Learning Directed Probabilistic Models

Special Bonus Topic: Markov chains for sequence data!

Daniel Lowd
CIS 473/573

What is P('covfefe')?

What is P('Yrsvjseubpihcovswnjghgfefxnlmnq')?

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('covfefe')?

Representation: Markov Chain

- We can use a Markov chain to describe a probability distribution over sequences:
  \[
  P(X_1, \ldots, X_n) = P(X_1)P(X_2|X_1) \cdots P(X_n|X_{n-1}, \ldots, X_1) \]
- First-order Markov property: next state $X_t$ only depends on previous state $X_{t-1}$:
  \[
  P(X_1)P(X_2|X_1) \cdots P(X_n|X_{n-1}) = P(X_1) \prod_{t=2}^{n} P(X_t|X_{t-1})
  \]

Markov Chain as a BN

- **CPDs**:
  - $P(X_i)$: 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

$$p(\text{heads} \mid \theta) = \theta$$
$$p(\text{tails} \mid \theta) = (1 - \theta)$$
$$p(\text{hhth...tth} \mid \theta) = \theta^{#h} (1 - \theta)^{#t}$$

(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

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

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

True probability $\theta$ is unknown

Bayesian probability density for $\theta$

Example: Application to Observation of Single “Heads”

Probability of Heads on Next Toss

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 $\neq$ Expected value
  • MAP $\rightarrow$ ML for infinite data (as long as prior $\neq 0$ everywhere)

Prior Distributions for $\theta$

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

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

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

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
\[ p(\text{heads}) = 0.2 \quad \text{and} \quad p(\text{heads} | 3 \text{ heads}) = 0.5 \Rightarrow \alpha_h = 1, \alpha_t = 4 \]
check: \[ 0.2 = \frac{1}{1+4}, \quad 0.5 = \frac{1+3}{1+3+4} \]

Generalization to \( m \) Outcomes (Multinomial Distribution)

Dirichlet distribution:
\[ p(\theta_1, \ldots, \theta_m) = \text{Dirichlet}(\alpha_1, \ldots, \alpha_m) \propto \prod_{i=1}^{m} \theta_i^{\alpha_i - 1} \quad \sum_{i=1}^{m} \theta_i = 1, \quad \alpha_i > 0 \]

Properties:
\[ E(\theta_i) = \frac{\alpha_i}{\sum_{i=1}^{m} \alpha_i} \]
\[ p(\theta | N_1, \ldots, N_m) \propto \prod_{i=1}^{m} \theta_i^{\alpha_i + N_i - 1} \]
Other Distributions
Likelihoods from the exponential family
- Binomial
- Multinomial
- Poisson
- Gamma
- Normal

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 Bayesian Networks From Data

From thumbtacks to Bayes nets
Thumbtack problem can be viewed as learning the probability for a very simple BN:

\[
X \rightarrow \Theta \\
X_1 \quad X_2 \quad \ldots \quad X_N
\]

The next simplest Bayes net

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

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

X

heads/tails

Y

heads/tails

θ₁

θ₂

X₁

X₂

Xₙ

Y₁

Y₂

Yₙ

case 1

case 2

case N

The next simplest Bayes net

heads/tails X

heads/tails Y

"parameter independence"

two separate thumbtack-like learning problems

X₁

X₂

Xₙ

Y₁

Y₂

Yₙ

case 1

case 2

case N

A bit more difficult...

heads/tails X

heads/tails Y

Three probabilities to learn:

• \( \theta_{X=heads} \)

• \( \theta_{Y=heads|X=heads} \)

• \( \theta_{Y=heads|X=tails} \)

A bit more difficult...

heads/tails X

heads/tails Y

θ₁

θ₂

θ₃

θ₄

θ₅

θ₆

X₁

X₂

Xₙ

Y₁

Y₂

Yₙ

case 1

case 2

A bit more difficult...

heads/tails X

heads/tails Y

θ₁

θ₂

θ₃

θ₄

θ₅

θ₆

X₁

X₂

Xₙ

Y₁

Y₂

Yₙ

case 1

case 2
A bit more difficult...

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('covfefe')...

Markov Chain as a BN

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"
  "eik"
  "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

Bayesian approach

Given data, which model is correct? more likely?

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

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

Given data, which model is correct? more likely?

\[ p(m_1 | d) = 0.1 \]
\[ p(m_2 | d) = 0.9 \]

Bayesian approach: Model selection

Given data, which model is correct? more likely?

\[ p(m_1 | d) = 0.1 \]
\[ p(m_2 | d) = 0.9 \]

To score a model, use Bayes’ theorem

Given data \( d \):

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

"marginal likelihood"

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

Thumbtack example

\[
p(d | m) = \int \theta^{a_0} (1 - \theta)^{a_†} p(\theta | m) d\theta
\]
\[
= \int \theta^{a_0 + a_i - 1} (1 - \theta)^{a_† + a_i - 1} d\theta
\]

conjugate prior

\[
= \frac{\Gamma(a_0 + a_i)}{\Gamma(a_0 + a_i + \# h + \# t)} \frac{\Gamma(a_† + \# h)}{\Gamma(a_† + \# t)} \frac{\Gamma(a_†)}{\Gamma(a_†)}
\]

More complicated graphs

3 separate thumbtack-like learning problems

\[
p(d | m) = \frac{\Gamma(a_0 + a_i)}{\Gamma(a_0 + a_i + \# h + \# t)} \frac{\Gamma(a_† + \# h)}{\Gamma(a_† + \# t)} \frac{\Gamma(a_†)}{\Gamma(a_†)}
\]

\[ X \quad Y \quad X \quad Y \]

heads/tails heads/tails

Model score for a discrete Bayes net

\[
p(d | m) = \prod_{i=1}^n \prod_{j=1}^q \frac{\Gamma(\alpha_{ij} + N_{ij})}{\Gamma(\alpha_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{jk} + N_{jk})}{\Gamma(\alpha_{jk})}
\]

\[ N_{\alpha i j}: \# \text{ cases where } X_i = x_i \text{ and } \mathbf{P}a_i = \mathbf{p}_i \]
\[ r_i: \text{ number of states of } X_i \]
\[ q_i: \text{ number of instances of parents of } X_i \]
\[ \alpha_{ij} = \sum_{k=1}^{r_i} \alpha_{jk} \quad N_{ij} = \sum_{k=1}^{r_i} N_{jk} \]
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

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

\[
\begin{array}{c}
\text{equivalent sample size} \\
\rightarrow \\
\text{imaginary count for any variable configuration}
\end{array}
\]
Combining knowledge & data

prior network=equivalent sample size

data

Improved network(s)

<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>false</td>
</tr>
<tr>
<td>true</td>
<td>true</td>
<td>false</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: \( P(\text{data} | \text{model}) \)
  (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
  \( \text{model} = \text{argmax}_{\text{model}} P(\text{data} | \text{model}) \)
• Example: Assuming data is i.i.d. (independent and identically distributed),
  \( P(\text{data}) = P(a')P(c')P(a')P(b')P(b') = P(a')P(b') \)

Maximum Likelihood Parameter Estimation

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