CPSC 440: Advanced Machine Learning

Mixture Models

Mark Schmidt

University of British Columbia

Winter 2021

Last Time: Properties of Multivariate Gaussian

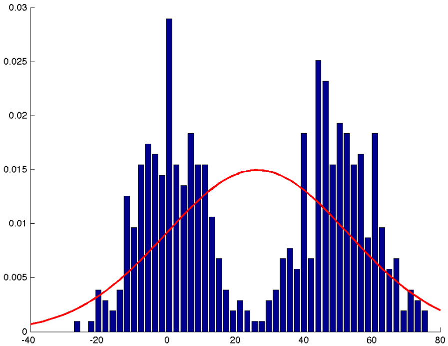
- Consider modeling density of "grades" data:

Math	Physics	Biology	English
72	57	53	87
88	84	73	75
64	70	75	70
⋮	⋮	⋮	⋮

  - Might expect $\Sigma_{\text{Math,Physics}} > 0$ and $\Sigma_{\text{Math,English}} < 0$.
- Last time we discussed how Gaussians are closed under many operations.
  - Affine transformation, marginalization, conditioning, product.
- These properties are what allow us to easily do inference with Gaussians.
  - We can compute likelihood of data $p(x)$ by plugging into formula.
    - What is likelihood of getting $\begin{bmatrix}80 & 80 & 80 & 80\end{bmatrix}$?
  - We can computie a marginal likelihood like $p(x_j)$?
    - What is likelihood of getting 75 in physics? What is probability of getting > 75?
  - Computing a conditional likelihood $p(x_j \mid x_{j'})$.
    - If I got 80 in math, what is likelihood if getting 75 in physics?

1 Gaussian for Multi-Modal Data

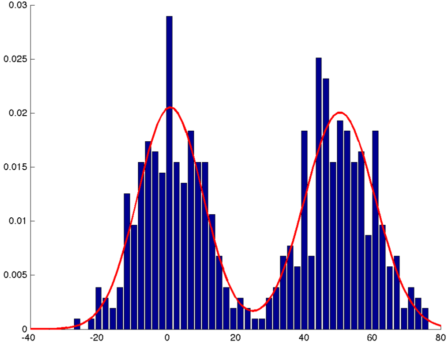
- Major drawback of Gaussian is that it's uni-modal.
  - It gives a terrible fit to data like this:



- If Gaussians are all we know, how can we fit this data?

2 Gaussians for Multi-Modal Data

- We can fit this data by using two Gaussians



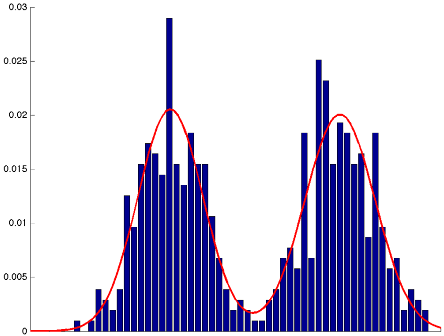
- Half the samples are from Gaussian 1, half are from Gaussian 2.

Mixture of Gaussians

- Our probability density in this example is given by

$$p(x^i \mid \mu_1, \mu_2, \Sigma_1, \Sigma_2) = \frac{1}{2}\underbrace{p(x^i \mid \mu_1, \Sigma_1)}_{\text{PDF of Gaussian 1}} + \frac{1}{2}\underbrace{p(x^i \mid \mu_2, \Sigma_2)}_{\text{PDF of Gaussian 2}},$$

  - We need the (1/2) factors so it still integrates to 1.

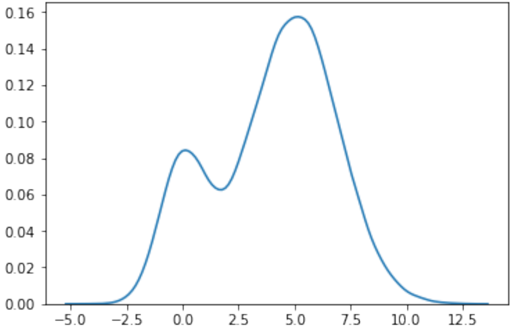


Mixture of Gaussians

- If data comes from one Gaussian more often than the other, we could use

$$p(x^i \mid \mu_1, \mu_2, \Sigma_1, \Sigma_2, \pi_1, \pi_2) = \pi_1 \underbrace{p(x^i \mid \mu_1, \Sigma_1)}_{\text{PDF of Gaussian 1}} + \pi_2 \underbrace{p(x^i \mid \mu_2, \Sigma_2)}_{\text{PDF of Gaussian 2}},$$

where π_1 and π_2 are non-negative and sum to 1.
  - π_1 represents "probability that we take a sample from Gaussian 1".

