Mixtures of Bernoulli Distributions

Sargur Srihari
srihari@cedar.buffalo.edu

Mixtures of Bernoulli Distributions

- GMMs are defined over continuous variables
- Now consider mixtures of discrete binary variables: Bernoulli distributions (BMMs)
- Sets foundation of HMM over discrete variables
- We begin by defining:
  1. Bernoulli
  2. Multivariate Bernoulli
  3. Mixture of Bernoulli
  4. Mixture of multivariate Bernoulli

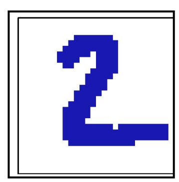
Bernoulli Distribution

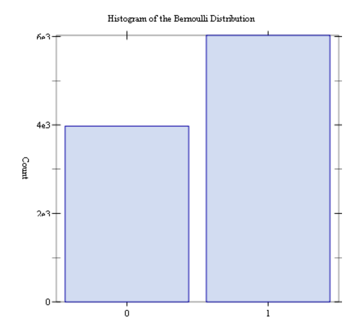
1. A coin has a Bernoulli distribution

$$p(x \mid \mu) = \mu^x (1-\mu)^{1-x}$$

where $x=0,1$

2. Each pixel of a binary image has a Bernoulli distribution.



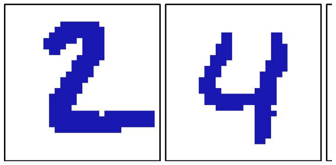


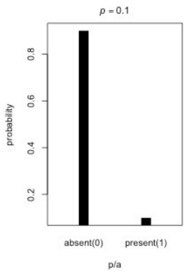
– All D pixels together define a multivariate Bernoulli distribution:

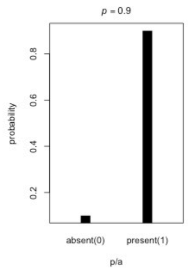
Mixture of Two Bernoullis

- Outcome is single value x, result of tossing one of $K=2$ coins, with parameters $\mu=[\mu_1,\mu_2]$ chosen with probabilities $\pi=[\pi_1,\pi_2]$
- Mixture distribution is

$$p(x \mid \mu,\pi) = \pi_1 \mu_1^x (1-\mu_1)^{1-x} + \pi_2 \mu_2^x (1-\mu_2)^{1-x}$$







Multivariate Bernoulli

- Set of D *independent* binary variables $x_i,\ i=1,..,D$
  - E.g., a set of D coins with heads and tails
- Each governed by parameter μ_i
- Multivariate distribution

$$p(\mathrm{x}|\mu) = \prod_{i=1}^{D} \mu_i^{x_i}(1-\mu_i)^{(1-x_i)}$$

where $\mathrm{x}=(x_1,..,x_D)^T$ and

$\boldsymbol{\mu}=(\mu_1,..,\mu_D)^T$

- Mean and covariance are

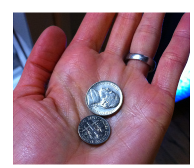
$\mathrm{E}[\mathrm{x}]=\mu,\ \mathrm{cov}[\mathrm{x}]=\mathrm{diag}\{\mu_i(1-\mu_i)\}$

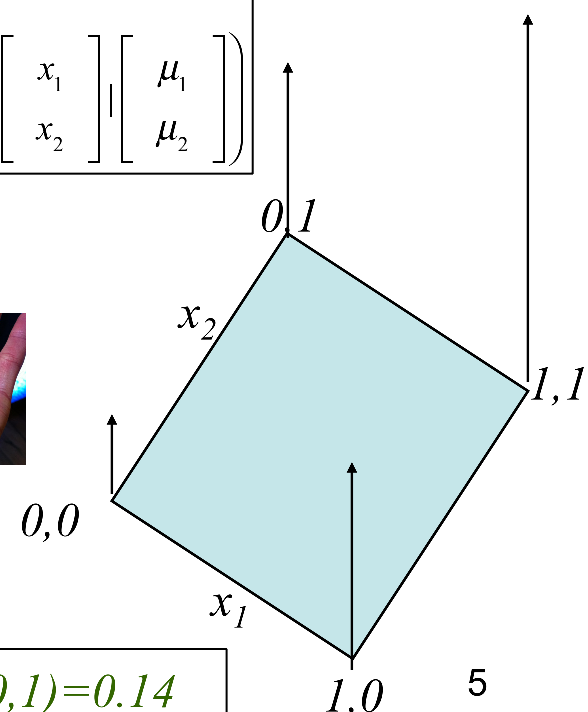
$\mu_1=0.8,\ \mu_2=0.7$: $p(0,0)= 0.06,\ p(1,0)=0.24,\ p(1,1)=0.56,\ p(0,1)=0.14$

