we have a list of words, sampled from the distribution we want to estimate:
  “ant”
  “bear”
  “cat”
  “dog”
  “elk”
  “emu”
  “goat”

- What is the probability that a letter ‘e’ is followed by ‘a’ in this list?
- What is the probability that a letter ‘e’ is followed by ‘a’ in this distribution?

More On Learning BNs...
Incomplete data makes parameters dependent

Solution: Use EM

- Initialize parameters ignoring missing data
- **E step**: Infer missing values using current parameters
- **M step**: Estimate parameters using completed data
- Can also use gradient descent
Learning Bayes-net structure

Given data, which model is correct?

model 1: $X \quad Y$

model 2: $X \rightarrow Y$

Bayesian approach

Given data, which model is correct? more likely?

model 1: $X \quad Y \quad p(m_1) = 0.7 \quad p(m_1 | d) = 0.1$

model 2: $X \rightarrow Y \quad p(m_2) = 0.3 \quad p(m_2 | d) = 0.9$
Bayesian approach: Model averaging

Given data, which model is correct? more likely?

model 1: \( \begin{align*} X & \quad Y \\ p(m_1) = 0.7 & \quad p(m_1 \mid d) = 0.1 \end{align*} \)

model 2: \( \begin{align*} X & \quad Y \\ p(m_2) = 0.3 & \quad p(m_2 \mid d) = 0.9 \end{align*} \)

average predictions

Bayesian approach: Model selection

Given data, which model is correct? more likely?

model 1: \( \begin{align*} X & \quad Y \\ p(m_1) = 0.7 & \quad p(m_1 \mid d) = 0.1 \end{align*} \)

model 2: \( \begin{align*} X & \quad Y \\ p(m_2) = 0.3 & \quad p(m_2 \mid d) = 0.9 \end{align*} \)

Keep the best model:
- Explanation
- Understanding
- Tractability
To score a model, use Bayes’ theorem

Given data \( d \):

\[
p(m \mid d) \propto p(m) p(d \mid m)
\]

"marginal likelihood"

\[
p(d \mid m) = \int p(d \mid \theta, m) p(\theta \mid m) d\theta
\]

Thumbtack example

\[
\begin{align*}
p(d \mid m) &= \int \theta^h (1 - \theta)^t p(\theta \mid m) d\theta \\
&= \int \theta^{h + \alpha_h^{-1}} (1 - \theta)^{t + \alpha_t^{-1}} d\theta \\
&= \frac{\Gamma(\alpha_h + \alpha_t)}{\Gamma(\alpha_h + \alpha_t + \#h + \#t)} \frac{\Gamma(\alpha_h + \#h)}{\Gamma(\alpha_h)} \frac{\Gamma(\alpha_t + \#t)}{\Gamma(\alpha_t)}
\end{align*}
\]
More complicated graphs

Heads/tails $X$ Heads/tails $Y$

3 separate thumbtack-like learning problems

$$p(d | m) = \frac{\Gamma(\alpha_x + \alpha_y)}{\Gamma(\alpha_x + \alpha_x + \# h + \# t)} \frac{\Gamma(\alpha_h + \# h)}{\Gamma(\alpha_h)} \frac{\Gamma(\alpha_x + \# t)}{\Gamma(\alpha_x)}$$

$\xrightarrow{X}$ $X$|$X=heads$ $\xrightarrow{Y}$ $Y$|$X=tails$

Model score for a discrete Bayes net

$$p(d | m) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + N_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + N_{ijk})}{\Gamma(\alpha_{ijk})}$$

$N_{ijk}$: # cases where $X_i = x_i^k$ and $Pa_i = pa_i^j$

$r_i$: number of states of $X_i$

$q_i$: number of instances of parents of $X_i$

$$\alpha_{ij} = \sum_{k=1}^{r_i} \alpha_{ijk} \quad N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$$
Computation of marginal likelihood

Efficient closed form if

• Local distributions from the exponential family (binomial, poisson, gamma, ...)
• Parameter independence
• Conjugate priors
• No missing data (including no hidden variables)

Structure search

• Finding the BN structure with the highest score among those structures with at most $k$ parents is NP hard for $k>1$ (Chickering, 1995)
• Heuristic methods
  • Greedy
  • Greedy with restarts
  • MCMC methods

流程图说明：
初始化结构 -> 计算所有可能的单个变化 -> 任何变化更好？
是 -> 执行最佳更改
否 -> 返回保存的结构
Structure priors

1. All possible structures equally likely
2. Partial ordering, required / prohibited arcs
3. Prior(m) $\propto$ Similarity(m, prior BN)

Parameter priors

- All uniform: Beta(1,1)
- Use a prior Bayes net
Parameter priors

Recall the intuition behind the Beta prior for the thumbtack:

• The hyperparameters $\alpha_h$ and $\alpha_t$ can be thought of as imaginary counts from our prior experience, starting from "pure ignorance"
• Equivalent sample size = $\alpha_h + \alpha_t$
• The larger the equivalent sample size, the more confident we are about the long-run fraction

Parameter priors for any Bayes net structure for $X_1...X_n$
Combining knowledge & data

prior network+equivalent sample size

x

1

x

2

x

3

x

4

x

5

x

6

x

7

x

8

x

9

improved network(s)

data

<table>
<thead>
<tr>
<th>x₁</th>
<th>x₂</th>
<th>x₃</th>
</tr>
</thead>
<tbody>
<tr>
<td>true</td>
<td>false</td>
<td>true</td>
</tr>
<tr>
<td>false</td>
<td>false</td>
<td>true</td>
</tr>
<tr>
<td>false</td>
<td>false</td>
<td>false</td>
</tr>
<tr>
<td>true</td>
<td>true</td>
<td>false</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Maximum Likelihood Principle

- We don’t know the true distribution – we can only guess. How do we know which guesses are good?
- **Likelihood**: \( P(\text{data} \mid \text{model}) \)
  (What’s the probability that a given model will generate a given output?)
- **Maximum likelihood principle**: Choose the model with highest probability of generating this data
  \[
  \hat{\text{model}} = \text{argmax}_{\text{model}} P(\text{data} \mid \text{model})
  \]
- **Example**: Assuming data is i.i.d. (independent and identically distributed),
  \[
  P(\text{data}) = P(a')P(b')P(a')P(b')P(b') = P(a')^2P(b')^3
  \]
Maximum Likelihood Parameter Estimation

- Let $P('a') = \theta, P('b') = 1 - \theta$
- $\hat{\theta} = \arg\max_{\theta} P('a', 'b', 'a', 'b', 'b'; \theta) = \arg\max_{\theta} P('a'; \theta)^2 P('b'; \theta)^3 = \arg\max_{\theta} \theta^2 (1 - \theta)^3$
- Differentiate, set to zero, solve:
  $\hat{\theta} = \frac{2}{5}$, $1 - \hat{\theta} = \frac{3}{5}$
- General case:
  $\hat{\theta} = \arg\max_{\theta} \theta^\alpha (1 - \theta)^\beta = \alpha / (\alpha + \beta)$