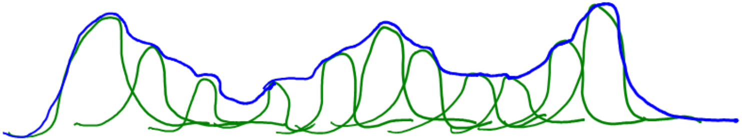


Mixture of Gaussians

- In general we might have a mixture of k Gaussians with different weights.

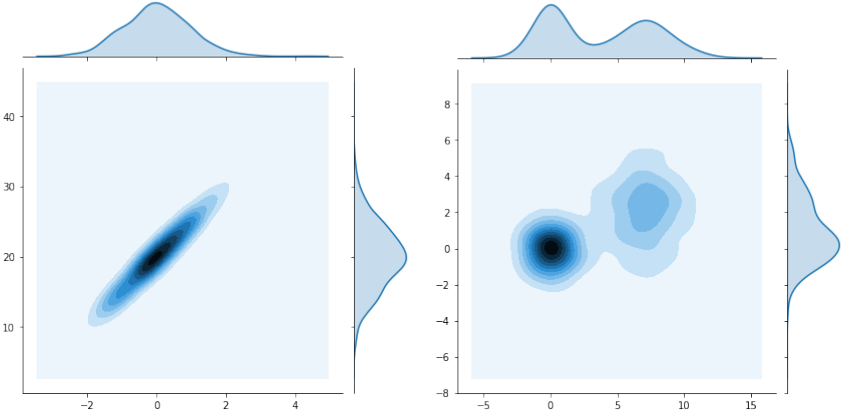
$$p(x \mid \mu, \Sigma, \pi) = \sum_{c=1}^{k} \pi_c \underbrace{p(x \mid \mu_c, \Sigma_c)}_{\text{PDF of Gaussian } c},$$

  - Where π_c are categorical distribution parameters (non-negative and sum to 1).
  - We can use it to model complicated densities with Gaussians (like RBFs).
    - "Universal approximator": can model any continuous density on compact set.



Mixture of Gaussians

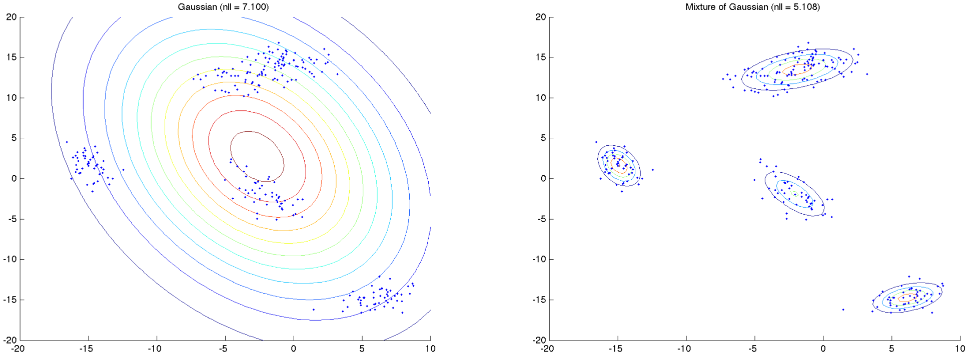
- Gaussian vs. mixture of 2 Gaussian densities in 2D:



- Marginals will also be mixtures of Gaussians.

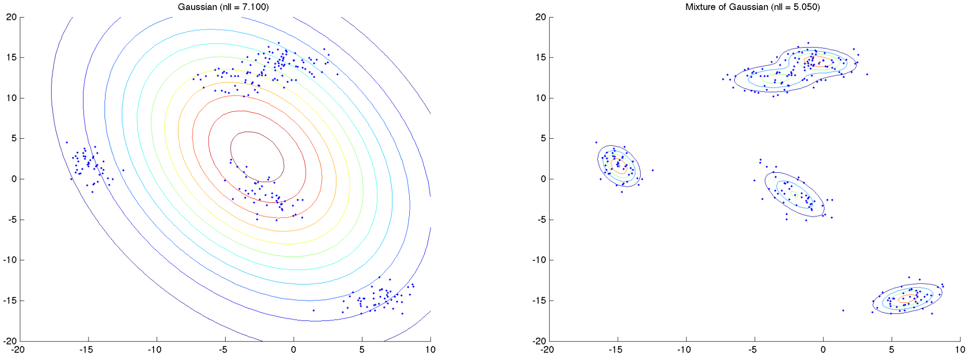
Mixture of Gaussians

- Gaussian vs. Mixture of 4 Gaussians for 2D multi-modal data:



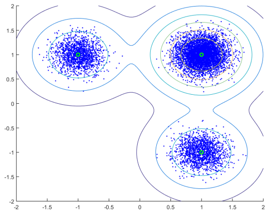
Mixture of Gaussians

- Gaussian vs. Mixture of 5 Gaussians for 2D multi-modal data:



Mixture of Gaussians

- Given parameters $\{\pi_c, \mu_c, \Sigma_c\}$, we can sample from a mixture of Gaussians using:
  1. Sample cluster c based on prior probabilities π_c (categorical distribution).
  2. Sample example x based on mean μ_c and covariance Σ_c.



