Parameter Estimation: Gradient Descent and Expectation Maximization

Saurabh Sinha (sinhas@uiuc.edu)

January 29, 2008

http://www.cs.uiuc.edu/homes/sinhas/

Department of Computer Science
University of Illinois Urbana-Champaign

Prepared By: Charles Blatti

Outline of Bayesian Analysis

- Construct a parameterized model, $M(\omega)$
- Decide how to calculate the joint $Pr(\omega, D)$ and the posterior $Pr(\omega|D)$
- Find the optimal setting of the parameters (ω) by maximizing $Pr(\omega|D)$
  - This is the MAP (Maximum a posteriori) estimate
- Assign a score $f(\omega, D)$ to the data based on the model
  - Example 1: gene prediction where every genetic window must have a score
  - Example 2: coin with scoring function $f(p) = (.5 - p)^2$
  - Note: if a particular set of values for ω is not a reliable answer than use

$$\sum_{\omega} f(\omega) Pr(\omega|D)$$

    rather than $f(\omega, D)$

$Pr(\omega|D)$ Optimization Algorithms

- Gradient Descent
- Expectation Maximization (E-M)
- Sampling or MCMC

Gradient Descent: Overview

- Gradient Descent is a general purpose algorithm that is widely used
- The method is useful if $f(\omega)$ is differentiable
- Gradient Descent terminates at a local optimum
- While the algorithm is easy to code, it can be slow to run
- There are variants available to speed the process up (conjugate gradients)

Gradient Descent: Procedure

- Gradient Descent is useful to minimize:
  - for the MAP (Maximum a posteriori) estimate, $-\log Pr(\omega|D)$
  - for the M-L (Maximum likelihood) estimate, $-\log Pr(D|\omega)$
- Gradient Descent takes steps in the direction of the slope of f (the derivative) in order to move towards a local minimum
- The algorithm's update procedure is

$$\omega^{t+1} = \omega^t - \eta \frac{\delta f}{\delta \omega^t}$$

- where η is the step size and $\frac{\delta f}{\delta \omega^t} = \frac{\delta f(\omega)}{\delta \omega}$ evaluated at $\omega = t$

E-M with Hidden Variables

- Expectation Maximization is useful in models with hidden variables

- If D is the data and H is the hidden variables and there is a known procedure to calculate $Pr(D, H|\omega)$, then

$$Pr(D|\omega) = \sum_{H} Pr(D, H|\omega)$$

- In general, it is difficult to optimize $Pr(D|\omega)$ when there are hidden variables

E-M Overview

- The idea of Expectation Maximization is to use the current estimate of the parameters (ω^t) to impose a distribution,

  $$\sigma^*(H) = Pr(H|D, \omega^t)$$

  and optimize

  $$E_{\sigma^*}[\log Pr(D, H|\omega)]$$

- In other words, E-M seeks to find the

  $$\omega^{t+1} = argmax \sum_H Pr(H|D, \omega^t) \log Pr(D, H|\omega)$$

- Maximizing $E_{\sigma^*}[\log Pr(D, H|\omega)]$ in many cases may be easier because it turns products into sums and allows the use of the linearity of expectation

Hidden Markov Models (HMM)

- An Hidden Markov Model is abstract construction that contains k hidden states that emit symbols
- The states are called hidden variables because it is unknown which state emitted the observed symbols
- An HMM is similar to a Markov Chain because there are transition probabilities, where each state has its own probability distribution over the states that will follow it
- HMMs also have emission probabilities, distributions defined for each state over the alphabet of symbols that are emitted

