CSE 40171:
Artificial Intelligence

Probabilistic Read-Out Layers for Artificial Neural Networks:
Bayesian Hierarchical Modeling

Homework #8 is due on 12/11 at 11:59PM

Final Project Deliverable are Due
12/18 at 11:59PM

(See Course Website for Instructions)

Quiz #2 will take place on 12/11 in class. See review checklist on course website.

How do we deploy Bayes' theorem for decision making?

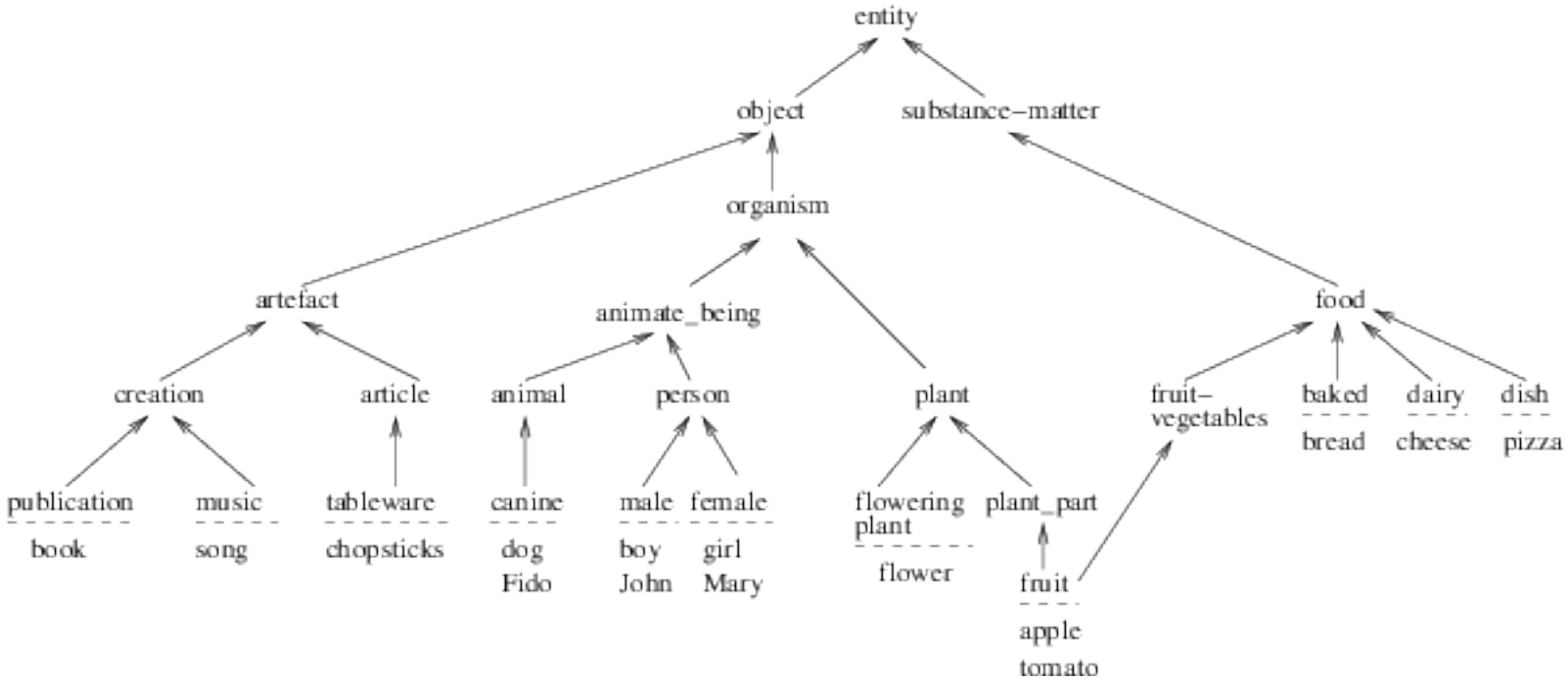
Course Instructor Feedback (CIF)