Bivariate Bernoulli distribution

$D=2,\ K=2$

$$p(\mathrm{x}\,|\,\boldsymbol{\mu}) = p\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \middle| \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}\right)$$





Mixture of multivariate Bernoulli

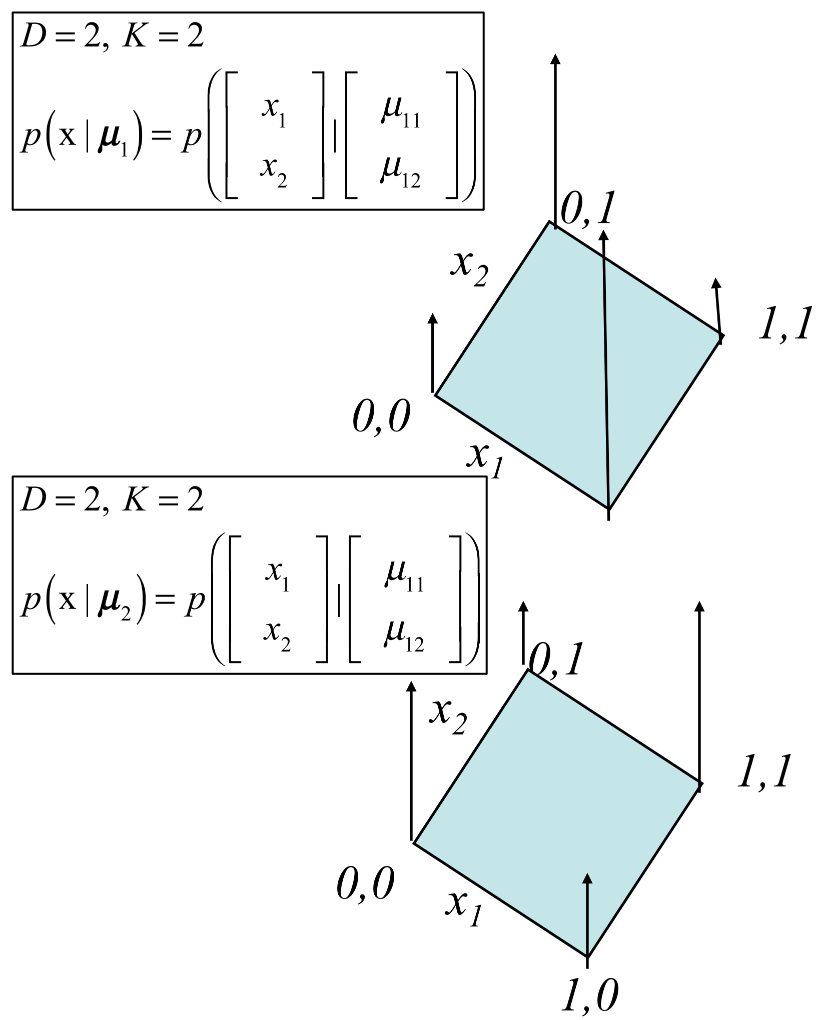
- Finite mixture of K Bernoulli distributions
  - E.g., K bags of D coins each where bag k is chosen with probability π_k

$$\boxed{p(\mathrm{x} \mid \mu, \pi) = \sum_{k=1}^{K} \pi_k \, p(\mathrm{x} \mid \mu_k)}$$

  - where $\boldsymbol{\mu}=\{\mu_1,..,\mu_K\}$, $\pi=\{\pi_1,..,\pi_K\}$, $\sum_k \pi_k = 1$ and

$$p(\mathrm{x} \mid \mu_k) = \prod_{i=1}^{D} \mu_{ki}^{x_i} (1-\mu_{ki})^{(1-x_i)}$$

Two bivariate Bernoulli distributions



Mean and Covariance of BMM

$$E[\mathrm{x}] = \sum_{k=1}^{K} \pi_k \boldsymbol{\mu}_k$$

$$\mathrm{cov}[x] = \sum_{k=1}^{K} \pi_k \left\{ \Sigma_k + \boldsymbol{\mu}_k \boldsymbol{\mu}_k^T \right\} - E[\mathrm{x}]E[\mathrm{x}]^T$$

where $\Sigma_k = diag\{\mu_{k1}(1-\mu_{k1})\}$

- Because the covariance matrix $\mathrm{cov}[\mathrm{x}]$ is no longer diagonal,
  - the mixture distribution can capture correlation between the variables,
  - unlike a single Bernoulli distribution

