CPSC 540: Machine Learning

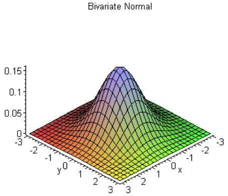
Mixture Models

Mark Schmidt

University of British Columbia

Winter 2020

Last Time: Multivariate Gaussian



http://personal.kenyon.edu/hartlaub/MellonProject/Bivariate2.html

- The multivariate normal/Gaussian distribution models PDF of vector x^i as

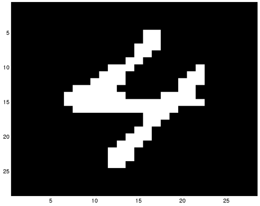
$$p(x^i \mid \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x^i - \mu)^\top \Sigma^{-1}(x^i - \mu)\right)$$

  where $\mu \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d \times d}$ and $\Sigma \succ 0$.
- Density for a linear transformation of a product of independent Gaussians.
- Diagonal Σ implies independence between variables.

Example: Multivariate Gaussians on Digits

- Recall the task of density estimation with handwritten images of digits:

$x^i = \text{vec}\Big($

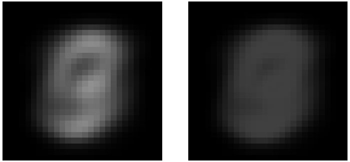


$\Big),$

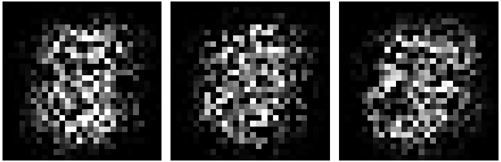
- Let's treat this as a continuous density estimation problem.

Example: Multivariate Gaussians on Digits

- MLE of parameters using independent Gaussians (diagonal Σ):
  - Mean μ_j (left) and variance σ_j^2 (right) for each feature.



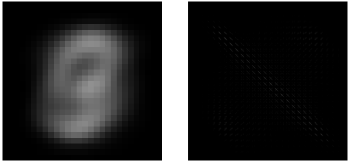
- Samples generate from this model:



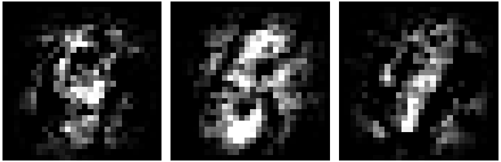
- Because Σ is diagonal, doesn't model dependencies between pixels.

Example: Multivariate Gaussians on Digits

- MLE of parameters using multivariate Gaussians (dense Σ):



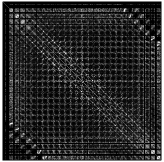
  - μ is the same, the $d \times d$ matrix Σ is degenerate (need to zoom in to see anything).
- Samples generate from this model:



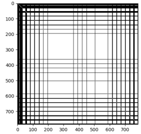
- Captures some pairwise dependencies between pixels, but not expressive enough.

Graphical LASSO on Digits

- MAP estimate of precision matrix Θ with regularizer $\lambda \text{Tr}(\Theta)$ (with $\lambda = 1/n$).



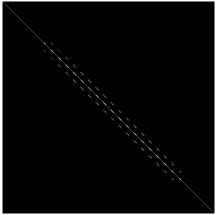
- Sparsity pattern using this "L1-regularization of the trace":



- Doesn't yield a sparse matrix (only zeroes are with pixels near the boundary).

Graphical LASSO on Digits

- Sparsity pattern if we instead use the graphical LASSO:
  - MAP estimate of precision matrix Θ with regularizer $\lambda\|\Theta\|_1$ (with $\lambda = 1/8$).



- The graph represented by this adjacency matrix is (roughly) the 2d image lattice.
  - Pixels that are near each other in the image end up being connected by an edge.