Hidden Markov Models: Example

- An example of a HMM would be two four-sided dice where you see the result of a roll, but not the die that was thrown
- The hidden states are the identity of the die, either uniform (U) or G/C rich (G)
- $\Sigma = \{A, C, G, T\}$ is the set of emission symbols
- $A = \{a_{kl}\}$ is the set of probabilities of changing from state k to state l for all k and l. For this example:

$$a_{UU} = p_U, a_{UG} = 1 - p_U, a_{GG} = p_G, a_{GU} = 1 - p_G$$

- $E = \{e_k(b)\}$: the set of probabilities of emitting symbol b while in state k for all k and b. For this example:

$$e_U(A) = 1/4, e_U(C) = 1/4, e_U(G) = 1/4, e_U(T) = 1/4$$

$$e_G(A) = 1/6, e_G(C) = 1/3, e_G(G) = 1/3, e_G(T) = 1/6$$

HMMs: Calculations

- The sequence of rolls is $D \in \{\Sigma\}^L = D_1 D_2 ... D_L$
- The sequence of hidden states (a parse) is $H \in \{U, G\}^L = H_1 H_2 ... H_L$
- Remember

$$Pr(D|\omega) = \sum_H Pr(D, H|\omega)$$

- In terms of of HMM model

$$Pr(D|\omega) = \sum_H \prod_{i=1}^{L} Pr(D_i|H_i) Pr(H_i|H_{i-1})$$

the product of the emission probability and the transition probabilities over the entire sequence

E-M Justification: Calculations I

- The idea of the Expectation Maximization method is to iteratively improve ω such that

$$\log Pr(D|\omega^{t+1}) \geq \log Pr(D|\omega^{t})$$

- The justification of this result is derived from the formula

$$\log Pr(D|\omega) = \log Pr(D, H|\omega) - \log Pr(H|D, \omega)$$

- Multiply both sides by $Pr(H|D, \omega^t)$ and sum over H than:
  - LHS $= \sum_H Pr(H|D, \omega^t) \log Pr(D|\omega)$
  - RHS $= \sum_H Pr(H|D, \omega^t) \log Pr(D, H|\omega) - \sum_H Pr(H|D, \omega^t) \log Pr(H|D, \omega)$

- Define the first term of the RHS as

$$Q(\omega, \omega^t) = \sum_H Pr(H|D, \omega^t) \log Pr(D, H|\omega)$$

E-M Justification: Calculations II

- From the previous slide

$$\log Pr(D|\omega) = Q(\omega, \omega^t) - \sum_H Pr(H|D, \omega^t) \log Pr(H|D, \omega)$$

- The following equation can be written using the above substitution

$$\log Pr(D|\omega) - \log Pr(D|\omega^t)$$

$$= Q(\omega, \omega^t) - \sum_H Pr(H|D, \omega^t) \log Pr(H|D, \omega)$$

$$-Q(\omega^t, \omega^t) + \sum_H Pr(H|D, \omega^t) \log Pr(H|D, \omega^t)$$

$$= Q(\omega, \omega^t) - Q(\omega^t, \omega^t) + \sum_H Pr(H|D, \omega^t) \frac{\log Pr(H|D, \omega^t)}{Pr(H|D, \omega)}$$

- Notice that the sum in the last equation is just the relative entropy between $Pr(H|D, \omega^t)$ and $Pr(H|D, \omega)$. Because of the property of relative entropy, this term must be non-negative.

E-M Justification: Calculations III

- Non-negative relative entropy means that

$$\log Pr(D|\omega) - \log Pr(D|\omega^t) \geq Q(\omega, \omega^t) - Q(\omega^t, \omega^t)$$

- Let $\omega^{t+1} = argmax Q(\omega, \omega^t)$
- Since ω^{t+1} is optimal,

$$Q(\omega^{t+1}, \omega^t) - Q(\omega^t, \omega^t) \geq 0$$

- Which implies that

$$\log Pr(D|\omega^{t+1}) \geq \log Pr(D|\omega^t)$$

- Remember that

$$Q(\omega, \omega^t) = \sum_H Pr(H|D, \omega^t) \log Pr(D, H|\omega) = E_{\sigma^*}[\log Pr(D, H|\omega)]$$

- Thus, through the maximization that E-M prescribes above, there is a mathematical guarantee that ω^{t+1} will be better than ω^t