Log likelihood of Bernoulli mixture

- Given data set $X=\{x_1,..,x_N\}$ log-likelihood of model is

$$\ln p(X \mid \boldsymbol{\mu},\pi)=\sum_{n=1}^{N}\ln\left\{\sum_{k=1}^{K}\pi_k p(x_n \mid \boldsymbol{\mu}_k)\right\}$$

Summation due to logarithm

Summation due to mixture

- Due to summation inside logarithm there is no closed form m.l.e. solution

To define EM, introduce latent variables

- One of K representation $\mathrm{z} = (z_1,..z_K)^{\mathrm{T}}$
- Conditional distribution of x given the latent variable is

$$p(\mathrm{x}|\mathrm{z},\mu) = \prod_{k=1}^{K} p(\mathrm{x}|\mu_k)^{z_k}$$

- Prior distribution for the latent variables is same as for GMM

$$p\left(\mathrm{z} \mid \pi)\right) = \prod_{k=1}^{K} \pi_k^{z_k}$$

- If we form product of $p(\mathrm{x}|\mathrm{z},\mu)$ and $p(\mathrm{z}|\mu)$ and marginalize over z, we recover

$$p(\mathrm{x} \mid \mu,\pi) = \sum_{k=1}^{K} \pi_k p(\mathrm{x}|\mu_k)$$

Complete Data Log-likelihood

- To derive the EM algorithm we first write down the complete data log-likelihood function

$$\ln p\left(X,Z \mid \mu,\pi\right) = \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\left\{\ln \pi_k + \sum_{i=1}^{D}\left\{x_{ni}\ln \mu_{ki} + \left(1-x_{ni}\right)\ln\left(1-\mu_{ki}\right)\right\}\right\}$$

  - where $X=\{\mathrm{x}_\mathrm{n}\}$ and $Z=\{\mathrm{z}_\mathrm{n}\}$

- Taking its expectation w.r.t. posterior distribution of the latent variables

$$E_Z\left[\ln p\left(X,Z \mid \boldsymbol{\mu},\boldsymbol{\pi}\right)\right] = \sum_{n=1}^{N}\sum_{k=1}^{K} \gamma\left(z_{nk}\right)\left\{\ln \pi_k + \sum_{i=1}^{D}\left[x_{ni}\ln \mu_{ki} + \left(1-x_{ni}\right)\ln\left(1-\mu_{ki}\right)\right]\right\}$$

  - where $\gamma(z_{nk})=E[z_{nk}]$ is the posterior probability or responsibility of component k given data point x_n.
- In the E-step these responsibilities are evaluated

E-step: Evaluating Responsibilities

- In the E step the responsibilities $\gamma(z_{nk})=E[z_{nk}]$ are evaluated using Bayes theorem

$$\gamma(z_{nk}) = E[z_{nk}] = \frac{\sum_{z_{nk}} z_{nk}\left[\pi_k p(\mathrm{x}_n \mid \boldsymbol{\mu}_k)\right]^{z_{nk}}}{\sum_{z_{nj}}\left[\pi_j p(\mathrm{x}_n \mid \boldsymbol{\mu}_j)\right]^{z_{nj}}} = \frac{\pi_k p(\mathrm{x}_n \mid \boldsymbol{\mu}_k)}{\sum_{j=1}^{K} \pi_j p(\mathrm{x}_n \mid \boldsymbol{\mu}_j)}$$

- If we consider the sum over n in

$$E_Z\left[\ln p(X,Z \mid \boldsymbol{\mu},\boldsymbol{\pi})\right] = \sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(z_{nk})\left\{\ln \pi_k + \sum_{i=1}^{D}\left[x_{ni}\ln \mu_{ki} + (1-x_{ni})\ln(1-\mu_{ki})\right]\right\}$$

  - Responsibilities enter only through two terms

$$N_k = \sum_{n=1}^{N}\gamma(z_{nk})$$ Effective no. of data points associated with component k

$$\overline{\mathrm{x}}_k = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})\mathrm{x}_n$$