- Examples:
  - https://normaldeviate.wordpress.com/2012/09/17/high-dimensional-undirected-graphical-models

Closedness of Multivariate Gaussian

- Multivariate Gaussian has nice properties of univariate Gaussian:
  - Closed-form MLE for μ and Σ given by sample mean/variance.
  - Central limit theorem: mean estimates of random variables converge to Gaussians.
  - Maximizes entropy subject to fitting mean and covariance of data.
- A crucial computational property: Gaussians are closed under many operations.
  1. Affine transformation: if $p(x)$ is Gaussian, then $p(Ax + b)$ is a Gaussian[1].
  2. Marginalization: if $p(x, z)$ is Gaussian, then $p(x)$ is Gaussian.
  3. Conditioning: if $p(x, z)$ is Gaussian, then $p(x \mid z)$ is Gaussian.
  4. Product: if $p(x)$ and $p(z)$ are Gaussian, then $p(x)p(z)$ is proportional to a Gaussian.
- Most continuous distributions don't have these nice properties.

[1] Could be degenerate with $|\Sigma| = 0$, dependending on particular A.

Affine Property: Special Case of Shift

- Assume that random variable x follows a Gaussian distribution,
$$x \sim \mathcal{N}(\mu, \Sigma).$$
- And consider an shift of the random variable,
$$z = x + b.$$
- Then random variable z follows a Gaussian distribution
$$z \sim \mathcal{N}(\mu + b, \Sigma),$$
where we've shifted the mean.

Affine Property: General Case

- Assume that random variable x follows a Gaussian distribution,

$$x \sim \mathcal{N}(\mu, \Sigma).$$

- And consider an affine transformation of the random variable,

$$z = Ax + b.$$

- Then random variable z follows a Gaussian distribution

$$z \sim \mathcal{N}(A\mu + b, A\Sigma A^\top),$$

although note we might have $|A\Sigma A^\top| = 0$.

Marginalization of Gaussians

- Consider a dataset where we've partitioned the variables into two sets:

$$X = \begin{bmatrix} | & | & | & | \\ x_1 & x_2 & z_1 & z_2 \\ | & | & | & | \end{bmatrix}.$$

- It's common to write multivariate Gaussian for partitioned data as:

$$\begin{bmatrix} x \\ z \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_x \\ \mu_z \end{bmatrix}, \begin{bmatrix} \Sigma_{xx} & \Sigma_{xz} \\ \Sigma_{zx} & \Sigma_{zz} \end{bmatrix} \right),$$

- If I want the marginal distribution $p(x)$, I can use the affine property,

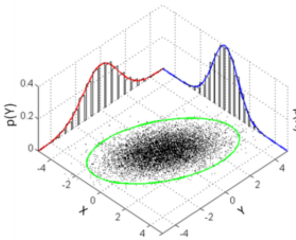
$$x = \underbrace{\begin{bmatrix} I & 0 \end{bmatrix}}_{A} \begin{bmatrix} x \\ z \end{bmatrix} + \underbrace{0}_{b},$$

  to get that

$$x \sim \mathcal{N}(\mu_x, \Sigma_{xx}).$$

Marginalization of Gaussians

- In a picture, ignoring a subset of the variables gives a Gaussian:



https://en.wikipedia.org/wiki/Multivariate_normal_distribution

- This seems less intuitive if you use rules of probability to marginalize:

$$p(x) = \int_{z_1}\int_{z_2}\cdots\int_{z_d} \frac{1}{(2\pi)^{\frac{d}{2}}\left|\begin{bmatrix}\Sigma_{xx} & \Sigma_{xz}\\ \Sigma_{zx} & \Sigma_{zz}\end{bmatrix}\right|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\left(\begin{bmatrix}x\\ z\end{bmatrix} - \begin{bmatrix}\mu_x\\ \mu_z\end{bmatrix}\right)\begin{bmatrix}\Sigma_{xx} & \Sigma_{xz}\\ \Sigma_{zx} & \Sigma_{zz}\end{bmatrix}^{-1}\left(\begin{bmatrix}x\\ z\end{bmatrix} - \begin{bmatrix}\mu_x\\ \mu_z\end{bmatrix}\right)\right) dz_d dz_{d-1}\ldots dz_1.$$