E-M Applied to Example HMM

- In the example HMM,

$$Pr(D, H|\omega) = \prod_{i=1}^{L} Pr(H_i|H_{i-1})Pr(D_i|H_i)$$

- Let $A_{k,l}(H)$ be the number of transitions from state k to l in H. Also, let $E_k(b, H)$ be the number of emissions of the b character from a state k in H

- Rewrite the above with the following transformations

$$Pr(D, H|\omega) = (\prod_{k\in\{U,G\}} \prod_{l\in\{U,G\}} a_{k,l}^{A_{k,l}(H)})(\prod_{k\in\{U,G\}} \prod_{b\in\{\Sigma\}} e_k(b)^{E_k(b,H)})$$

$$\log Pr(D, H|\omega) = \sum_{k\in\{U,G\}} \sum_{l\in\{U,G\}} A_{k,l}(H) \log a_{k,l} + \sum_{k\in\{U,G\}} \sum_{b\in\{\Sigma\}} E_k(b, H) \log e_k(b)$$

$$Q(\omega, \omega^t) = E_{\sigma^*}[\log Pr(D, H|\omega)]$$

$$= E_{\sigma^*}[\sum_{k\in\{U,G\}} \sum_{l\in\{U,G\}} A_{k,l}(H) \log a_{k,l} + \sum_{k\in\{U,G\}} \sum_{b\in\{\Sigma\}} E_k(b, H) \log e_k(b)]$$

E-M Applied to Example HMM

- Using the Linearity of Expectation on the previous formula,

$$= E[\sum_{k \in \{U,G\}} \sum_{l \in \{U,G\}} A_{k,l}(H) \log a_{k,l}] + E[\sum_{k \in \{U,G\}} \sum_{b \in \{\Sigma\}} E_k(b, H)]$$

$$= \sum_{k \in \{U,G\}} \sum_{l \in \{U,G\}} E[A_{k,l}(H)] \log a_{k,l} + \sum_{k \in \{U,G\}} \sum_{b \in \{\Sigma\}} E[E_k(b, H)] \log e_k(b)$$

- Since we maximize $Q(\omega, \omega^t)$ over ω where the parameters are $\omega = a_{k,l}$,
- The term $\sum_{k \in \{U,G\}} \sum_{b \in \{\Sigma\}} E[E_k(b, H)] \log e_k(b)$ is a constant
- So we must only maximize

$$\sum_{k \in \{U,G\}} \sum_{l \in \{U,G\}} E[A_{k,l}(H)] \log a_{k,l}$$

E-M Applied to Example HMM

- $E[A_{k,l}(H)]$ can be treated as a constant vector when maximizing the parameters in the example HMM $\sum_{k\in\{U,G\}} \sum_{l\in\{U,G\}} E[A_{k,l}(H)] \log a_{k,l}$

- Let $\bar{A}_{k,l} = E[A_{k,l}(H)]$ be the mean number of transitions over all possible H. In our HMM setting, we can calculate this $\bar{A}_{k,l}$ with dynamic programming.

- In summary, we are maximizing

$$\sum_{k\in\{U,G\}} \sum_{l\in\{U,G\}} \bar{A}_{k,l} \log a_{k,l}$$

  such that $\sum_l a_{k,l} = 1$ for all k

- With the appropriate derivatives and using LaGrange Multipliers, we will calculate that the maximum ω is the one with

$$a_{k,l} = \frac{\bar{A}_{k,l}}{\sum_{l'} A_{k,l'}}$$

- This is the maximization result, ω^{t+1}, in Expectation Maximization.

E-M Summary

- The E-M approach guarantees to iteratively improve your objective function ω at each update because

$$\log Pr(D|\omega^{t+1}) \geq \log Pr(D|\omega^{t})$$

by maximizing

$$Q(\omega, \omega^{t}) = E_{\sigma^{*}}[\log Pr(D, H|\omega)]$$

where the expectation is over the probability distribution

$$\sigma^{*}(H) = Pr(H|D, \omega^{t})$$