Learning Probability Distributions

CPS 271
Ron Parr

The Problem

- Observe a sequence of events
- Predict the probability of future events based upon observations
- Classical statistical problem
- Surprisingly subtle issues arise

Event Spaces

- We first consider learning with a small event space
- Recall: An event space is the space of all possible outcomes
- What is the event space for coin tosses?

Classical Approaches

- In classical (frequentist) statistics, we count the number of outcomes of each type and divide by the total number of events
- 642 heads, 375 tails
- \( P(\text{Heads})=0.6312 \)
- What do we know about this approach?

Bayesian Approaches

- Bayesians hold that probabilities come not just from frequencies, but from our beliefs
- Suppose \( P(\text{Heads})=X \), why not use Bayes rule to figure out \( P(X) \) given observations \( O \)?

\[
P(X \mid O) = \frac{P(O \mid X)P(X)}{P(O)}
\]

Issues

- \( X \) is now continuous
- What is \( P(X) \)?
- What is \( P(O) \)?
- What is the form of \( P(X \mid O) \)?
Priors

- Choosing a prior distribution over model parameters is a subtle question
- What is a plausible prior for X?
- Priors must also be mathematically convenient: P(X|O) must be something we can work with
- Mathematical convenience and plausibility are often competing objectives

Multinomial Distributions

- Binomial Distribution
  - Two random events, e.g., coin flip
- Multinomial
  - Multiple random events
  - N-sided die
  - Multi-valued nodes in a Bayes net
  - Next state in a Markov process

Dirichlet Distribution

Define a probability density function over n parameters:

$$f(x|\alpha) = \frac{\Gamma(\alpha_1 + \ldots + \alpha_n)}{\Gamma(\alpha_1) \ldots \Gamma(\alpha_n)} x_1^{\alpha_1-1} \ldots x_n^{\alpha_n-1}$$

Where:

$$\Gamma(\alpha) = \int_0^\infty u^{\alpha-1} e^{-u} du$$

Dirichlet for 2 variables

$$P(\text{Heads}) = x$$

$$p(x|\alpha_1 = 1, \alpha_2 = 1)$$

Dirichlet for 2 variables

$$p(x|\alpha_1 = 2, \alpha_2 = 2)$$

Dirichlet for 2 variables

$$p(x|\alpha_1 = 10, \alpha_2 = 10)$$
Dirichlet for 2 variables

\[ p(x|\alpha_1 = 5, \alpha_2 = 10) \]

Dirichlet: General

\[ \alpha_0 = \sum_\alpha \alpha_i \]

\[ E(X_i) = \frac{\alpha_i}{\alpha_0} \]

\[ Var(X_i) = \frac{\alpha_i(\alpha_0 - \alpha_i)}{\alpha_0(\alpha_0 + 1)} \]

- Magnitudes of the alphas indicate the “strength” of the average
- Small alphas indicate “weak” preference for mean

Dirichlet: General

- Suppose we have a dirichlet(1,3) prior for \( P(\text{Heads}) \)
- The mean of our prior on \( P(\text{Heads}) \) is 1/4
- In general, the mean of the dirichlet prior gives a distribution over the parameters equivalent to what we would get if we had seen the corresponding number of random events and done classical statistics
- It gets better…

Dirichlet as Conjugate Prior

Suppose we have \( P(X) \) dirichlet, and \( P(O|X) \) multinomial in \( X \). (e.g. \( X=P(\text{Heads}) \), \( P(X) \) is dirichlet(2,2), \( O=\text{HTT} \))

\[ P(X | O) = \frac{P(O | X)P(X)}{P(O)} \]

It’s a (complicated) theorem that \( P(X|O) \) is also dirichlet. Moreover, if \( P(X) \) is dirichlet(\( \alpha \)) and we observe random Events with counts(\( \beta \)), then \( P(X|O) \) is dirichlet(\( \alpha + \beta \))

Dirichlet Example

- Suppose our prior on the coin is dirichlet(2,2) and we have observed one head and two tails.
- Our posterior is dirichlet(3,4)
- The mean for \( P(\text{Heads})=3/7 \)
- Classical methods would yield 1/3

Why This is Nice

- For multinomials, we can be Bayesian without learning lots of extra math
- Our prior corresponds to a certain number of “phantom” examples that we pretend to have seen in the past
- We can express the same mean with different “strengths” by increasing the number of phantom examples, e.g., (10,10) vs. (10000, 10000)
- For large amounts of data, our phantom transitions will be overwhelmed and not matter
- Gives us a way of introducing prior knowledge and getting (possibly) more robust results for small data sets
Learning Bayes Nets

- So far, we have gone back to the atomic event view of the world
- In Bayes nets, we have random variables and structure
- Let’s assume that we are told the structure
- How do we learn the parameters?

Learning BN Parameters

- For every node, every CPT, we maintain a set of counts
- Iterate over the data and collect the counts for every combination of parents, children we see
- e.g. For every combination of Flu, Allergy in the data, we count the number of times we have Sinus=t and Sinus=f, etc.
- Why is this sufficient?

Priors in Bayes Nets

- Each node in a Bayes net is like a multi-sided die
- We can do the Dirichlet trick for Bayes nets too!
- How this can work in practice:
  - Goal: Adaptive medical diagnosis system
  - Doctors initialize system with dirichlet priors
    - From doctors’ estimates of probabilities, -or-
    - Statistics from a different area
  - System adds new data as it is used, adjusts to characteristics of new area

Missing Data

- Suppose that for some data, we observe flu, allergy, headache and nose, but not sinus
- Can we still learn, or should we throw data from these patients anyway?
- Idea: Why not guess the missing values based on the other patients?

EM Algorithm

- Expectation maximization
  - Guess probabilities for nodes with missing values (e.g. based on other patients)
  - Compute the probability distribution over the missing values given our guess
  - Update the probabilities based upon the filled in values
  - Repeat until convergence

EM Example

- Suppose we have not observed Sinus for patient 1
- We estimate the CPTs based upon the other patients
- We then estimate P(Sinus) for patient 1
- Now we recompute our CPTs as if the guessed values had been observed
- Repeat
Advanced Topics

- EM can also be used to learn structure and infer the presence of hidden variables (these are simply additional unobserved parameters)
- EM is a general algorithm that applies to estimating hidden parameters in many types of models
  - HMMs
  - Mixtures of Gaussians, etc.
- Structure learning typically requires regularization