Conditioning in Gaussians

- Again consider a partitioned Gaussian,

$$\begin{bmatrix} x \\ z \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_x \\ \mu_z \end{bmatrix}, \begin{bmatrix} \Sigma_{xx} & \Sigma_{xz} \\ \Sigma_{zx} & \Sigma_{zz} \end{bmatrix}\right).$$

- The conditional probabilities are also Gaussian,

$$x \mid z \sim \mathcal{N}(\mu_{x \mid z}, \Sigma_{x \mid z}),$$

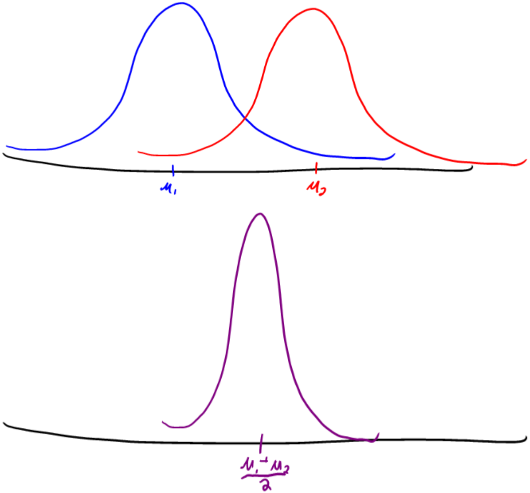
where

$$\mu_{x \mid z} = \mu_x + \Sigma_{xz}\Sigma_{zz}^{-1}(z - \mu_z), \quad \Sigma_{x \mid z} = \Sigma_{xx} - \Sigma_{xz}\Sigma_{zz}^{-1}\Sigma_{zx}.$$

- "For any fixed z, the distribution of x is a Gaussian".
  - Notice that if $\Sigma_{xz} = 0$ then x and z are independent ($\mu_{x \mid z} = \mu_x$, $\Sigma_{x \mid z} = \Sigma_x$).
  - We previously saw the special case where Σ is diagonal (all variables independent).

Product of Gaussian Densities

- If $\Sigma_1 = I$ and $\Sigma_2 = I$ then product of PDFs has $\Sigma = \frac{1}{2}I$ and $\mu = \frac{\mu_1+\mu_2}{2}$.



Product of Gaussian Densities

- Let $f_1(x)$ and $f_2(x)$ be Gaussian PDFs defined on variables x.
- The product of the PDFs $f_1(x)f_2(x)$ is proportional to a Gaussian density,
  - With (μ_1, Σ_1) as parameters of f_1 and (μ_2, Σ_2) for f_2:

$$\text{covariance of} \quad \Sigma = (\Sigma_1^{-1} + \Sigma_2^{-1})^{-1}.$$

$$\text{mean of } \mu = \Sigma\Sigma_1^{-1}\mu_1 + \Sigma\Sigma_2^{-1}\mu_2,$$

although this density may not be normalized (may not integrate to 1 over all x).

Product of Gaussian Densities

- So if we can write a probability as $p(x) \propto f_1(x) f_2(x)$ for 2 Gaussians, then p is a Gaussian with known mean/covariance.

