Cognitive Modeling

Lecture 12: Bayesian Inference

Sharon Goldwater

School of Informatics
University of Edinburgh
sgwater@inf.ed.ac.uk

February 18, 2010

Reading: Griffiths and Yuille (2006).

Bayesian Models of Cognition

Much of cognition can be viewed as *prediction* based on data.

- decision-making
- categorization
- causal inference
- word learning
- language processing

Probability theory provides techniques for making *optimal predictions*, so rational analysis approach suggests we use them.

Intuitions

Last class we developed some intuitions about Bayesian inference.

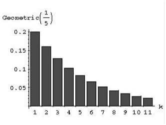
- Probabilities reflect degrees of belief.
- In real situations, probabilities are unknown and must be estimated (inferred).
- Estimates depend both on prior beliefs and on observations.
- As more observations accrue, estimates converge to relative frequencies.

Today we will discuss some of the mathematics.

Distributions

So far, we have discussed *discrete distributions*.

- Sample space S is finite or countably infinite (integers).
- Distribution is a *probability mass function*, defines probability of RV taking on a particular value.
- Ex: $P(X = x) = (1 - p)^{x-1}p$ (Geometric distribution):

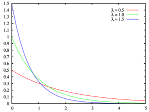


(Image from http://eom.springer.de/G/g044230.htm)

Distributions

Today we will also see *continuous distributions*.

- Sample space is uncountably infinite (real numbers).
- Distribution is a *probability density function*, defines relative probabilities of different values (sort of).
- Ex: $p(x) = \lambda e^{-\lambda x}$ (Exponential distribution):



(Image from Wikipedia)

Discrete vs. Continuous

Discrete distributions:

- $0 \leq P(X = x) \leq 1$ for all $x \in S$
- $\sum_{x \in S} P(x) = 1$.
- $P(Y) = \sum_{X_i} P(Y|X_i)P(X_i)$ (Law of Total Prob.)
- $E[X] = \sum_x x \cdot P(X = x)$ (Expectation)

Continuous distributions:

- $p(x) \geq 0$ for all x
- $\int_{-\infty}^{\infty} p(x) = 1$.
- $p(y) = \int p(y|x)p(x)dx$ (Law of Total Prob.)
- $E[X] = \int_x x \cdot p(x)dx$ (Expectation)

Prediction

Simple inference task: estimate the probability that a particular coin shows heads. Let

- θ: the probability we are estimating.
- H: hypothesis space (values of θ between 0 and 1).
- D: observed data (previous coin flips).
- n_h, n_t: number of heads and tails in D.

Bayes' Rule tells us:

$$p(\theta|D) = \frac{P(D|\theta)p(\theta)}{p(D)} \propto P(D|\theta)p(\theta)$$

How can we use this?

Maximum-likelihood Estimation

1. Choose θ that makes D most probable, i.e., ignore $p(\theta)$:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, P(D|\theta)$$

This is the *maximum-likelihood* (ML) estimate of θ, and turns out to be equivalent to relative frequencies:

$$\hat{\theta} = \frac{n_h}{n_h + n_t}$$

- Insensitive to sample size, and does not generalize well (overfits).

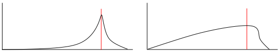
Maximum a Posteriori Estimation

2. Choose θ that is most probable given D:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, P(\theta|D) = \underset{\theta}{\operatorname{argmax}}\, P(D|\theta)p(\theta)$$

This is the *maximum a posteriori* (MAP) estimate of θ, and is equivalent to ML when $p(\theta)$ is uniform.

- Non-uniform priors can reduce overfitting, but MAP still doesn't account for the shape of $p(\theta|D)$:



Bayesian integration

3. Take the expected value of θ instead of maximizing:

$$E[\theta] = \int \theta \frac{P(D|\theta)p(\theta)}{p(D)} d\theta \propto \int \theta P(D|\theta)p(\theta)d\theta$$

This is the *posterior mean*, an average over hypotheses. When prior is uniform, we have

$$E[\theta] = \frac{n_h + 1}{n_h + n_t + 2}$$

- Automatic smoothing effect: unseen events have non-zero probability.
- Non-uniform prior favoring $\theta = .5$ adds more "pseudo-counts", requires more observations to overcome.

