

# CSE 6740 Lecture 4

## *How Do I Learn a Mixture of Gaussians? (Parametric Estimation)*

Alexander Gray

agray@cc.gatech.edu

Georgia Institute of Technology

# Today

Today we'll look at the parts of one machine learning problem, learning a mixture of Gaussians:

**Task:** density estimation

**Model class:** set of all possible *mixtures of Gaussians* with  $K$  components

**Loss:** *likelihood*

**Optimizer:** *EM algorithm*

**Generalization mechanism:** *cross-validation*

# Today

1. Mixture of Gaussians and graphical models (*What's an example of a machine learning model?*)
2. The likelihood (*An important loss function*)
3. The EM algorithm (*What's an example of an optimizer?*)
4. Generalization and model selection (*How do we minimize future error?*)

# Mixture of Gaussians and graphical models

What's an example of a more complex model?

# Gaussian Distribution

Recall the univariate Gaussian:  $X \sim \mathcal{N}(\mu, \sigma^2)$  if

$$f(x) = \mathcal{N}(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}(x - \mu)^2\right\} \quad (1)$$

for  $\mu \in \mathbb{R}, \sigma > 0$ .

# Multivariate Gaussian Distribution

Recall the univariate Gaussian:  $X \sim \mathcal{N}(\mu, \sigma^2)$  if

$$f(x) = \mathcal{N}(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}(x - \mu)^2\right\} \quad (2)$$

for  $\mu \in \mathbb{R}, \sigma > 0$ .

Multivariate Gaussian:  $X \sim \mathcal{N}(\mu, \Sigma)$  if

$$f(x) = \frac{1}{|\Sigma|^{1/2}(2\pi)^{D/2}} \exp\left\{-\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu)\right\} \quad (3)$$

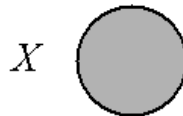
for  $\mu \in \mathbb{R}^D, \Sigma$  a  $D \times D$  symmetric, positive definite matrix.  
Then  $\mathbb{E}(X) = \mu$  and  $\mathbb{V}(X) = \Sigma$ .

# Graphical models

*Graphical models* refers to a graph-based representation of statistical models.

Circles denote continuous-valued random variables, squares denote discrete rv's, clear means hidden, and shaded means observed.

$$f(x) = f_{\Theta}(x) = f(x; \Theta) = f(x; \mu, \Sigma) = \mathcal{N}(\mu, \Sigma) \quad (4)$$



# Mixtures of Gaussians

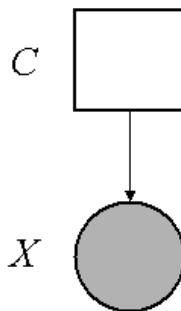
What if the underlying density is something more complicated? We could consider a *mixture of Gaussians* with  $K$  components:

$$P(C = k) = \pi_k \quad (5)$$

$$f(X|C = k) = \mathcal{N}(\mu_k, \sigma_k^2) \quad (6)$$

$$f(X) = \sum_{k=1}^K f(X|C = k)P(C = k) = \sum_{k=1}^K \pi_k \mathcal{N}(\mu_k, \sigma_k^2) \quad (7)$$

where  $\sum_k \pi_k = 1$



# The Likelihood

An important loss function.

# Parametric Estimation

We assume a model class  $\mathcal{F}$ , such as set of all possible Gaussians, which has a *parameter space*

$\Theta = \{(\mu, \sigma) : \mu \in \mathbb{R}, \sigma > 0\}$ . We will call this *parametric estimation* since there is a finite number of parameters.

Often we are only interested in some function  $T(\theta)$ . For example, the mean, or some function of it. Then  $\mu$  is the *parameter of interest* and  $\sigma$  is called a *nuisance parameter*.

Many ways exist to create an estimator for a model. One is called the method of moments. But the most popular is *maximum likelihood estimation*.