M-step

- In the M step we maximize the expected complete data log-likelihood wrt parameters μ_k and π
- If we set the derivative of

$$E_Z\left[\ln p\left(X,Z\mid \boldsymbol{\mu},\boldsymbol{\pi}\right)\right]=\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma\left(z_{nk}\right)\left\{\ln \pi_k+\sum_{i=1}^{D}\left[x_{ni}\ln \mu_{ki}+\left(1-x_{ni}\right)\ln\left(1-\mu_{ki}\right)\right]\right\}$$

  - wrt μ_k equal to zero and rearrange the terms, we obtain

$$\boldsymbol{\mu}_k=\overline{\boldsymbol{x}}_k$$

  - Thus mean of component k is equal to weighted mean of data with weighting given by responsibilities of component k for the data points
- For maximization wrt π_k, we use a Lagrangian to ensure $\sum_k \pi_k = 1$
  - Following steps similar to GMM we get the reasonable result

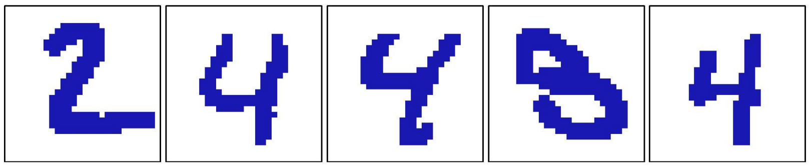
$$\pi_k=\frac{N_k}{N}$$

No singularities in BMMs

- In contrast to GMMs no singularities with BMMs
- Can be seen as follows
  - Likelihood function is bounded above because $0 \leq p(\mathbf{x}_n|\boldsymbol{\mu}_k) \leq 1$
  - There exist singularities at which the likelihood function goes to zero
    - But these will not be found by EM provided it is not initialized to a pathological starting point
      - Because EM always increases value of likelihood function

Illustrate BMM with Handwritten Digits

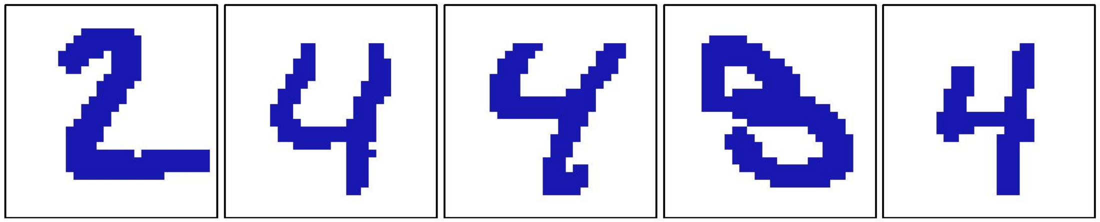
- Given a set of unlabeled digits 2,3 and 4



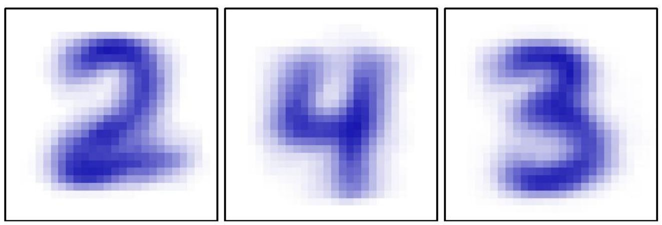
- Images turned to binary:
  set values >0.5 to 1, rest to 0
- Fit a data set of $N = 600$ digits, $K=3$
- 10 iterations of EM
- Mixing coefficients initialized with $\pi_k=1/K$

Parameters μ_{ki} were set to random values chosen uniformly in range $(0.25,0.75)$ and normalized so that $\Sigma_j \mu_{kj}=1$

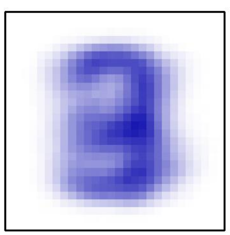
Result of BMM



EM finds the three clusters corresponding to different digits



Parameters μ_{ki} for each of three components of mixture model (gray scale since mean for each pixel varies)



Using single multivariate Bernoulli and maximum likelihood amounts to averaging counts in each pixel

Bayesian EM for Discrete Case

- Conjugate prior of the parameters of Bernoulli is given by the beta distribution
- Beta prior is equivalent to introducing additional effective observations of $\mathbf{x}$
- Also introduce priors into the Bernoulli mixture model
- Use EM to maximize posterior probability of distribution
- Can be extended to multinomial discrete variables
  - Introduce Dirichlet priors over model parameters if desired