Deadline: 11:59PM, 12/15/19

Hierarchies in Vision

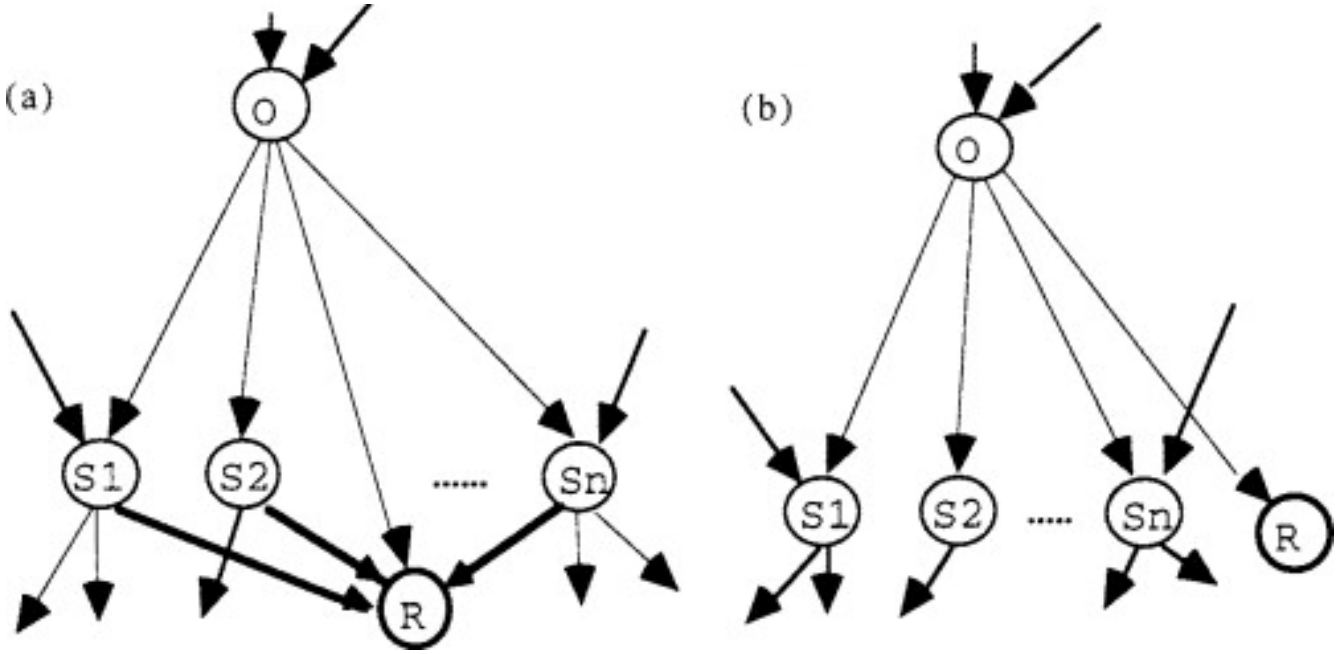
Tenenbaum et al. Science 2011

Hierarchies in Language



Kazakov and Dobnik Topics in Phonetics and Computational Linguistics 2003

Hierarchies in Abstract Knowledge



Roberto 1997

Bayesian Hierarchical Modeling

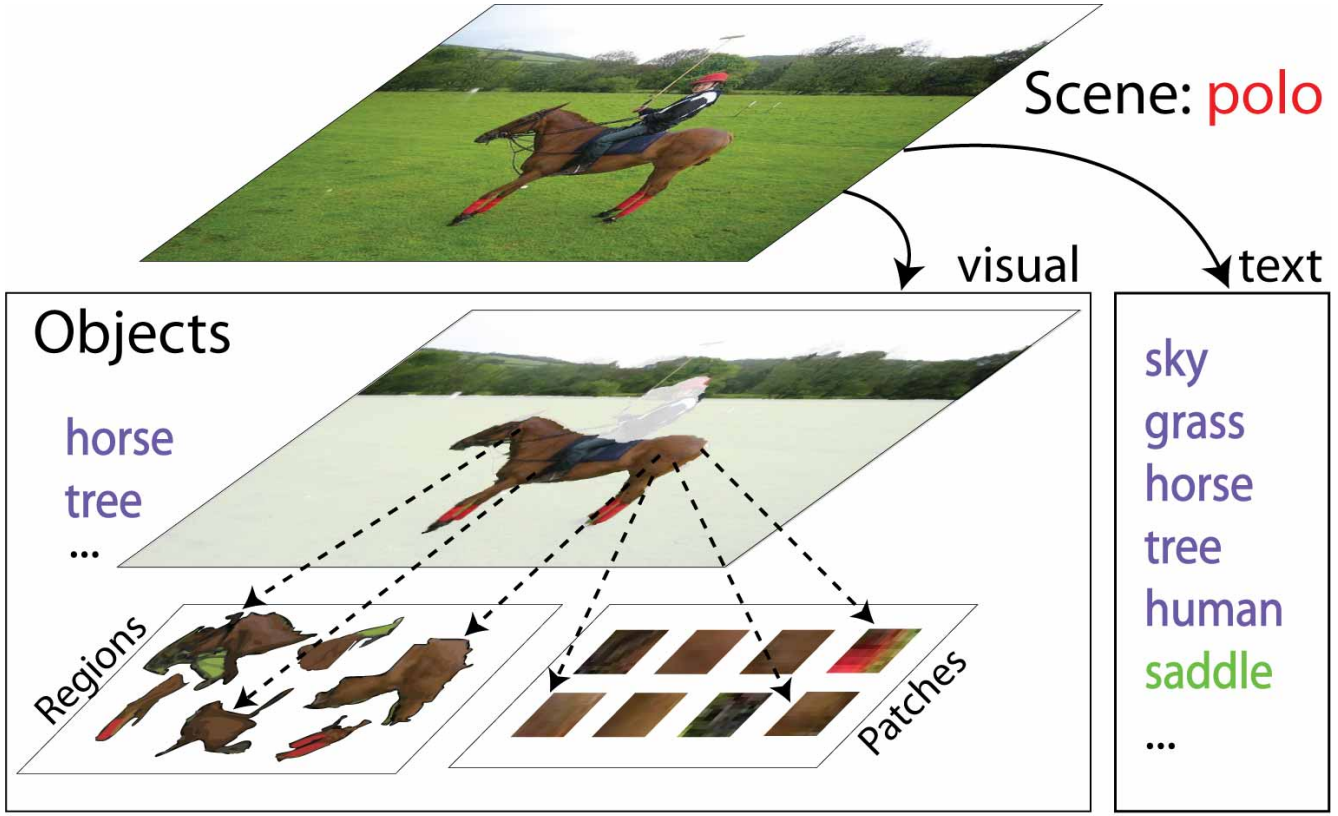
- A statistical model written at multiple levels (i.e., a hierarchy)
- Estimates the parameters of the posterior distribution using Bayesian inference
- Sub-models combine to form the hierarchical model
  - Bayes' theorem is used to integrate them with the observed data (accounting for uncertainty)
  - Result is a probability estimate (the read-out value we want)

Advantages of Bayesian Hierarchical Modeling

- The right approach when information is available at different levels
- Hierarchical form of analysis and organization helps in the understanding of multiparameter problems
- **Important for developing computational strategies**

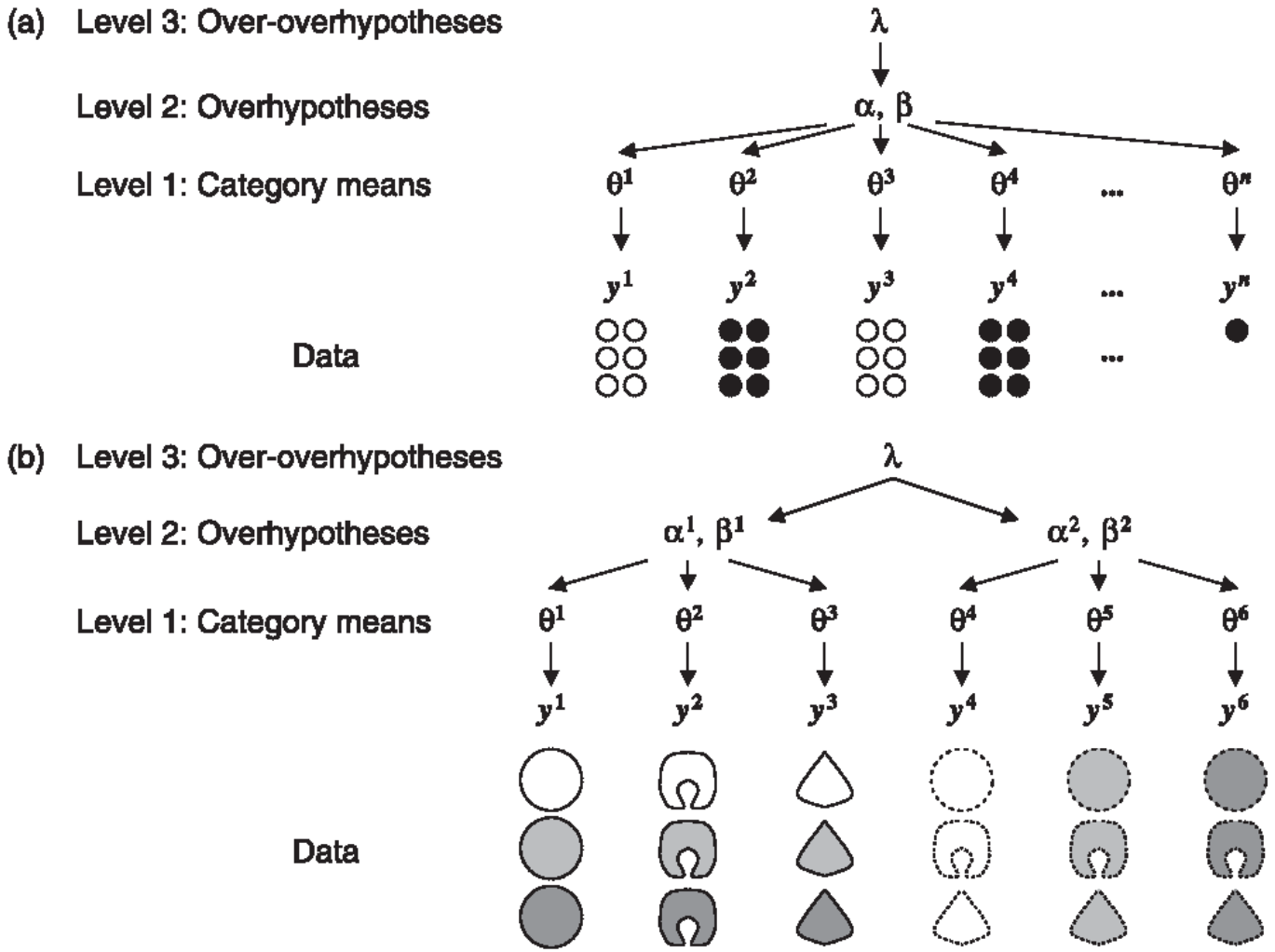
Forms of hierarchical models

Coherent Object Model



Li et al. CVPR 2009

Constraints on the hypotheses considered by the learner



Kemp et al. Developmental Science 2007

Exchangeability

Primer:

Let's assume n values $y_1, y_2, \ldots y_n$ are exchangeable

Let's also assume that θ_j is a parameter vector associated with each y_j

If no information is available to distinguish any of the θ_js, and no order or grouping of the associated parameters can be made, one must assume symmetry among the parameters in their prior distribution

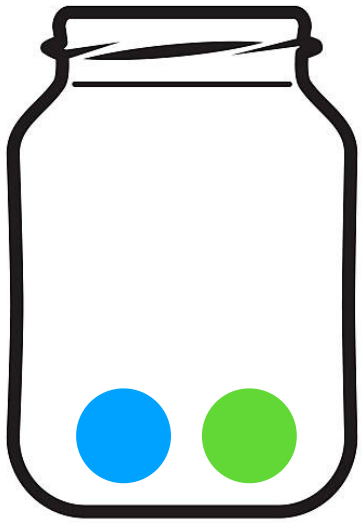
The symmetry is represented probabilistically by **exchangeability**

Finite Exchangeability

We typically model data from an exchangeable distribution as i.i.d., but there is some nuance to this

Example:

How do the probabilities work out if we sample without replacement?



Finite Exchangeability

In the previous example, y_1 and y_2 are exchangeable, but they aren't independent

If $x_1, x_2, \ldots x_n$ are i.i.d., then they are exchangeable, but the converse is not necessarily true

Infinite Exchangeability

Infinite exchangeability is the property that every finite subset of a sequence $y_1, y_2, \ldots$ is exchangeable

For any n, the sequence $y_1, y_2, \ldots y_n$ is exchangeable

Formulating Hierarchical Models

Components

Need two important pieces to derive the posterior distribution:

1. **Hyperparameters:** parameters of the prior distribution
2. **Hyperpriors:** distributions of hyperparameters

Pre-requisites

Suppose a random variable Y follows a normal distribution with θ as the mean and 1 as the variance:

$$Y \mid \theta \sim N(\theta, 1)$$

Suppose θ is normally distributed with mean μ and variance 1:

$$\theta \mid \mu \sim N(\mu, 1)$$

μ is a **hyperparameter**, and its distribution is a standard normal, which is a **hyperprior**

Multiple stages? μ follows another normal distribution with mean β and variance ϵ (these are also hyperparameters with hyperpriors)

Notation

Let y_j be an observation and θ_j a parameter governing the data generating process of y_j.

Assume parameters $\theta_1, \theta_2 \ldots \theta_j$ are generated exchangeably from a common population, with a distribution governed by a hyper parameter ϕ

Bayesian Hierarchical Model Stages

$$\text{Stage I: } y_j \mid \theta_j, \phi \sim P(y_j \mid \theta_j, \phi) \longleftarrow \textbf{likelihood}$$

$$\text{Stage II: } \theta_j \mid \phi \sim P(\theta_j \mid \phi) \longleftarrow \textbf{prior distribution}$$

$$\text{Stage III: } \phi \sim P(\phi) \longleftarrow \textbf{likelihood depends on hyperparameter through } \theta_j$$

Bayesian Hierarchical Model Stages

Prior distribution from Stage I can be broken down into:

$$P(\theta_j, \phi) = P(\theta_j \mid \phi)P(\phi)$$

With ϕ as its hyperparameter with hyperprior distribution $P(\phi)$

Bayesian Hierarchical Model Stages

Posterior distribution is proportional to:

$$P(\phi, \theta_j \mid y) \propto P(y_j \mid \theta_j, \phi) P(\theta_j, \phi)$$ ← **using Bayes' theorem**

$$P(\phi, \theta_j \mid y) \propto P(y_j \mid \theta_j) P(\theta_j \mid \phi) P(\phi)$$

Example

A teacher wants to estimate how well a female student did on the SAT. He uses information on the student's high school grades and her current GPA to make an estimate

Y = current GPA

θ = SAT score

Y has a likelihood given by some probability function with θ:

$$Y \mid \theta \sim P(Y \mid \theta)$$

Example

The SAT score is a sample coming from a common population distribution indexed by another parameter ϕ, which is the high school grade of the student:

$$\theta \mid \phi \sim P(\theta \mid \phi)$$

The hyperparameter ϕ follows its own distribution given by $P(\phi)$, a hyperprior

To solve for the SAT score given information on the GPA:

$$P(\theta, \phi \mid Y) \propto P(Y \mid \theta, \phi) P(\theta, \phi)$$
$$P(\theta, \phi \mid Y) \propto P(Y \mid \theta) P(\theta \mid \phi) P(\phi)$$

2-stage Hierarchical Model

The joint posterior distribution of interest in 2-stage hierarchical models is:

$$P(\theta, \phi \mid Y) = \frac{P(Y \mid \theta, \phi)P(\theta, \phi)}{P(Y)} = \frac{P(Y \mid \theta)P(\theta \mid \phi)P(\phi)}{P(Y)}$$

$$P(\theta, \phi \mid Y) \propto P(Y \mid \theta)P(\theta \mid \phi)P(\phi)$$

3-stage Hierarchical Model

For 3-stage hierarchical models, the posterior distribution is given by:

$$P(\theta, \phi, X \mid Y) = \frac{P(Y \mid \theta)P(\theta \mid \phi)P(\phi \mid X)P(X)}{P(Y)}$$

$$P(\theta, \phi, X \mid Y) \propto P(Y \mid \theta)P(\theta \mid \phi)P(\phi \mid X)P(X)$$

Software: PyMC3

https://docs.pymc.io