- Example of a Gaussian likelihood $p(x^i \mid \mu, \Sigma)$ and Gaussian prior $p(\mu \mid \mu_0, \Sigma_0)$.
  - Posterior for μ will be Gaussian:

$$\begin{aligned} p(\mu \mid x^i, \Sigma, \mu_0, \Sigma_0) &\propto p(x^i \mid , \mu, \Sigma) p(\mu \mid \mu_0, \Sigma_0) \\ &= p(\mu \mid x^i, \Sigma) p(\mu \mid \mu_0, \Sigma_0) \qquad \text{(symmetry of } x^i \text{ and } \mu) \\ &= \text{(some Gaussian)}. \end{aligned}$$

- Non-example of $p(x_2 \mid x_1)$ being Gaussian and $p(x_1 \mid x_2)$ being Gaussian.
  - Product $p(x_2 \mid x_1) p(x_1 \mid x_2)$ may not be a proper distribution.
  - Although we saw it will be a Gaussian if they are independent.

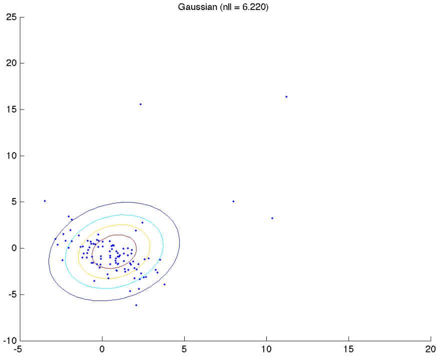
- "Product of Gaussian densities" will be used later in Gaussian Markov chains.

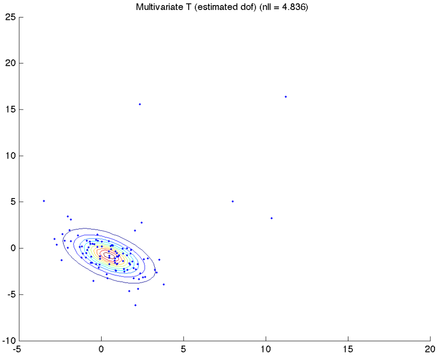
Properties of Multivariate Gaussians

- A multivariate Gaussian "cheat sheet" is here:
  - https://ipvs.informatik.uni-stuttgart.de/mlr/marc/notes/gaussians.pdf
- For a careful discussion of Gaussians, see the playlist here:
  - https://www.youtube.com/watch?v=TC0ZAX3DA88&t=2s&list=PL17567A1A3F5DB5E4&index=34

Problems with Multivariate Gaussian

- Why not the multivariate Gaussian distribution?
  - Still not robust, may want to consider multivariate Laplace or multivariate T.

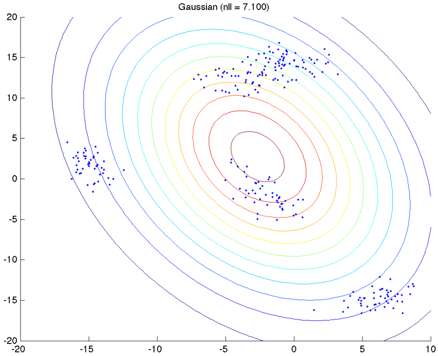




  - These require numerical optimization to compute MLE/MAP.

Problems with Multivariate Gaussian

- Why not the multivariate Gaussian distribution?
  - Still not robust, may want to consider multivariate Laplace of multivariate T.
  - Still unimodal, which often leads to very poor fit.



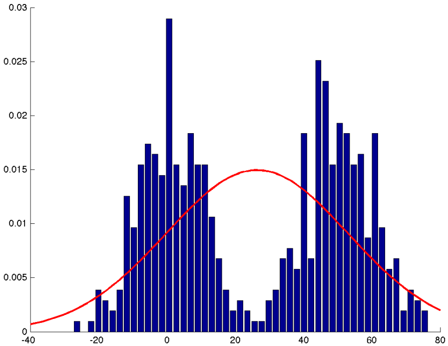
Outline

1 Properties of Multivariate Gaussian

2 Mixture Models

1 Gaussian for Multi-Modal Data

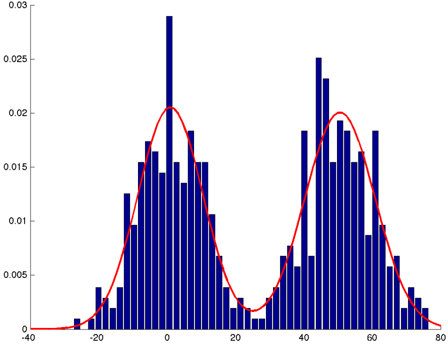
- Major drawback of Gaussian is that it's uni-modal.
  - It gives a terrible fit to data like this:



- 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



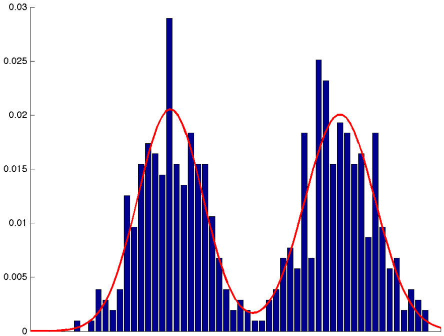
- Half the samples are from Gaussian 1, half are from Gaussian 2.

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.

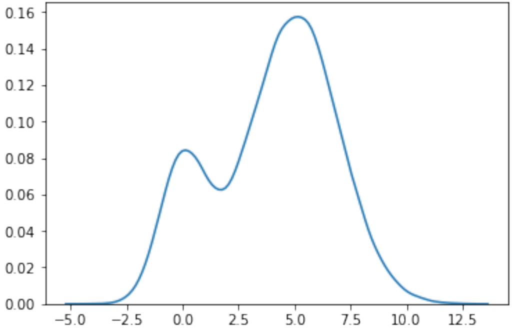


Mixture of Gaussians

- 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".

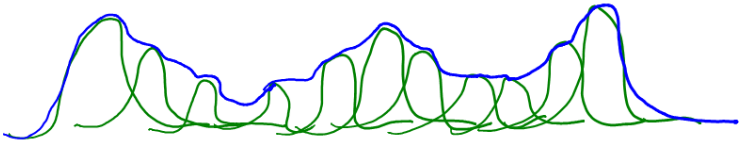


Mixture of Gaussians

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

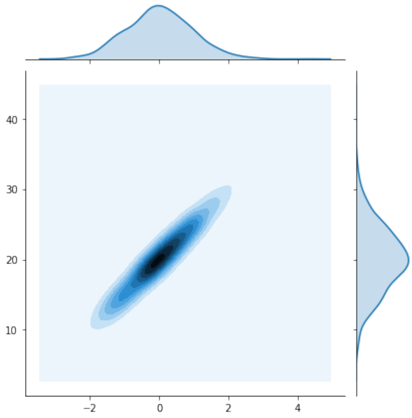
$$p(x \mid \mu, \Sigma, \pi) = \sum_{c=1}^{k} \pi_c \underbrace{p(x \mid \mu_c, \Sigma_c)}_{\text{PDF of Gaussian } c},$$

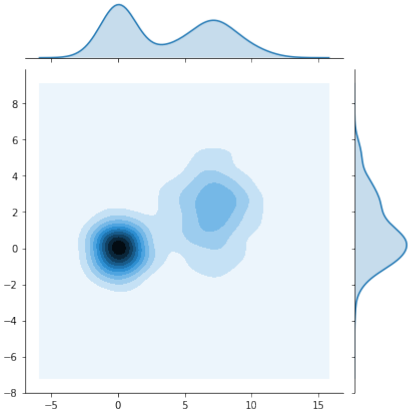
  - Where π is a categorical variable (the π_c are non-negative and sum to 1).
  - We can use it to model complicated densities with Gaussians (like RBFs).
    - "Universal approximator": can model any continuous density on compact set.



Mixture of Gaussians

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

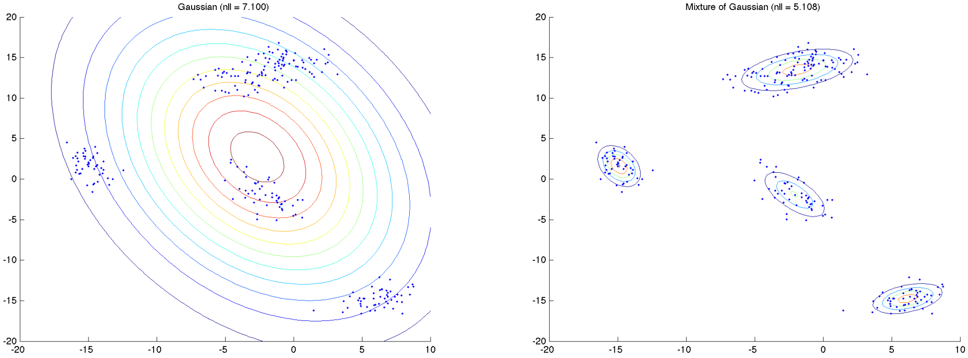




- Marginals will also be mixtures of Gaussians.

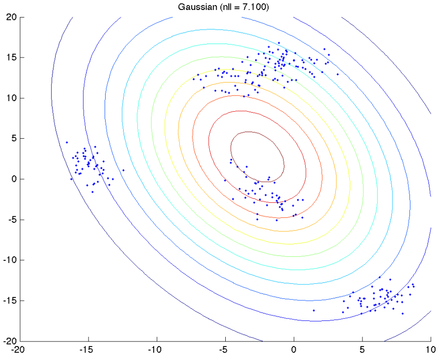
Mixture of Gaussians

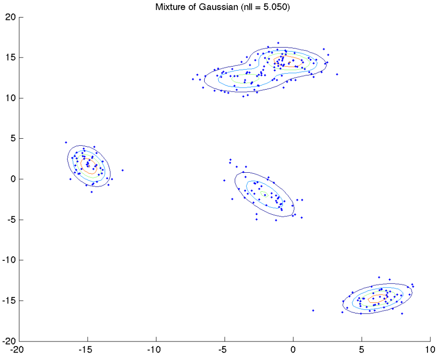
- Gaussian vs. Mixture of 4 Gaussians for 2D multi-modal data:



Mixture of Gaussians

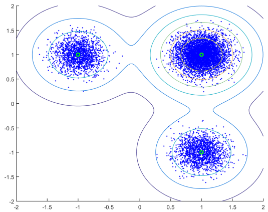
- Gaussian vs. Mixture of 5 Gaussians for 2D multi-modal data:





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.
  - 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:

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

    Easy to fit but strong independence assumption:
    - Knowing x^i_j tells you nothing about x^i_k.
  - 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:

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

  where Θ contains all the model parameters.

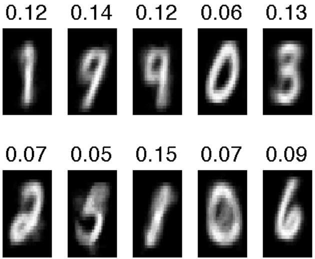
- Mixture of Bernoullis can model dependencies between variables
  - Individual mixtures act like clusters of the binary data.
  - Knowing cluster of one variable gives information about other variables.

- 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)



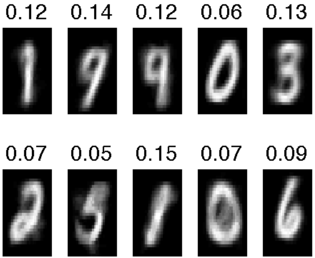
http://pmtk3.googlecode.com/svn/trunk/docs/demoOutput/bookDemos/%2811%29-Mixture_models_and_the_EM_algorithm/mixBerMnistEM.html

- 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.

Summary

- Properties of multivariate Gaussian:
  - Closed under affine transformations, marginalization, conditioning, and products.
  - But unimodal and not robust.
- 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.