Tools & Models for Data Science

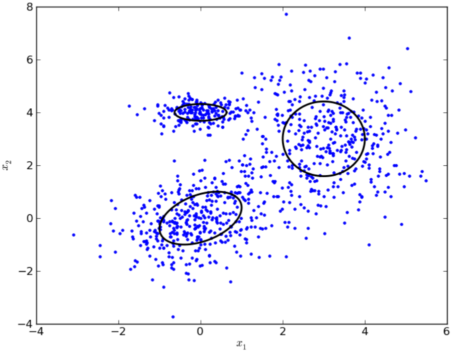
Mixture Models

Chris Jermaine & Risa Myers

Rice University

Mixture Model Introduction

- At highest level:
  - Have a set of data
  - And a set of random variables
  - Don't know which one produced which point
- This is a mixture model!
- In one line:
  - MM: "hierarchical," stochastic, latent variable model

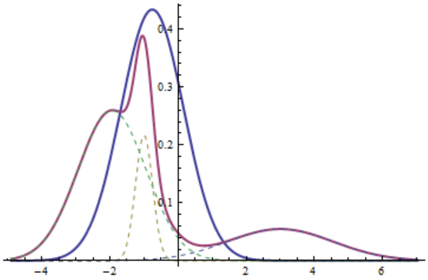


Why Use Them?

- Sometimes, we want to segment the data
  - Observe a set of test scores
  - Want 3 types of students: good, average, bad
  - Associate each with a different Normal distribution
- Sometimes, we just want a very flexible model
  - A mixture can give a complicated, multi-modal distribution
  - We can construct these from simple distributions
  - In reality, not much data is purely normally distributed

GMM Example

- Blue curve is what you get if you fit a single normal distribution to the purple data
- Note the shift in the mean
- The dashed curves are the mixture distributions
- Note there are actually 3 normal distributions in this plot



Mixture Model Introduction

- Choose the parameters for each normal distribution
- And a weighting for each mixture
- Use these to produce a set of data $x_1, x_2, \ldots, x_n$
  - And a set of hidden (latent) indicators $c_1, c_2, \ldots, c_n$ specifying which mixture was selected
- MM begins with a distribution function f
  - Common f: Gaussian, Multinomial, Gamma, etc.
  - We have k sets of parameters for f: $\theta_1, \theta_2, \ldots, \theta_k$
  - And a probability vector π that tells us how important each mixture component is
  - Note: we can have a mixture model where each component is from a different distribution
  - This is uncommon, though

Mixture Model Introduction

- Pseudo-code to produce n observations is:

for i = 1 to n **do:**

$c_i \sim \text{Categorical}(\pi)$

$x_i \sim f(\theta_{c_i})$

Mixture Model PDF

- In general, PDF is:

$$P(x_1, x_2, \ldots, x_n) = \prod_{i}^{n} \left(\sum_{j}^{K} \pi_j f(x_i | \theta_j) \right)$$

- Where n is the number of data points
- Where K is the number of mixtures
- Where f is the PDF for distribution k

? Why?

Mixture Model PDF

- In general, PDF is:

$$P(x_1, x_2, \ldots, x_n) = \prod_i^n \left(\sum_j^K \pi_j f(x_i|\theta_j) \right)$$

- Why?
  - Likelihood of choosing component j, then j producing x is:

$$\pi_j f(x|\theta_j)$$

  - Since n data points are independent, we have a product over the likelihoods

Mixture Model PDF

- For each data point, the PDF is:

$$\sum_{j}^{k} \pi_j f(x_i|\theta_j)$$

- We can represent the choice, c as a 1-of-k valued vector, where one dimension is 1 and the rest are all 0
- The dimension is set to 1 if that mixture model generated the data point
- and set to 0 otherwise
- It follows that $p(c_j = 1) = \pi_j$
- and that $\sum_j^k \pi_j = 1$
- To get the marginal distribution of x we sum the joint distribution over all possible states of c

Learning

- Would be easy if we knew $c_1, c_2, \ldots, c_n$
- Where c_j is the indicator of which mixture distribution produced data point j
  - In that case, we'd have k different MLE problems
  - That is, to learn, partition the data points according to c values
  - Then perform an MLE separately on each
  - For standard distributions, this is easy
  - Example, for 1-D Normal
    - MLE for μ_j is mean of points with $c_i = j$
    - MLE for σ_j is std. dev. of points with $c_i = j$
- But there are complications in doing this
- And we don't know the c_j values anyway!

Learning

- But we don't!
- Standard methods:
    - MCMC (sample $c_1, c_2, ..., c_n$): Stochastic, Bayesian approach, not covered in this course
    - EM

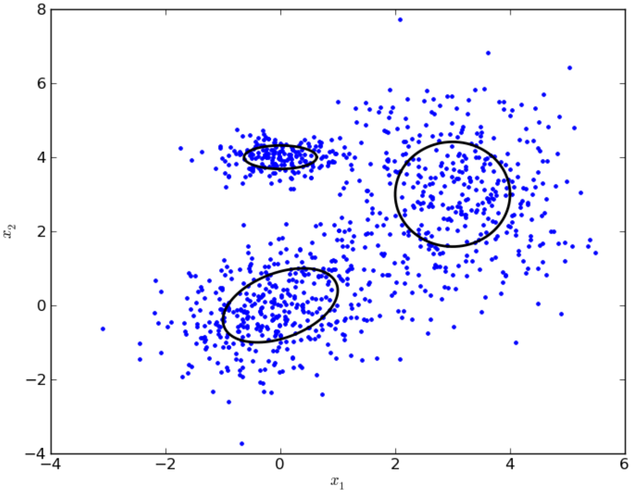
EM for Mixture Models...

- The Patriarch: The Gaussian Mixture Model (GMM)

$$P(x_1, x_2, \ldots, x_n) = \prod_i \left(\sum_j \pi_j \mathrm{Normal}(x_i | \mu_j, \sigma_j^2) \right)$$

1. Make non-stupid guesses for the means, variances, and the mixing parameters
2. E-step: Use the current parameter values to asign each data point to the most likely mixture component
3. M-step: Use the expected values to update the mixture parameters

Example GMM



Mixture of Dirichlets

- Imagine that I have a large corpus of text documents
- And I want to understand the various types of documents present
- Basically, I want to cluster the documents
- We might view the TF (term frequency) vector as being produced as a sample from a Dirichlet distribution

? What's a Dirichlet distribution?

Dirichlet Distribution

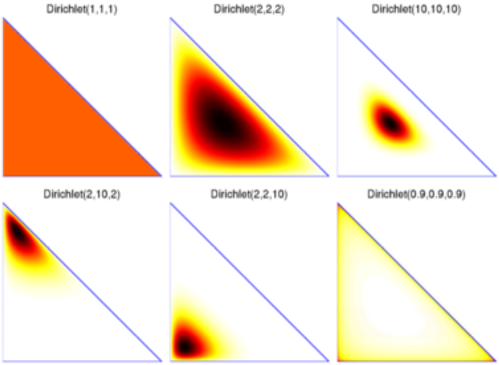
- Generalization of Beta to many dimensions
- Produces vectors on the “simplex”
  - That is, vectors whose entries sum to 1
  - Hence, used as a distribution to produce vectors of probabilities
- Takes params $\langle v_1, v_2, ..., v_m \rangle$
  - Produces m values that sum to 1
  - “looks” like a probability vector
  - To generate, sample from m Gammas, each with shape v_j
  - ith value is ith Gamma’s fractional contribution to total

Dirichlet Distribution

- Expected value of ith is $\frac{v_j}{\sum_{j'} v_{j'}}$
  - Then what's the difference between parameter vector $\langle 1, 2, 3 \rangle$ and $\langle 10, 20, 30 \rangle$?
- Both result in Dirichlets with same mean
  - But smaller parameters allow more variance
  - So the former has much more variance... resulting Dirichlet could produce $\langle 0.2, 0.1, 0.7 \rangle$
  - Latter will always produce something like $\langle 0.1666, 0.3333, 0.5 \rangle$
  - If we normalize the parameter vectors (divide by their sum) we get $6 \times \langle 0.166, 0.333, 0.5 \rangle$ and $60 \times \langle 0.166, 0.333, 0.5 \rangle$
  - The higher the normalization constant is, the closer the samples will be to the mean

Dirichlet Distribution

1. $\langle 1, 1, 1 \rangle$ Everything is equally likely
2. $\langle 2, 2, 2 \rangle$ As the parameters grow, on expectation, the samples are more concentrated
3. $\langle 10, 10, 10 \rangle$ As the numbers continue to get bigger, the samples are more likely to be $\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$
4. $\langle 2, 10, 2 \rangle$ If a parameter is larger, that dimension get more weight
5. $\langle 2, 2, 10 \rangle$
6. $\langle 0.9, 0.9, 0.9 \rangle$ When parameters are < 1, the samples go to the outside (more extreme differences)



[a]

[a] www.cs.cmu.edu/~epxing/Class/10701-08s/recitation/dirichlet.pdf

Mixture of Dirichlets Application

- Imagine that I have a large corpus of text documents
- And I want to understand the various types of documents present
- Basically, I want to cluster the documents
- We might view the TF (term frequency) vector as being produced as a sample from a Dirichlet distribution
- TF vector is basically a vector of probabilities

Mixture of Dirichlets Example

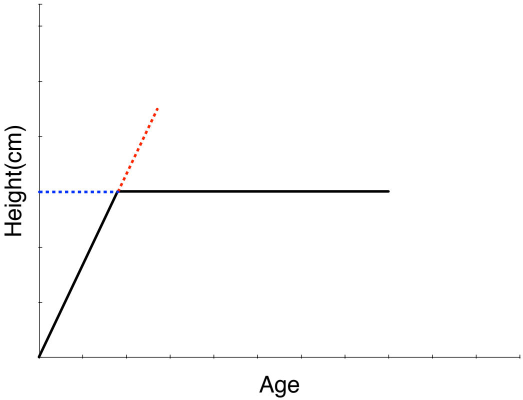
- If param vector for jth Dirichlet is α_j, then PDF is:

$$P(x_1, x_2, \ldots, x_n) = \prod_i \left(\sum_j \pi_j \text{Dirichlet}(x_i | \alpha_j) \right)$$

- Take all the papers Chris has written, all the papers Luay has written
- Dictionary has words: $\langle$ biology, databases, phylogenetics, statistics $\rangle$
- Might get $\alpha_{\text{Chris}} = \langle .5, 12, 1.2, 8 \rangle$
- Means Chris' mean probability vector is $\langle \frac{.5}{21.7}, \frac{12}{21.7}, \frac{1{,}2}{21.7}, \frac{8}{21.7} \rangle = \langle 0.02, 0.55, 0.055, 0.37 \rangle$
- For most docs, Chris is most likely to write the word "databases"
- Might get $\alpha_{\text{Luay}} = \langle 11, 2, 18, 3.2 \rangle$
- His mean probability vector is $\langle 0.32, 0.06, 0.53, 0.09 \rangle$
- Means for most docs, he's most likely to write the word "phylogenetics"
- So, we have learned how likely each topic is to appear in each author's publications

Mixture of Experts

- Used for regression/classification
- Uses a combination of simpler models to form a better model
- Each simpler model covers a range of input values
- We have a switch that determines which model is in place



Mixture of Experts

- In “classical” GLMs
- Use dot product $x \cdot r = \sum_j x_j \times r_j$ to get “natural” parameter for error distribution
- Ex: Normal (least squares) regression:

$$f(y|x) = \text{Normal}(y|x \cdot r, \sigma^2)$$

Mixture of Experts

- In “classical” GLMs
- Use dot product $x \cdot r = \sum_j x_j \times r_j$ to get “natural” parameter for error dist
- Ex: Normal (least squares) regression:

$$f(y|x) = \text{Normal}(y|x \cdot r, \sigma^2)$$

- Now, let’s have a mixture of these
  - Instead of r, we have $r_1, r_2, \ldots, r_k$
  - Instead of σ^2, we have $\sigma_1^2, \sigma_2^2, \ldots, \sigma_k^2$

Mixture of Experts

- Mixture proportions computed by looking at x
  - Each component has a gating vector η
  - We use a "softmax gating network" to get π for a given x... that is:

$$\pi_j = \frac{\exp(x \cdot \eta_j)}{\sum_{j'} \exp(x \cdot \eta_{j'})}$$

- In other words, we normalize the mixture selections to sum to 1

Mixture of Experts

- If param vector for jth Dirichlet is α_j, then PDF is:

$$P(x_1, x_2, \ldots, x_n) = \prod_i \left(\sum_j \pi_j \text{Dirichlet}(x_i | \alpha_j) \right)$$

- So PDF is:

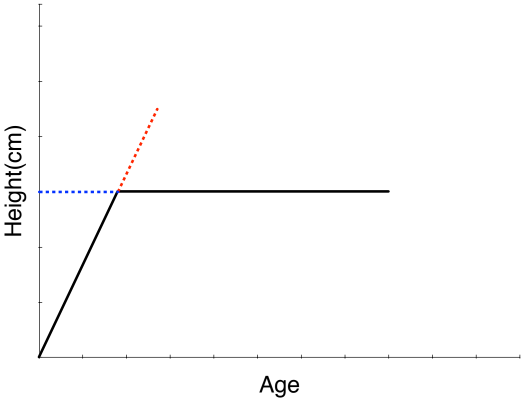
$$P(x_1, x_2, \ldots, x_n) = \prod_i \left(\sum_j \frac{\exp(x_i \cdot \eta_j)}{\sum_j' \exp(x_i \cdot \eta_{j'})} \text{Normal}(x_i | x_i \cdot r_j, \sigma_j^2) \right)$$

Mixture of Experts: What's It Good For?

- Allows more flexible regression/classification models
- Example: we want to predict height y in cm using age x
  - We know people grow as children, then stop somewhere between age 16 and 21
  - Model as a mixture of two Normal regression models
  - Allow bias by making input 2-D $\langle \text{age}, 1 \rangle$
  - Then second regression coef is the bias
  - $r_1 = \langle 6.5, 50 \rangle$ (person starts off at 50cm at birth, grows 6.5cm per year)
  - $r_2 = \langle 0, 175 \rangle$ (person hits around 175cm and stops growing)
- Then gating function
  - $\eta_0 = \langle 0, 8 \rangle$, $\eta_1 = \langle 1, -8 \rangle$... you can do the math...
  - At age 10, prob of first Normal is 0.9995
  - At age 14, prob of first Normal reduces to 0.88
  - At age 16, prob is 50/50
  - At age 18, down to 0.12
  - Down to 0.007 at age 21... means almost everyone has stopped growing

Mixture of Experts Motivation

- In reality, there are really 2 or 3 linear relationships
- Typically, we know what part of the line you are on



Mixture of Experts Summary

- Used for regression/classification
  - In "classical" GLMs
  - Use dot product $x \cdot r = \sum_j x_j \times r_j$ to get "natural" parameter for error distribution
  - Ex: Normal (least squares) regression:

$$f(y|x) = \text{Normal}(y|x \cdot r, \sigma^2)$$

- Now, let's have a mixture of these
  - Instead of r, we have $r_1, r_2, ..., r_k$
  - Instead of σ^2, we have $\sigma_1^2, \sigma_2^2, ..., \sigma_k^2$
- And then we use a "softmax gating network" to get π... that is:

$$\pi_j = \frac{\exp(x \cdot \eta_j)}{\sum_{j'} \exp(x \cdot \eta_{j'})}$$

Mixture of Experts Summary

- So PDF is:

$$P(x_1, x_2, \ldots, x_n) = \prod_i \left(\sum_j \frac{\exp(x_i \cdot \eta_j)}{\sum_j' \exp(x_i \cdot \eta_{j'})} \text{Normal}(x_i | x_i \cdot r_j, \sigma_j^2) \right)$$

- η_j is the set of gating coefficients for mixture j
- The denominator normalizes the values, so we have probabilities
- Basically, we are using a softmax function to decide the mixture
- π was a vector before
- Now, π is a vector whose value is dependent on the vector of xs
- If we have a large dot product value it's highly likely that the y value (output) will be generated by the jth component
- Smaller ages are to the left, larger are to the right

General Question: How to Choose Mixture Size

- In last example...
  - With four mixture components, could have two growth trends (women and men)
  - So probably, four is better than two here
- How to choose in general case?
  - Guess! Then see if it is useful
- Not good enough? Are some alternatives
  - Information theoretic methods (choose model that is able to encode data most efficiently)
  - Bayesian methods (put a prior on the number of components... “infinite mixture models”)

Questions?

- What do we know now that we didn't know before?
  - We understand that data can be generated by /represented as a mixture of different distributions
  - We know about a Dirichlet distribution that can generate probability vectors
- How can we use what we learned today?
  - We can better model complex data by using mixture models
  - We can combine simpler models into more complex models to get better predictions