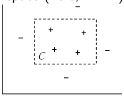
Concept Learning

Tenenbaum (1999) addresses the question of how people quickly learn new concepts.

- Concepts could be categories (dog, chair) or more vague ("healthy level" for a specific hormone, "ripe" for a pear).
- Generalization: given a small number of positive examples, which other examples are also members of the concept?
- In machine learning, often called *classification*.

Formalization

Assume that concept C can be represented as a rectangle in n-dimensional space (here, $n = 2$):



(Figure from Tenenbaum (1999))

- Dimensions could be levels of cholesterol, insulin; concept is "healthy levels".
- Learner does not know the boundaries of the concept rectangle.
- Given examples $X = \{x_1 \ldots x_n\}$ with $x_i \in C$, predict $p(y \in C|X)$ for new example y.

Related Work

Most classification methods/models are *discriminative*:

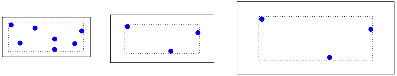
- Require both positive and negative examples.
- Usually require large numbers of examples.
- Ex: neural networks, decision trees, support vector machines.

Simple early model: MIN (Bruner et al., 1956).

- Works with positive examples only.
- Assumes smallest possible category that contains all observed examples.

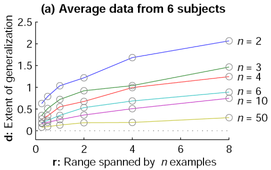
Human Data

- Subjects generalize further when fewer examples are available.
- Subjects generalize further when examples span a larger range.



Human Data

Findings from Tenenbaum (1999):



Bayesian Model

- Goal: Given examples $X = \{x_1 \dots x_n\}$ with $x_i \in C$, predict $p(y \in C|X)$ for new example y.
- C is a rectangle, so hyp. space H is all possible rectangles.
- Make prediction by summing over hypotheses:

$$\begin{aligned} p(y \in C|X) &= \int p(y \in C|h, X)p(h|X)dh && \text{Tot. Prob.} \\ &= \int p(y \in C|h)p(h|X)dh && \text{Cond. Indep.} \\ &\propto \int p(y \in C|h)p(X|h)p(h)dh && \text{Bayes' Rule} \end{aligned}$$

Likelihood

Assume X are sampled uniformly at random from C. Then

$$P(X|h) = \begin{cases} \frac{1}{|h|^n} & \text{if } \forall j, x_j \in h \\ 0 & \textit{otherwise} \end{cases}$$

- Smaller hypotheses have higher likelihood (*size principle*).
- Maximum likelihood chooses smallest h consistent with X: equivalent to MIN.

Prior

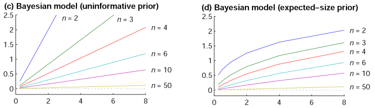
Tenenbaum (1999) considers two different priors.

- *Uninformative* prior: all rectangles are equally probable.
- Prior based on expected size of rectangles.

Since stimuli are presented on a computer screen, expected size makes sense: rectangles are presumably not larger than screen.

Model Results

Results from Tenenbaum (1999):



Discussion

- Model predicts behavior of concept learning from positive examples.
- Captures effects of number of examples and range of examples.
- Best fit uses expected-size prior.
- Suggests that humans make optimal Bayesian predictions.
- Says nothing about mechanisms that might implement inference in the mind.

Summary

- Many cognitive tasks involve prediction.
- Bayesian techniques for making optimal predictions: use of priors, hypothesis averaging.
- Permits generalization to unseen examples.
- Predicts human behavior in concept learning task.

References

Griffiths, Tom L. and Alan Yuille. 2006. A primer on probabilistic inference. *Trends in Cognitive Sciences* 10(7).

Tenenbaum, J. 1999. Bayesian modeling of human concept learning. In M. Kearns, S. Solla, and D. Cohn, editors, *Advances in Neural Information Processing Systems 11*. MIT press, Cambridge.