- We usually fit these models with expectation maximization (EM):
  - An optimization method that gives closed-form updates for this model.
    - We'll cover EM later.
  - To choose k, we might use domain knowledge or test set likelihood.

Previously: Independent vs. General Discrete Distributions

- We previously considered density estimation with discrete variables,

$$X = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix},$$

  and considered two extreme approaches:
  - Product of independent Bernoullis:

$$p(x^i \mid \theta) = \prod_{j=1}^{d} p(x_j^i \mid \theta_j).$$

    Easy to fit but strong independence assumption:
    - Knowing x_j^i tells you nothing about x_k^i.
  - General discrete distribution:

$$p(x^i \mid \theta) = \theta_{x^i}.$$

    No assumptions but hard to fit:
    - Parameter vector θ_{x^i} for each possible x^i.
- A model in between these two is the mixture of Bernoullis.

Mixture of Bernoullis

- Consider a coin flipping scenario where we have two coins:
  - Coin 1 has $\theta_1 = 0.5$ (fair) and coin 2 has $\theta_2 = 1$ (biased).
- Half the time we flip coin 1, and otherwise we flip coin 2:

$$\begin{aligned} p(x^i = 1 \mid \theta_1, \theta_2) &= \pi_1 p(x^i = 1 \mid \theta_1) + \pi_2 p(x^i = 1 \mid \theta_2) \\ &= \frac{1}{2}\theta_1 + \frac{1}{2}\theta_2 = \frac{\theta_1 + \theta_2}{2} \end{aligned}$$

- With one variable this mixture model is not very interesting:
  - It's equivalent to flipping one coin with $\theta = 0.75$.
- But with multiple variables mixture of Bernoullis can model dependencies...

Mixture of Independent Bernoullis

- Consider a mixture of independent Bernoullis:

$$p(x \mid \theta_1, \theta_2) = \frac{1}{2} \underbrace{\prod_{j=1}^{d} p(x_j \mid \theta_{1j})}_{\text{first set of Bernoullis}} + \frac{1}{2} \underbrace{\prod_{j=1}^{d} p(x_j \mid \theta_{2j})}_{\text{second set of Bernoulli}} .$$

- Conceptually, we now have two sets of coins:
  - Half the time we throw the first set, half the time we throw the second set.

- With $d = 4$ we could have $\theta_1 = \begin{bmatrix} 0 & 0.7 & 1 & 1 \end{bmatrix}$ and $\theta_2 = \begin{bmatrix} 1 & 0.7 & 0.8 & 0 \end{bmatrix}$.
  - Half the time we have $p(x_3^i = 1) = 1$ and half the time it's 0.8.

- Have we gained anything?

Mixture of Independent Bernoullis

- Example from the previous slide: $\theta_1 = \begin{bmatrix} 0 & 0.7 & 1 & 1 \end{bmatrix}$ and $\theta_2 = \begin{bmatrix} 1 & 0.7 & 0.8 & 0 \end{bmatrix}$.
- Here are some samples from this model:

$$X = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix}$$

- Unlike product of Bernoullis, notice that features in samples are not independent.
  - In this example knowing $x_1 = 1$ tells you that $x_4 = 0$.

- This model can capture dependencies: $\underbrace{p(x_4 = 1 \mid x_1 = 1)}_{0} \neq \underbrace{p(x_4 = 1)}_{0.5}$.

Mixture of Independent Bernoullis

- General mixture of independent Bernoullis:

$$p(x^i \mid \Theta) = \sum_{c=1}^{k} \pi_c p(x^i \mid \theta_c) = \sum_{c=1}^{k} \pi_c \prod_{j=1}^{d} \theta_{cj},$$

where Θ contains all the model parameters.

  - Θ has k values of π_c and $k \times d$ values of θ_{cj}.

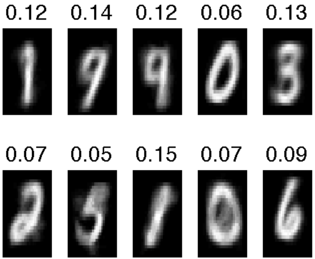
- Mixture of Bernoullis can model dependencies between variables
  - Individual mixtures act like clusters of the binary data.
  - Knowing cluster of one variable gives information about other variables.