# The Likelihood Function

Let  $X_1, \dots, X_N$  be IID with PDF  $f(x|\theta)$  or  $f(x; \theta)$ . The *likelihood function* is defined by  $L_N(\theta|x)$  or  $L_N(\theta; x)$  or

$$L_N(\theta) = \prod_{i=1}^N f(X_i; \theta), \quad (8)$$

or  $L(\theta|x) = f(x|\theta)$ .

The likelihood function is just the joint density of the data, except that we treat it as a function of the parameter  $\theta$ ,  $L_N : \Theta \mapsto [0, \infty)$ .

It is not a density function in general; it does not necessarily integrate to 1 with respect to  $\theta$ .

# Using/Interpreting the Likelihood

If  $X$  is discrete, then  $L(\theta|x) = \mathbb{P}_\theta(X = x)$ . If we compare the likelihood at two parameter values  $\theta_1$  and  $\theta_2$  and find that

$$\mathbb{P}_{\theta_1}(X = x) = L(\theta_1|x) > L(\theta_2|x) = \mathbb{P}_{\theta_2}(X = x) \quad (9)$$

then the sample we actually observed is more likely to have occurred if  $\theta = \theta_1$  than if  $\theta = \theta_2$ . This can be interpreted as saying that  $\theta_1$  is a more plausible guess than  $\theta_2$ .

For continuous  $X$ , we have

$$\frac{\mathbb{P}_{\theta_1}(x - \epsilon < X < x + \epsilon)}{\mathbb{P}_{\theta_2}(x - \epsilon < X < x + \epsilon)} \approx \frac{L(\theta_1|x)}{L(\theta_2|x)}. \quad (10)$$

However, in general we don't interpret likelihoods as probabilities for  $\theta$ .

# Maximum Likelihood Estimator

The *maximum likelihood estimate* (MLE)  $\hat{\theta}_N$ , is the value of  $\theta$  that maximizes  $L_N(\theta)$ .

The *log-likelihood function* is defined by  $l_N(\theta) = \log L_N(\theta)$ . Its maximum occurs at the same place as that of the likelihood function.

The same is true of the likelihood function times any constant. Thus we shall often drop constants in the likelihood function.

The log-likelihood is also sometimes called the *cross-entropy* or *deviance* in the context of classification.

# Likelihood as Loss Function

We can think of maximizing the likelihood as also fitting in our framework of minimizing loss (error), since maximizing the likelihood

$$L_N(\theta) = \prod_{i=1}^N f(X_i; \theta), \quad (11)$$

is equivalent to minimizing

$$-l_N(\theta) = \sum_{i=1}^N \log f(X_i; \theta). \quad (12)$$

Thus, the likelihood is a loss function summed over the data, like any other. Maximizing the likelihood is also asymptotically equivalent to minimizing a quantity called the *Kullback-Liebler divergence*, to be described.

# Gaussian Example

Let  $X_1, \dots, X_N \sim \mathcal{N}(\mu, \sigma^2)$ . The parameter is  $\theta = (\mu, \sigma)$  and the likelihood function (ignoring some constants) is

$$L_N(\mu, \sigma) = \prod_i \frac{1}{\sigma} \exp \left\{ -\frac{1}{2\sigma^2} (X_i - \mu)^2 \right\} \quad (13)$$

$$= \sigma^{-N} \exp \left\{ -\frac{1}{2\sigma^2} \sum_i (X_i - \mu)^2 \right\} \quad (14)$$

$$= \sigma^{-N} \exp \left\{ -\frac{NS^2}{2\sigma^2} \right\} \exp \left\{ -\frac{(\bar{X} - \mu)^2}{2\sigma^2} \right\} \quad (15)$$

where  $\bar{X}$  is the sample mean and  $S^2$  is the sample variance, because

$$\sum_i (X_i - \mu)^2 = \sum_i (X_i - \bar{X} + \bar{X} - \mu)^2 = NS^2 + \mathcal{N}(\bar{X} - \mu)^2.$$

# Gaussian Example

The log-likelihood is

$$l(\mu, \sigma) = -N \log \sigma - \frac{NS^2}{2\sigma^2} - \frac{\mathcal{N}(\bar{X} - \mu)^2}{2\sigma^2}. \quad (16)$$

Solving (analytically) the equations

$$\frac{\delta l(\mu, \sigma)}{\delta \mu} = 0 \quad \text{and} \quad \frac{\delta l(\mu, \sigma)}{\delta \sigma} = 0, \quad (17)$$

we find that  $\hat{\mu} = \bar{X}$  and  $\hat{\sigma} = S$ . It can be verified that these are global maxima of the likelihood.

# Numerical maximization

Now consider a mixture of Gaussians:

$$f(x) = \sum_{k=1}^K \alpha_k \phi(x; \mu_k, \sigma_k) \quad (18)$$

where  $\sum_k \alpha_k = 1$ .

The MLE can't be found analytically in this case. So we maximize numerically. The EM algorithm (to be described later) is commonly used for likelihood maximization in this kind of model.

Estimators that are the result of maximizing something are called *M-estimators*.

# The EM Algorithm

What's an example of an optimizer?

# Mixture of Gaussians

The *EM algorithm* is an example of an optimizer, for a special case: maximizing the likelihood, for a model with *hidden variables*. In the mixture of Gaussians model, the “hidden” variable is the class label:

$$P(C = k) = \pi_k, \quad \sum_k \pi_k = 1 \quad (19)$$

$$f(X|C = k) = \mathcal{N}(\mu_k, \Sigma_k^2) \quad (20)$$

$$f(X) = \sum_{k=1}^K f(X|C = k)P(C = k) = \sum_{k=1}^K \pi_k \mathcal{N}(\mu_k, \Sigma_k^2) \quad (21)$$

# Mixture of Gaussians

Recall Bayes rule, which gives

$$P(C = k|x) = \frac{f(x|C = k)P(C = k)}{f(x)}. \quad (22)$$

This value is the probability that a particular component  $k$  was responsible for generating the point  $x$ , and satisfies  $\sum_{k=1}^K P(C = k|x) = 1$ . We'll use as a shorthand

$$w_{ik} \equiv P(C = k|x_i). \quad (23)$$

# Mixture of Gaussians

We'll consider a simplified case where the covariances are fixed to be diagonal with all dimensions equal,  $\Sigma_k = \sigma_k^2 I$ , so

$$f(x|C = k) = \mathcal{N}(\mu_k, \Sigma_k) = \frac{1}{(2\pi\sigma_k^2)^{D/2}} \exp \left\{ -\frac{\|x - \mu_k\|^2}{2\sigma_k^2} \right\} \quad (24)$$

and

$$f(x) = \sum_{k=1}^K \pi_k \frac{1}{(2\pi\sigma_k^2)^{D/2}} \exp \left\{ -\frac{\|x - \mu_k\|^2}{2\sigma_k^2} \right\}. \quad (25)$$

# Maximum Likelihood, Identifiability

We want to find the parameters  $\theta = \{\pi_k, \mu_k, \sigma_k\}$  which maximize the likelihood

$$L(\theta) = \prod_{i=1}^N f_{\theta}(X_i), \quad (26)$$

or  $L(\theta|x) = f(x|\theta)$ .

Unfortunately there exist parameter settings for which the likelihood goes to infinity, for example when one of the Gaussian components collapses onto one of the data points. Also there may be several parameter settings with identical likelihoods. We'll just ignore all this and proceed, because it turns out to be fine in practice.

# Minimizing the Negative Log-likelihood

It is equivalent to minimize the negative log-likelihood

$$E \equiv -\log L(\theta) = -\sum_{i=1}^N \log f_{\theta}(X_i) \quad (27)$$

$$= -\sum_{i=1}^N \log \left( \sum_{k=1}^K f(X_i|C = k)P(C = k) \right) \quad (28)$$

Since this error function is a smooth differentiable function of the parameters, we can employ its derivatives to perform unconstrained optimization on it.

# Minimizing the Negative Log-likelihood

For the centers  $\mu_k$  we obtain

$$\frac{\partial E}{\partial \mu_k} = \sum_{i=1}^N w_{ik} \frac{(\mu_k - x_i)}{\sigma_k^2} \quad (29)$$

and for the  $\sigma_k$  we obtain

$$\frac{\partial E}{\partial \sigma_k} = \sum_{i=1}^N w_{ik} \left( \frac{D}{\sigma_k} - \frac{\|x_i - \mu_k\|^2}{\sigma_k^3} \right). \quad (30)$$

# Minimizing the Negative Log-likelihood

Optimizing for the mixing parameters  $\pi_k$  must be done subject to the constraints

$$\sum_{k=1}^K \pi_k = 1, \quad (31)$$

$$0 \leq \pi_k \leq 1. \quad (32)$$

# Minimizing the Negative Log-likelihood

This can be done by representing the mixing parameters in terms of a set of auxiliary variables  $\gamma_k$  such that

$$\pi_k = \frac{\exp(\gamma_k)}{\sum_{k=1}^K \exp(\gamma_k)}. \quad (33)$$

This is called the *logistic* or *softmax* transformation. It ensures for  $-\infty \leq \gamma_k \leq \infty$  that the constraints on  $\pi_k$  hold.

# Minimizing the Negative Log-likelihood

Utilizing

$$\frac{\partial \pi_j}{\partial \gamma_k} = \pi_k I(j = k) - \pi_k \pi_j \quad (34)$$

and the chain rule consequence

$$\frac{\partial E}{\partial \gamma_k} = \sum_{j=1}^K \frac{\partial E}{\partial \pi_j} \frac{\partial \pi_j}{\partial \pi_k} \quad (35)$$

we can obtain

$$\frac{\partial E}{\partial \gamma_k} = - \sum_{i=1}^N (w_{ik} - \pi_k). \quad (36)$$

# Minimizing the Negative Log-likelihood

Note that, at this point we are armed with derivatives, so we can use standard optimizers.

The EM algorithm does not actually require these derivatives in its final form (though we will consider these derivatives in our derivation which gets us there).

# Conditions at the Optimum

It is insightful to see what the maximum likelihood solutions look like; when these derivatives are zero we have

$$\hat{\mu}_k = \frac{\sum_i w_{ik} x_i}{\sum_i w_{ik}} \quad (37)$$

$$\hat{\sigma}_k^2 = \frac{1}{D} \frac{\sum_i w_{ik} \|x_i - \mu_k\|^2}{\sum_i w_{ik}} \quad (38)$$

$$\hat{\pi}_k = \frac{1}{N} \sum_i w_{ik} \quad (39)$$

which represents the intuitively satisfying result that they are the usual mean and standard deviation where the points are weighted by the posterior probabilities of being generated by each component.

# EM: Recurrence Idea

The previous equations can be interpreted as a recurrence, since parameter values pop out on the left-hand side, and parameter values underlie the  $w_{ik}$  on the right-hand side. We can imagine an alternating scheme: on each iteration “new” values are computed based on the “old” values from the last iteration.

There is a more formal derivation which we will show in a later lecture, based on the idea of *bound optimization* or *majorization*.

# EM Algorithm for Mixture of Gaussians

We arrive at the following update equations:

$$\mu_k^{\text{new}} = \frac{\sum_i w_{ik}^{\text{old}} x_i}{\sum_i w_{ik}^{\text{old}}} \quad (40)$$

$$(\sigma_k^2)^{\text{new}} = \frac{1}{D} \frac{\sum_i w_{ik}^{\text{old}} \|x_i - \mu_k^{\text{new}}\|^2}{\sum_i w_{ik}^{\text{old}}} \quad (41)$$

$$\pi_k^{\text{new}} = \frac{1}{N} \sum_i w_{ik}^{\text{old}}. \quad (42)$$

# EM Algorithm Steps

These comprise the *Expectation-Maximization* (EM) algorithm for the case of mixtures of Gaussians. The “expectation step” (*E-step*) consists of evaluating the conditional probabilities  $w_{ik}$  using the last values of the parameters. The “maximization step” (*M-step*) consists of the updates to the parameters which move toward the local maximum.