- With k large enough, mixture of Bernoullis can model any discrete distribution.
  - Hopefully with $k << 2^d$.

Mixture of Independent Bernoullis

- Plotting parameters θ_c with 10 mixtures trained on MNIST digits (with "EM"):

(numbers above images are mixture coefficients π_c)



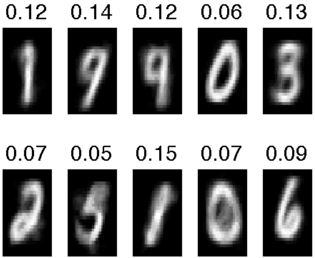
http://pmtk3.googlecode.com/svn/trunk/docs/demoOutput/bookDemos/%2811%29-Mixture_models_and_the_EM_algorithm/mixBerMnistEM.html

- Remember this is unsupervised: it hasn't been told there are ten digits.
  - Density estimation is trying to figure out how the world works.

Mixture of Independent Bernoullis

- Plotting parameters θ_c with 10 mixtures trained on MNIST digits (with "EM"):

(numbers above images are mixture coefficients π_c)

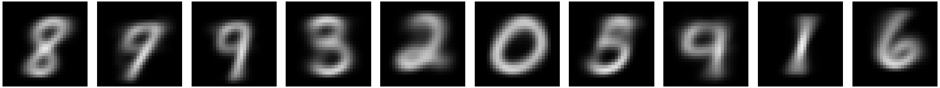


http://pmtk3.googlecode.com/svn/trunk/docs/demoOutput/bookDemos/%2811%29-Mixture_models_and_the_EM_algorithm/mixBerMnistEM.html

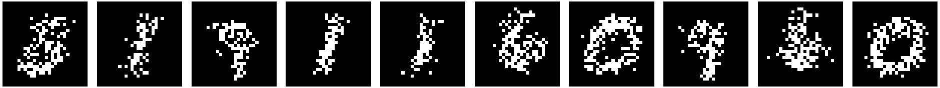
- You could use this model to "fill in" missing parts of an image:
  - By finding likely cluster/mixture, you find likely values for the missing parts.

Mixture of Bernoullis on Digits with $k > 10$

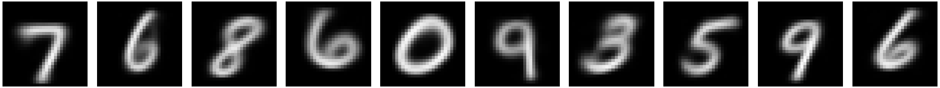
- Parameters of a mixture of Bernoulli model fit to MNIST with $k = 10$:



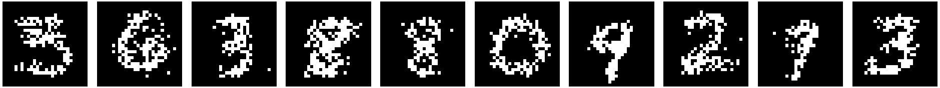
- Shapes of samples are better, but missing within-cluster dependencies:



- You get a better model with $k > 10$. First 10 components with $k = 50$:



- Samples from the $k = 50$ model (can have more than one "type" of a number):

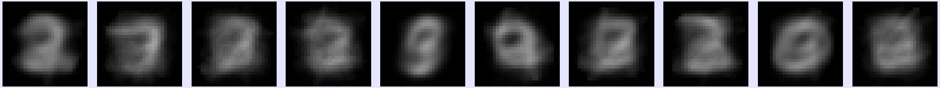


Summary

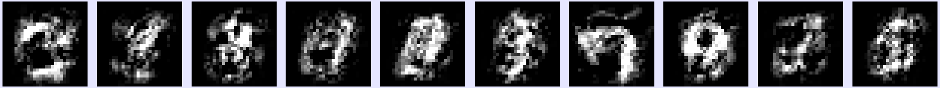
- Mixture of Gaussians writes probability as convex comb. of Gaussian densities.
  - Can model arbitrary continuous densities.
- Mixture of Bernoullis can model dependencies between discrete variables.
  - Probability of belonging to mixtures is a soft-clustering of examples.
- Next time: dealing with missing data.

Mixture of Gaussians on Digits

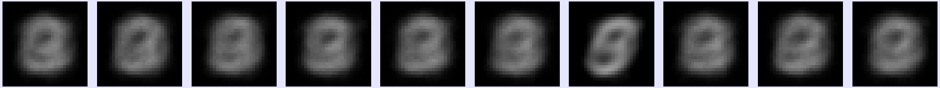
- Mean parameters of a mixture of Gaussians with $k = 10$:



- Samples:



- 10 components with $k = 50$ (I might need a better initialization):



- Samples: