15-388/688 - Practical Data Science: Anomaly detection and mixture of Gaussians

J. Zico Kolter
Carnegie Mellon University
Fall 2016

Announcements

Tutorial due tonight (max 2 late days, pushing it until Friday)

HW4 to be released tonight

Feedback on project proposals will be out tomorrow

What is an “anomaly”

Two views of anomaly detection

Supervised view: anomalies are what some user labels as anomalies

Unsupervised view: anomalies are outliers (points of low probability) in the data

In reality, you want a combination of both these viewpoints: not all outliers are anomalies, but all anomalies should be outliers

This lecture is going to focus on the unsupervised view, but this is only part of the full equation

What is an outlier?

Outliers are points of low probability

Given a collection of data points $x^{(1)}, \ldots, x^{(m)}$, describe the points using some distribution, then find points with lowest $p(x^{(i)})$

Since we are considering points with no labels, this is an *unsupervised* learning algorithm (could formulate in terms of hypothesis, loss, optimization, but instead for this lecture we'll be focusing on the probabilistic notation)

Multivariate Gaussian distributions

We have seen Gaussian distributions previously, but mainly focused on distributions over scalar-valued data $x^{(i)} \in \mathbb{R}$

$$p(x; \mu, \sigma^2) = (2\pi\sigma^2)^{-1/2} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Gaussian distributions generalize nicely to distributions over vector-valued random variables X taking values in $\mathbb{R}^n$

$$\begin{aligned} p(x; \mu, \Sigma) &= |2\pi\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right) \\ &\equiv \mathcal{N}(x; \mu, \Sigma) \end{aligned}$$

with parameters $\mu \in \mathbb{R}^n$ and $\Sigma \in \mathbb{R}^{n \times n}$, and were $|\cdot|$ denotes the determinant of a matrix (also written $X \sim \mathcal{N}(\mu, \Sigma)$)

Properties of multivariate Gaussians

Mean and variance

$$\mathbf{E}[X] = \int_{\mathbb{R}^n} x\mathcal{N}(x;\mu,\Sigma)dx = \mu$$

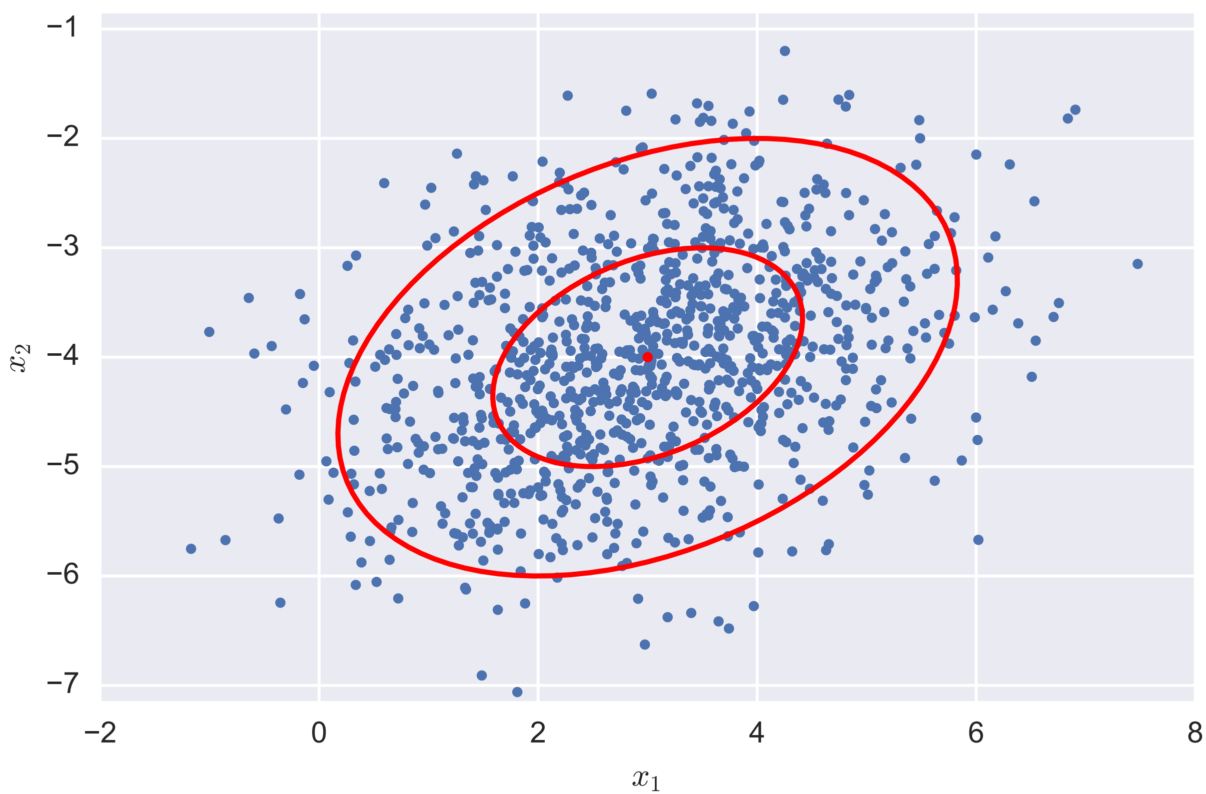
$$\mathbf{Cov}[X] = \int_{\mathbb{R}^n} (x-\mu)(x-\mu)^T\mathcal{N}(x;\mu,\Sigma)dx = \Sigma$$

(these are *not obvious)*

Creation from univariate Gaussians: for $x \in \mathbb{R}$, if $p(x_i) = \mathcal{N}(x;0,1)$ (i.e., each element x_i is an independent univariate Gaussian, then $y = Ax + b$ is also normal, with distribution

$$Y \sim \mathcal{N}(\mu = b, \Sigma = AA^T)$$

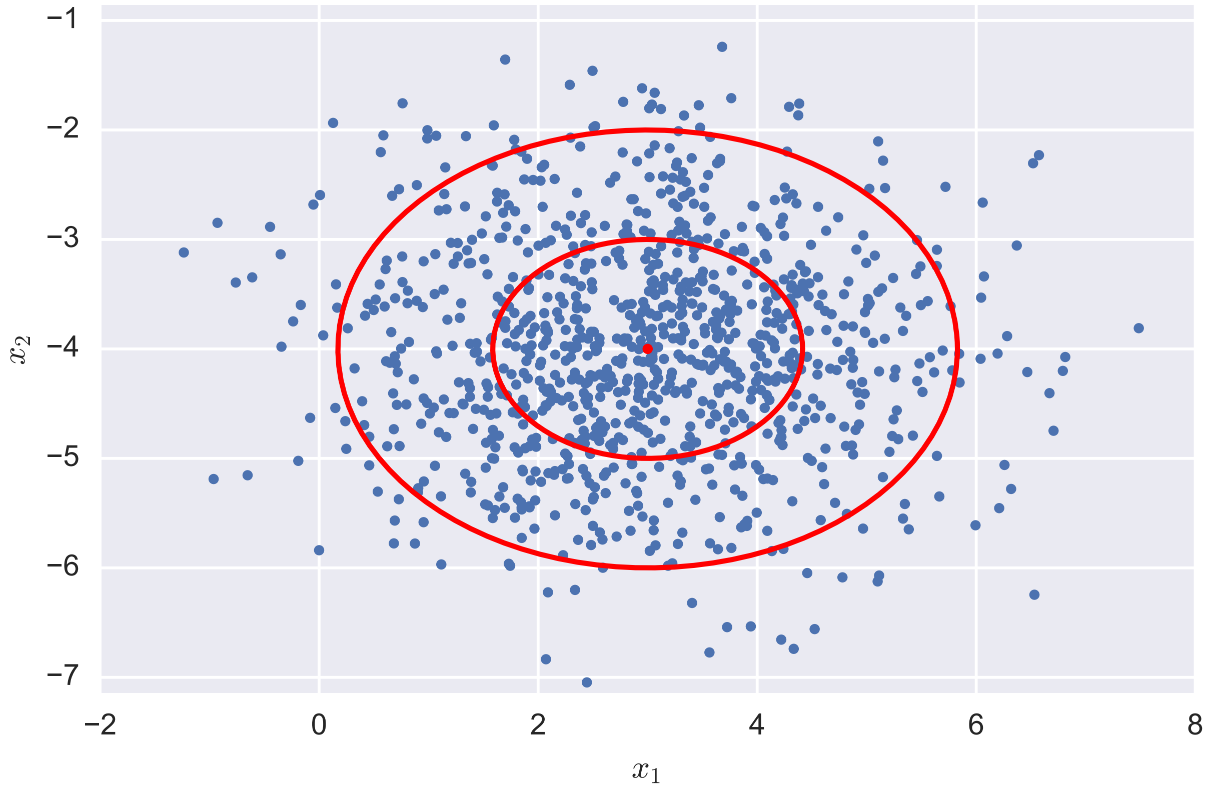
Multivariate Gaussians, graphically



$$\mu = \begin{bmatrix} 3 \\ -4 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} 2.0 & 0.5 \\ 0.5 & 1.0 \end{bmatrix}$$

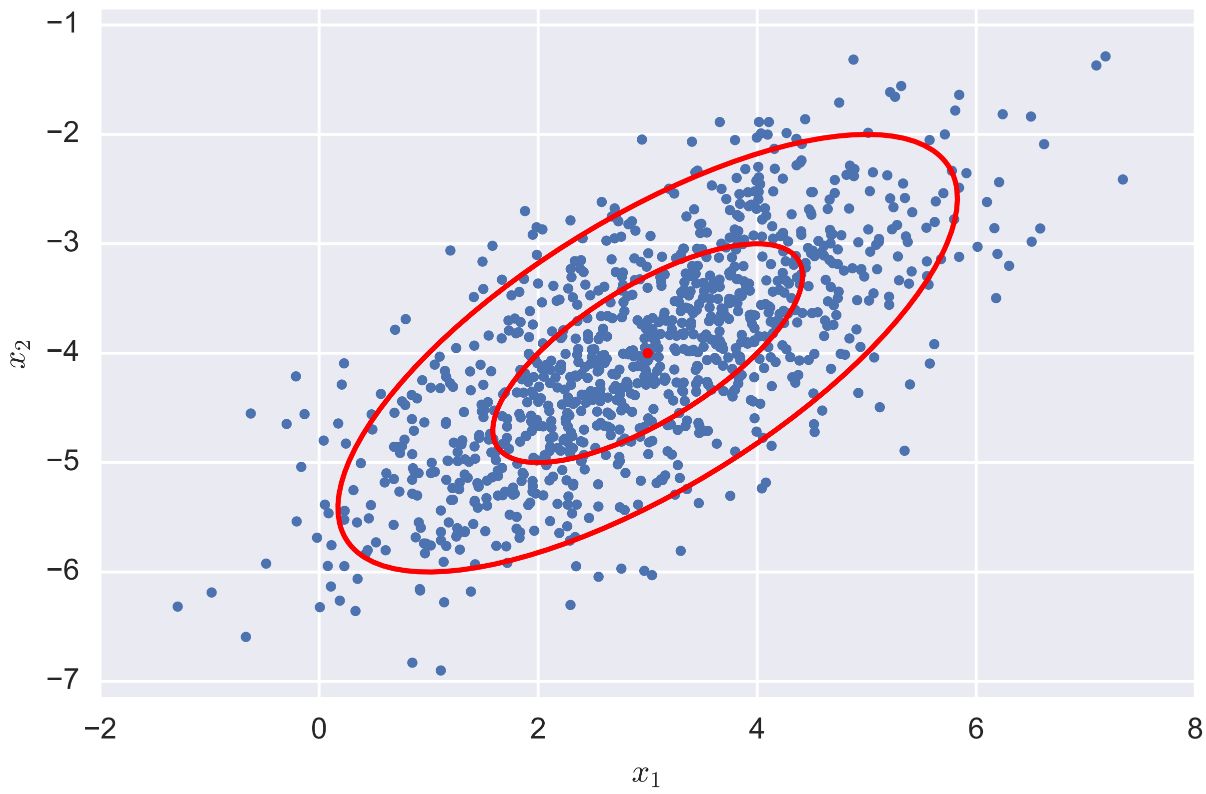
Multivariate Gaussians, graphically



$$\mu = \begin{bmatrix} 3 \\ -4 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} 2.0 & 0 \\ 0 & 1.0 \end{bmatrix}$$

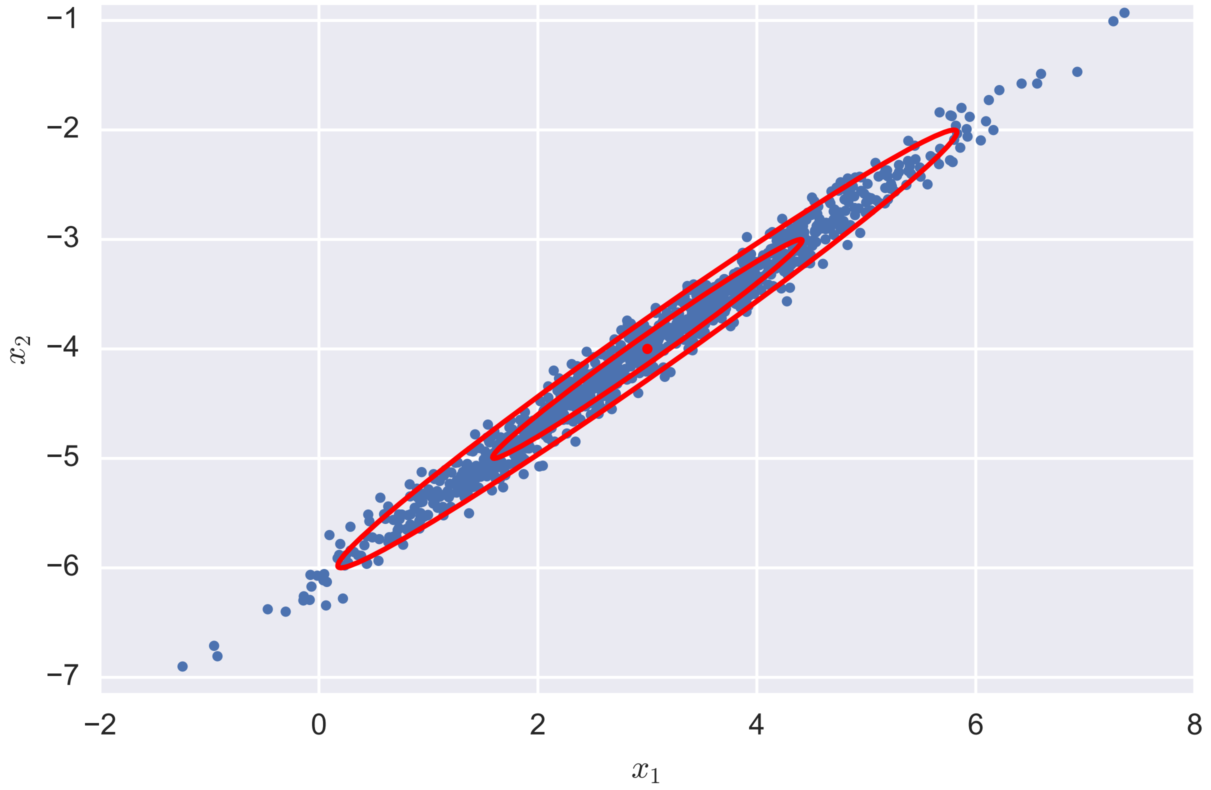
Multivariate Gaussians, graphically



$$\mu = \begin{bmatrix} 3 \\ -4 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} 2.0 & 1.0 \\ 1.0 & 1.0 \end{bmatrix}$$

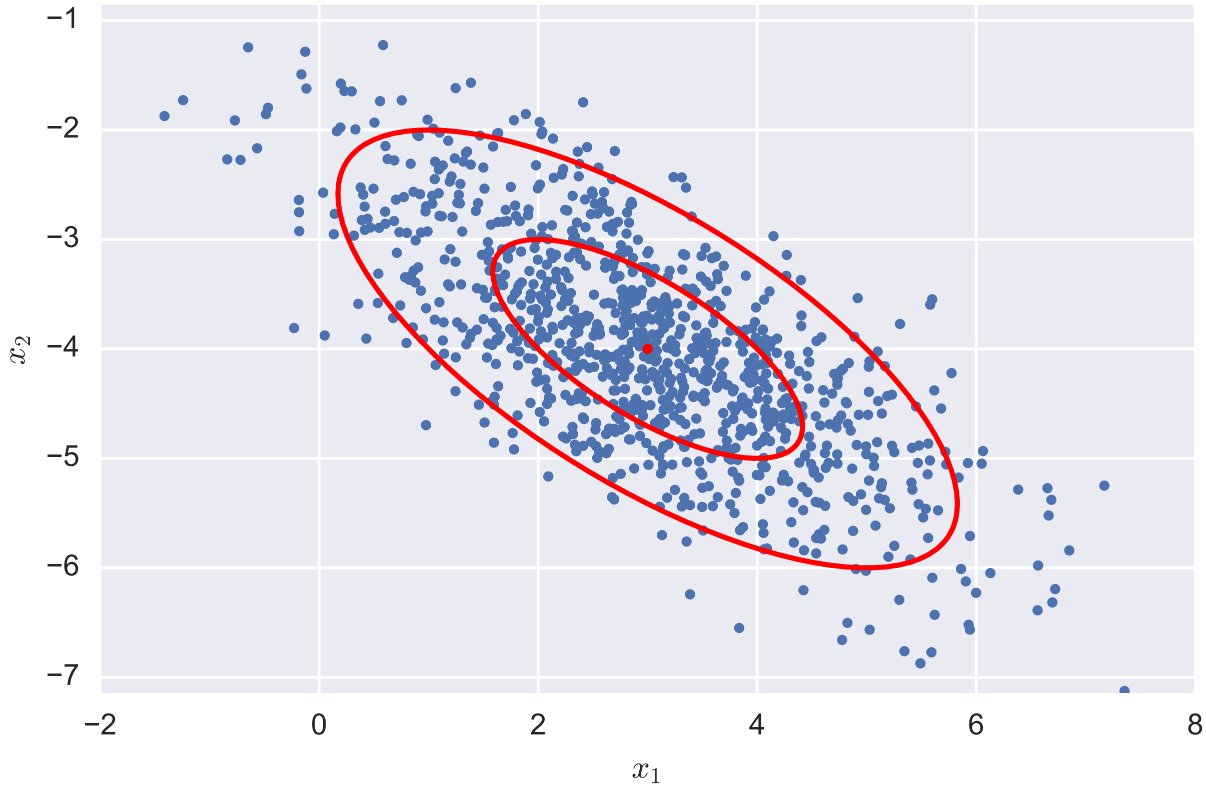
Multivariate Gaussians, graphically



$$\mu = \begin{bmatrix} 3 \\ -4 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} 2.0 & 1.4 \\ 1.4 & 1.0 \end{bmatrix}$$

Multivariate Gaussians, graphically



$$\mu = \begin{bmatrix} 3 \\ -4 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} 2.0 & -1.0 \\ -1.0 & 1.0 \end{bmatrix}$$

Maximum likelihood estimation

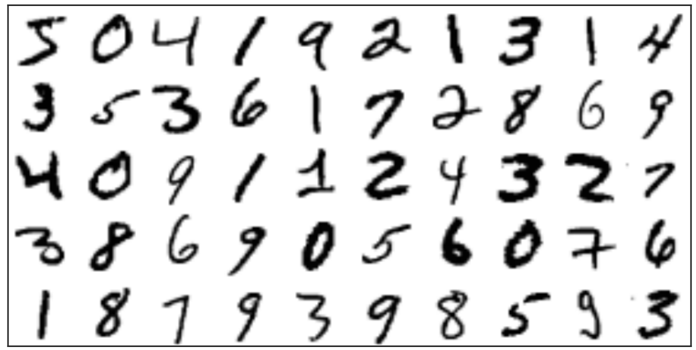
The maximum likelihood estimate of μ, Σ are what you would “expect”, but derivation is non-obvious

$$\begin{aligned} \underset{\mu,\Sigma}{\text{minimize}} \;\; \ell(\mu,\Sigma) &= \sum_{i=1}^{m} \log p(x^{(i)};\mu,\Sigma) \\ &= \sum_{i=1}^{m} \left(-\frac{1}{2}\log|2\pi\Sigma| - \frac{1}{2}(x^{(i)}-\mu)^T\Sigma^{-1}(x^{(i)}-\mu) \right) \end{aligned}$$

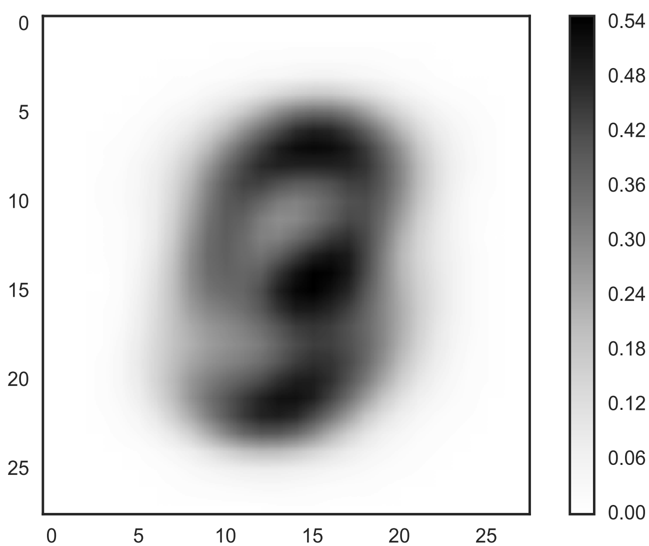
Taking gradients with respect to μ and Σ and setting equal to zero give the closed-form solutions

$$\mu = \frac{1}{m}\sum_{i=1}^{m} x^{(i)}, \qquad \Sigma = \frac{1}{m}\sum_{i=1}^{m}(x^{(i)}-\mu)(x^{(i)}-\mu)^T$$

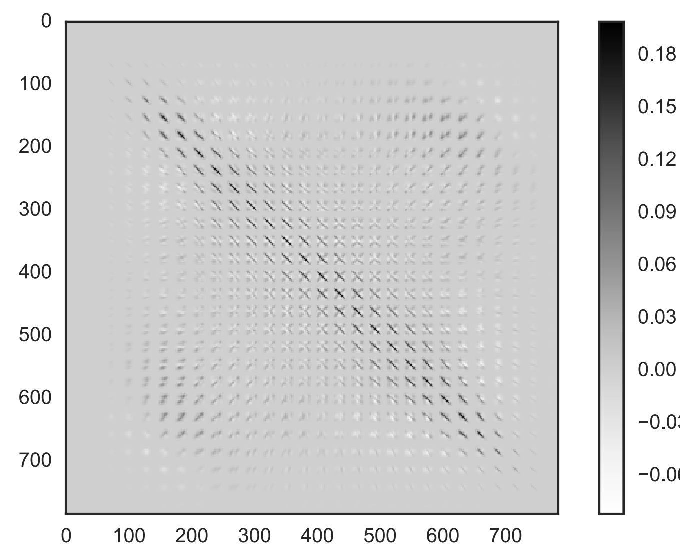
Fitting Gaussian to MNIST



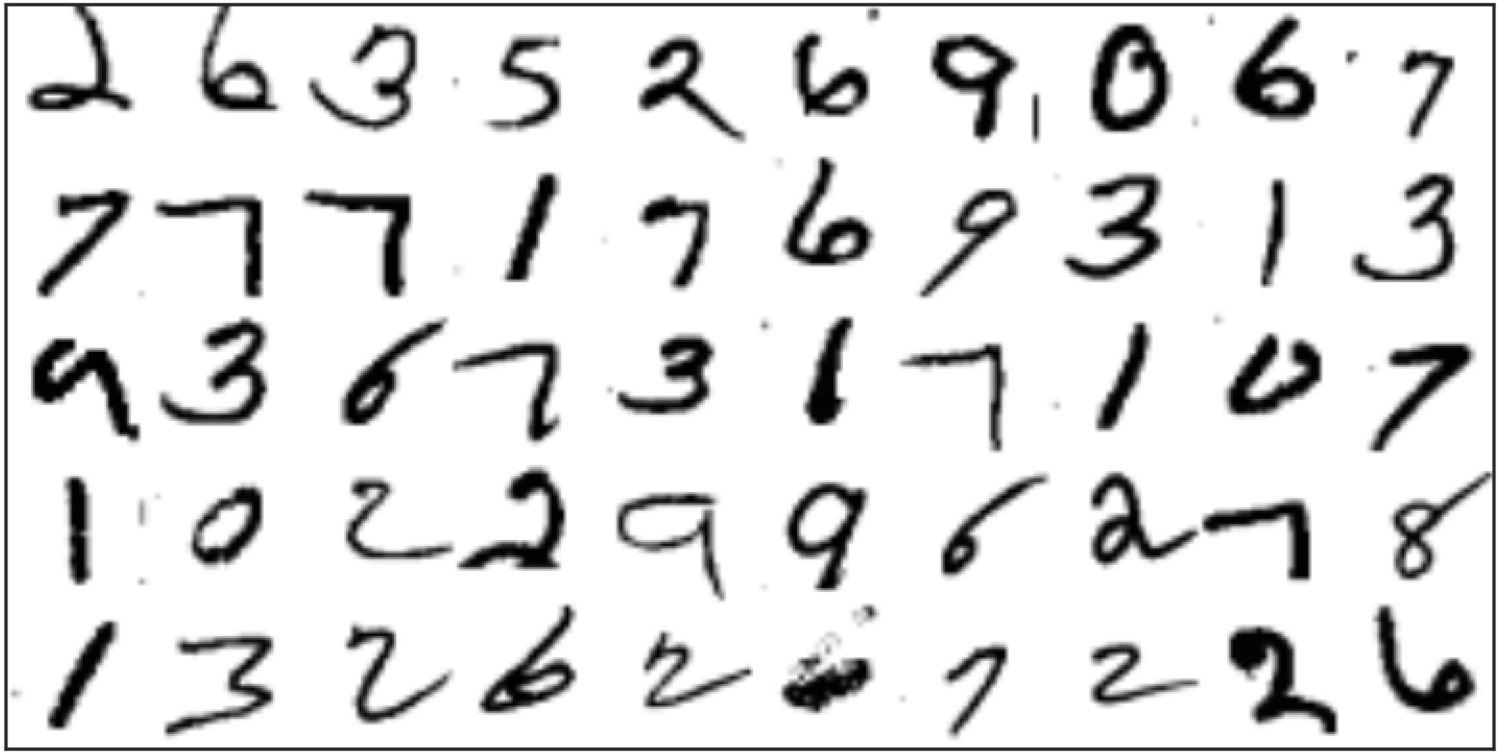
$\mu =$



$\Sigma =$

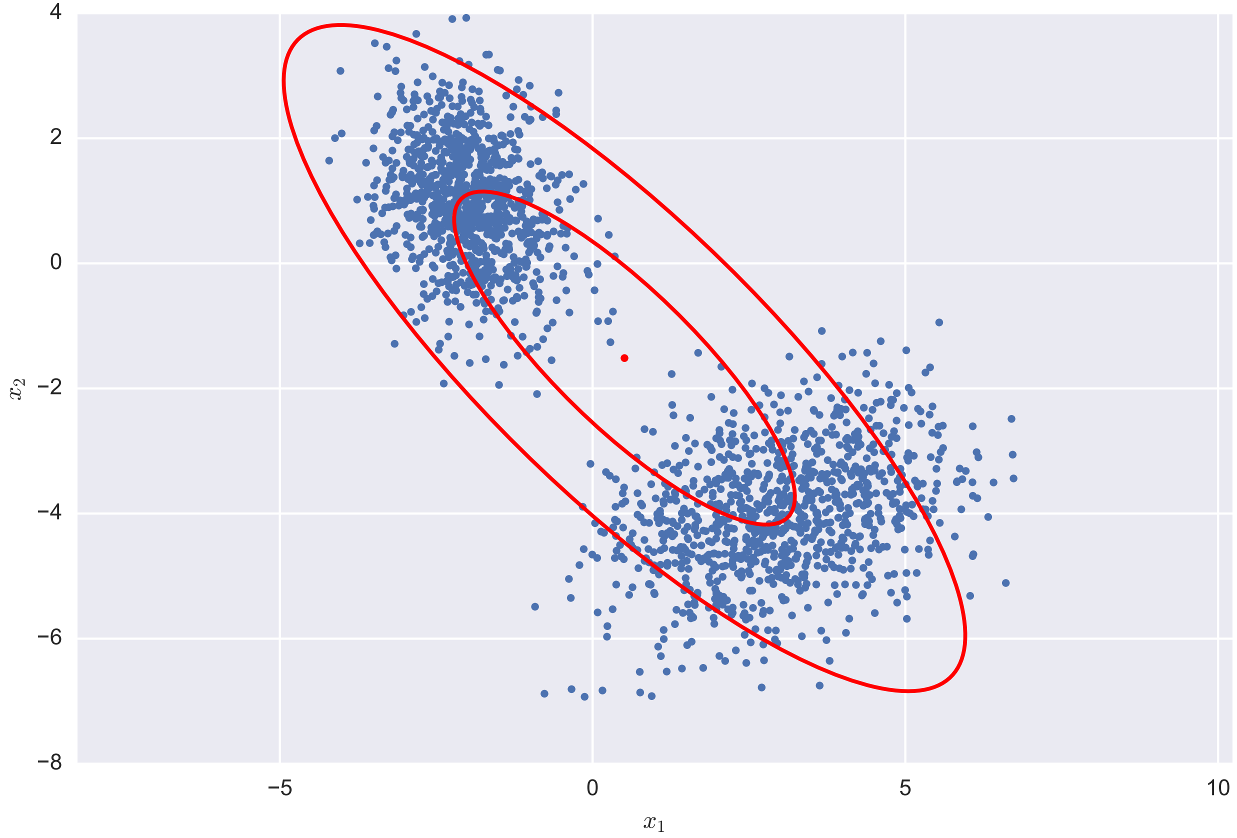


MNIST Outliers



Limits of Gaussians

Though useful, multivariate Gaussians are limited in the types of distributions they can represent



Mixture models

A more powerful model to consider is a *mixture* of Gaussian distributions, a distribution where we first consider a categorical variable

$$Z \sim \text{Categorical}(\phi), \qquad \phi \in [0,1]^k, \sum_i \phi_i = 1$$

i.e., z takes on values $\{1, \ldots, k\}$

For each potential value of Z, we consider a *separate* Gaussian distribution:

$$X|Z = z \sim \mathcal{N}(\mu^{(z)}, \Sigma^{(z)}), \qquad \mu^{(z)} \in \mathbb{R}^n, \Sigma^{(z)} \in \mathbb{R}^{n \times n}$$

Can write the distribution of X using marginalization

$$p(X) = \sum_z p(X|Z = z) p(Z = z) = \sum_z \mathcal{N}(x; \mu^{(z)}, \Sigma^{(z)}) \phi_z$$

Learning mixture models

To estimate parameters, suppose first that we can observe both X and Z, i.e., our data set is of the form $(x^{(i)}, z^{(i)}), i = 1, \ldots, m$

In this case, we can maximize the log-likelihood of the parameters:

$$\ell(\mu, \Sigma, \phi) = \sum_{i=1}^{m} \log p(x^{(i)}, z^{(i)}; \mu, \Sigma, \phi)$$

Without getting into the full details, it hopefully should not be too surprising that the solutions here are given by:

$$\phi_z = \frac{\sum_{i=1}^m \mathbf{1}\{z^{(i)} = z\}}{m}, \qquad \mu^{(z)} = \frac{\sum_{i=1}^m \mathbf{1}\{z^{(i)} = z\} x^{(i)}}{\sum_{i=1}^m \mathbf{1}\{z^{(i)} = z\}},$$

$$\Sigma^{(z)} = \frac{\sum_{i=1}^m \mathbf{1}\{z^{(i)} = z\}(x^{(i)} - \mu^{(z)})(x^{(i)} - \mu^{(z)})^T}{\sum_{i=1}^m \mathbf{1}\{z^{(i)} = z\}}$$

Latent variables and expectation maximization

In the unsupervised setting, $z^{(i)}$ terms will not be known, these are referred to as *hidden* or *latent* random variables

This means that to estimate the parameters, we can't use the function $1\{z^{(i)} = z\}$ anymore

Expectation maximization (EM) algorithm (at a high level): replace indicators $1\{z^{(i)} = z\}$ with probability estimates $p(z^{(i)} = z | x^{(i)}; \mu, \Sigma, \phi)$

When we re-estimate these parameter, probabilities change, so repeat:

E (expectation) step: compute $p(z^{(i)} = z | x^{(i)}; \mu, \Sigma, \phi), \forall i, z$

M (maximization) step: re-estimate μ, Σ, ϕ

EM for Gaussian mixture models

E step: using Bayes' rule, compute probabilities

$$\hat{p}_z^{(i)} \leftarrow p\big(z^{(i)} = z \big| x^{(i)}; \mu, \Sigma, \phi\big)$$
$$= \frac{p\big(x^{(i)} \big| z^{(i)} = z; \mu, \Sigma\big) p\big(z^{(i)} = z; \phi\big)}{\sum_{z'} p\big(x^{(i)} \big| z^{(i)} = z'; \mu, \Sigma\big) p\big(z^{(i)} = z'; \phi\big)}$$
$$= \frac{\mathcal{N}\big(x^{(i)}; \mu^{(z)}, \Sigma^{(z)}\big)\phi_z}{\sum_{z'} \mathcal{N}\big(x^{(i)}; \mu^{(z')}, \Sigma^{(z')}\big)\phi_{z'}}$$

M step: re-estimate parameters using these probabilities

$$\phi_z \leftarrow \frac{\sum_{i=1}^m \hat{p}_z^{(i)}}{m}, \qquad \mu^{(z)} \leftarrow \frac{\sum_{i=1}^m \hat{p}_z^{(i)} x^{(i)}}{\sum_{i=1}^m \hat{p}_{i,z}},$$
$$\Sigma^{(z)} \leftarrow \frac{\sum_{i=1}^m \hat{p}_z^{(i)} \big(x^{(i)} - \mu^{(z)}\big)\big(x^{(i)} - \mu^{(z)}\big)^T}{\sum_{i=1}^m \hat{p}_z^{(i)}}$$

Local optima

Like k-means, EM is effectively optimization a *non-convex* problem

Very real possibility of local optima (seemingly moreso than k-means, in practice)

Same heuristics work as for k-means (in fact, common to initialize EM with clusters from k-means)

Illustration of EM algorithm

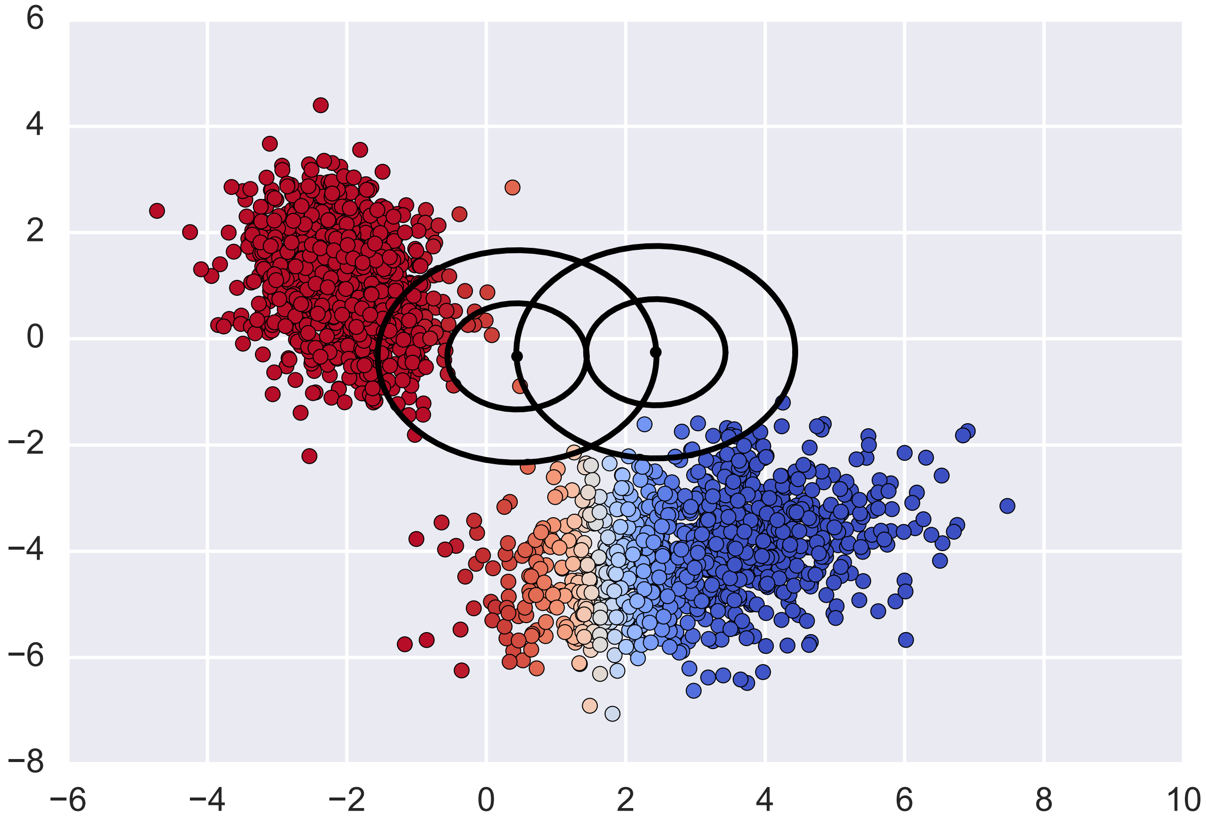


Illustration of EM algorithm

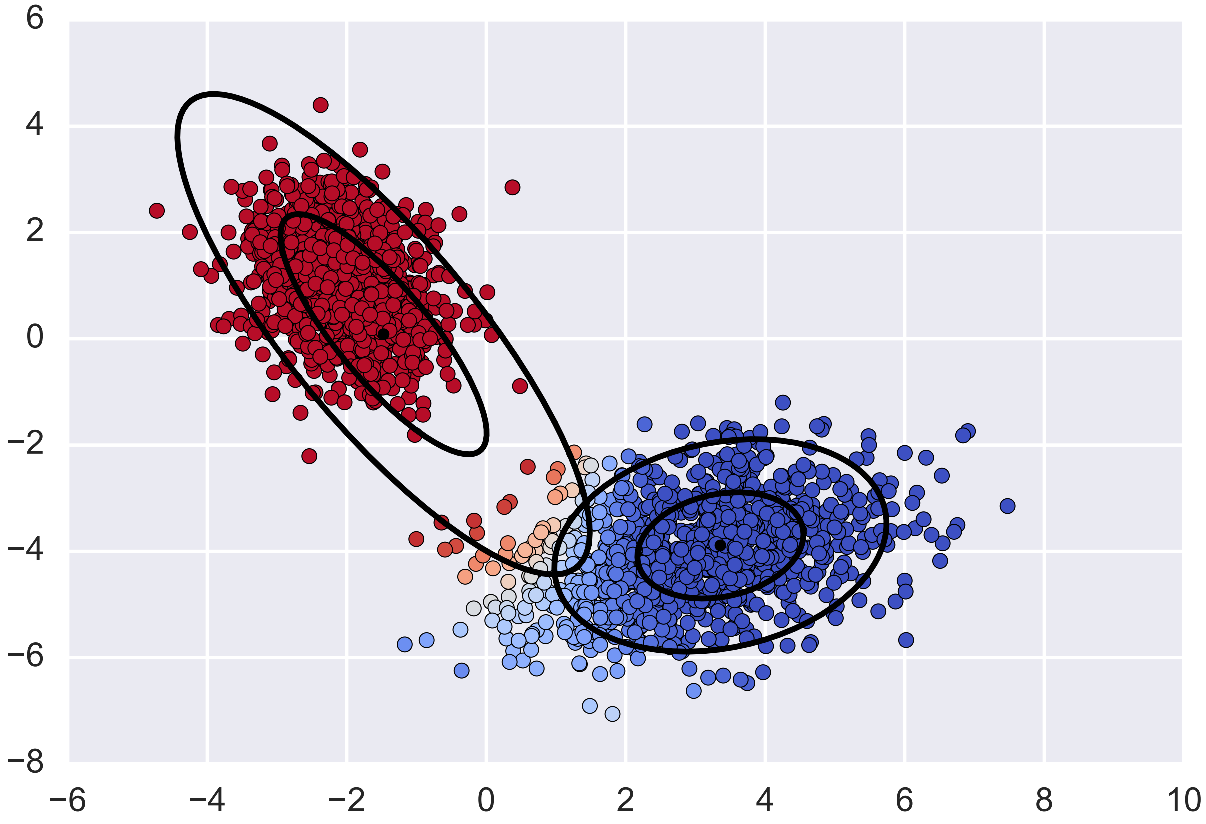


Illustration of EM algorithm

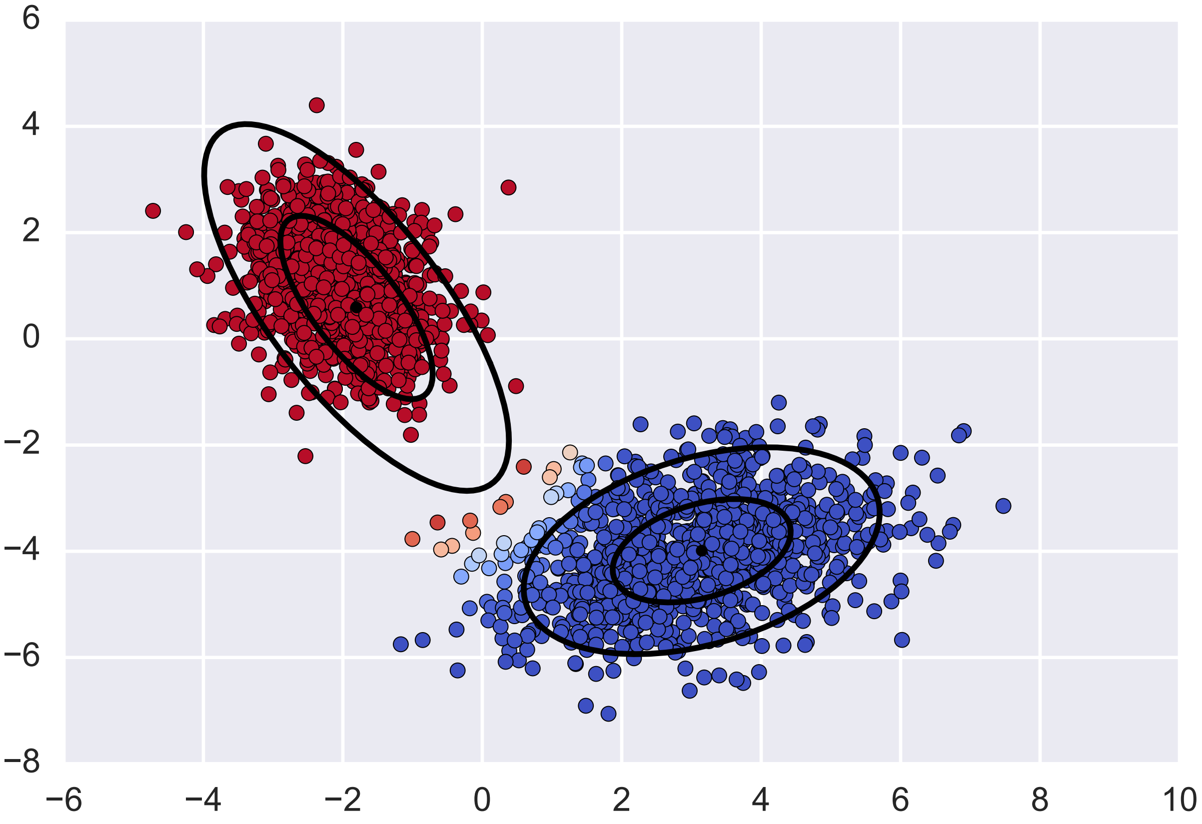


Illustration of EM algorithm

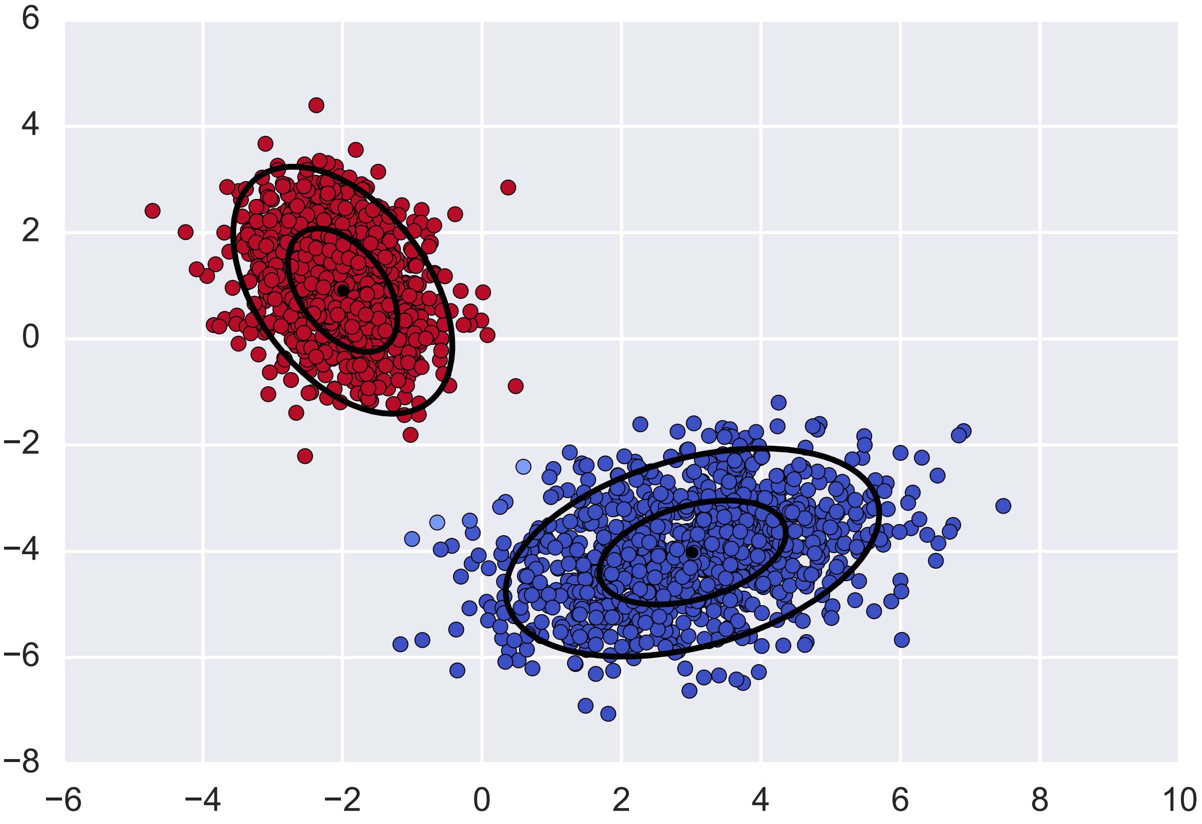
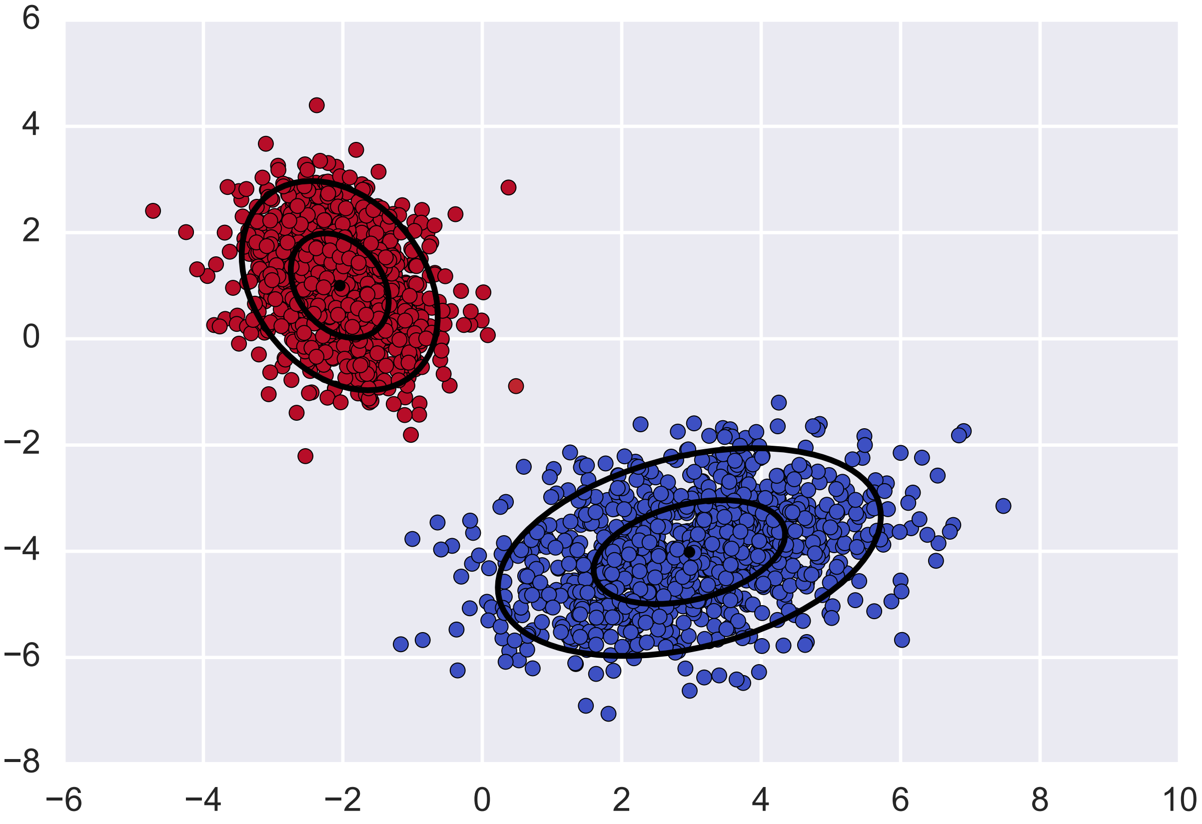
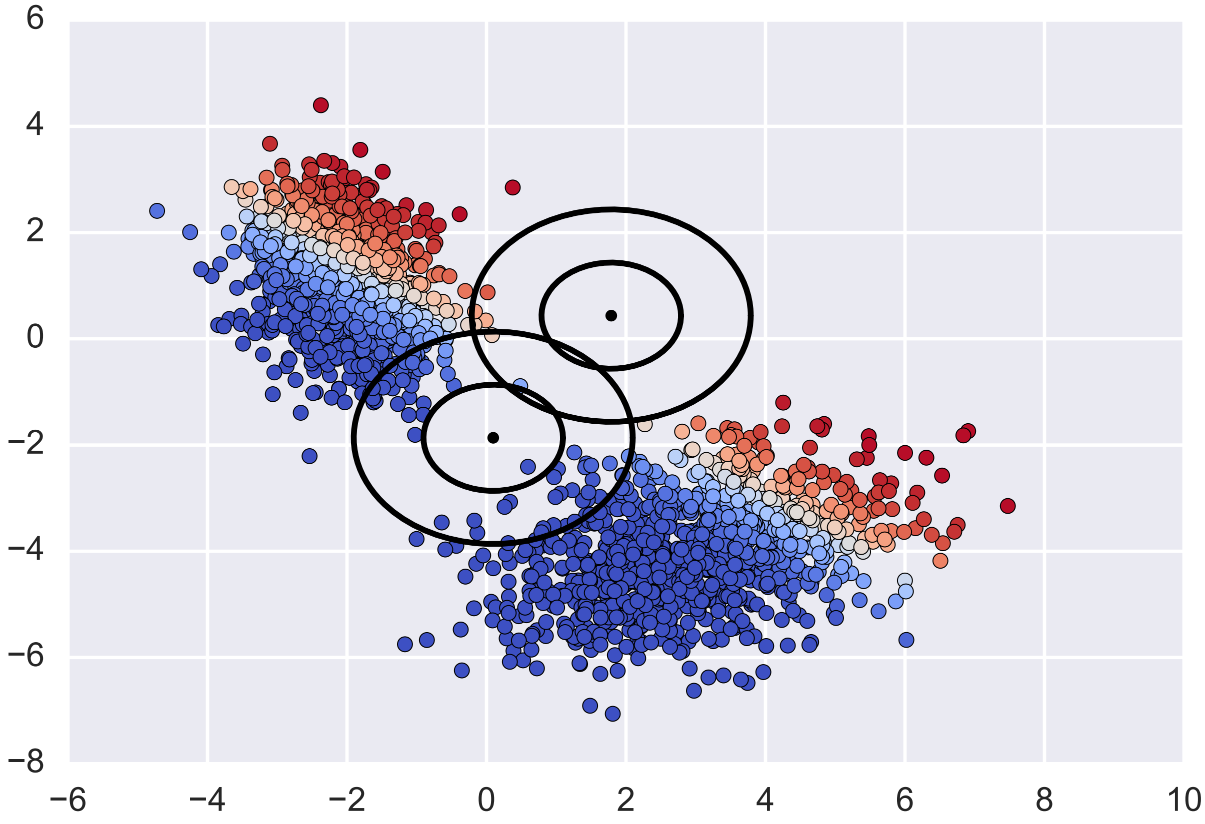


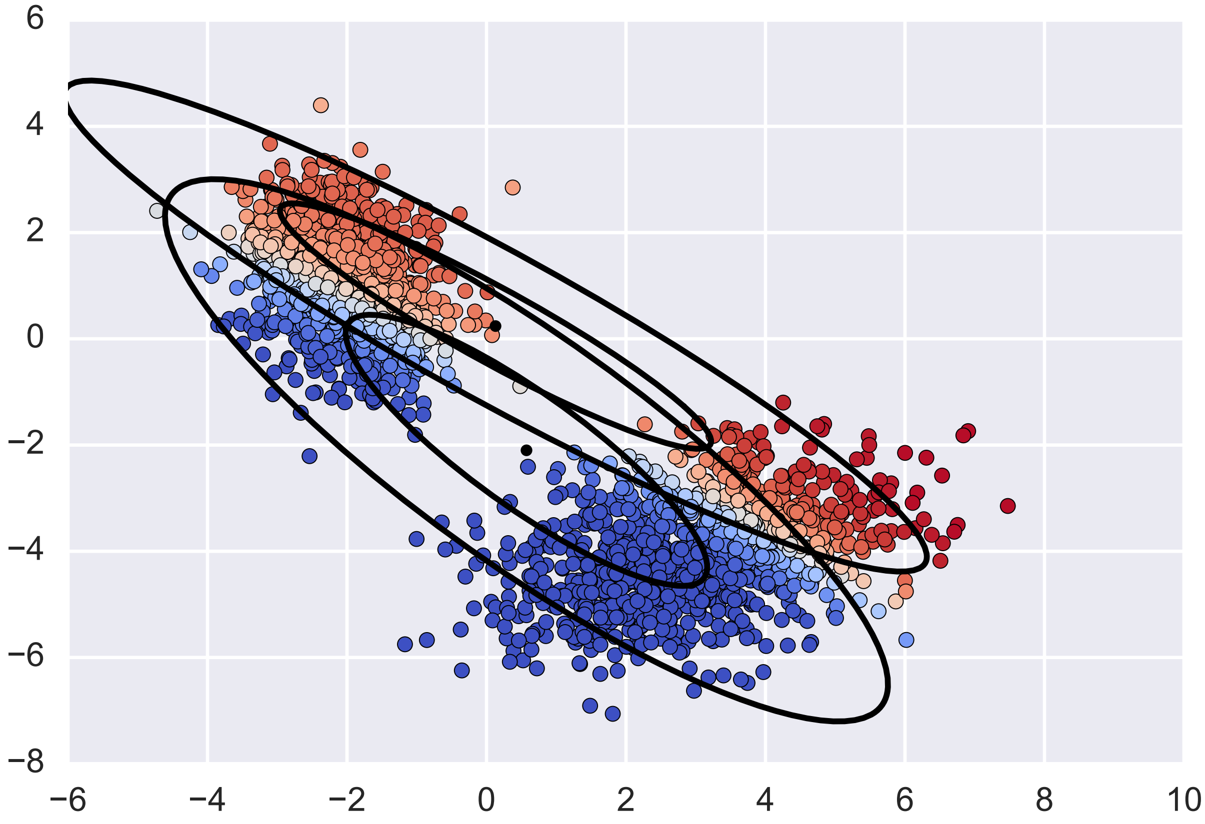
Illustration of EM algorithm



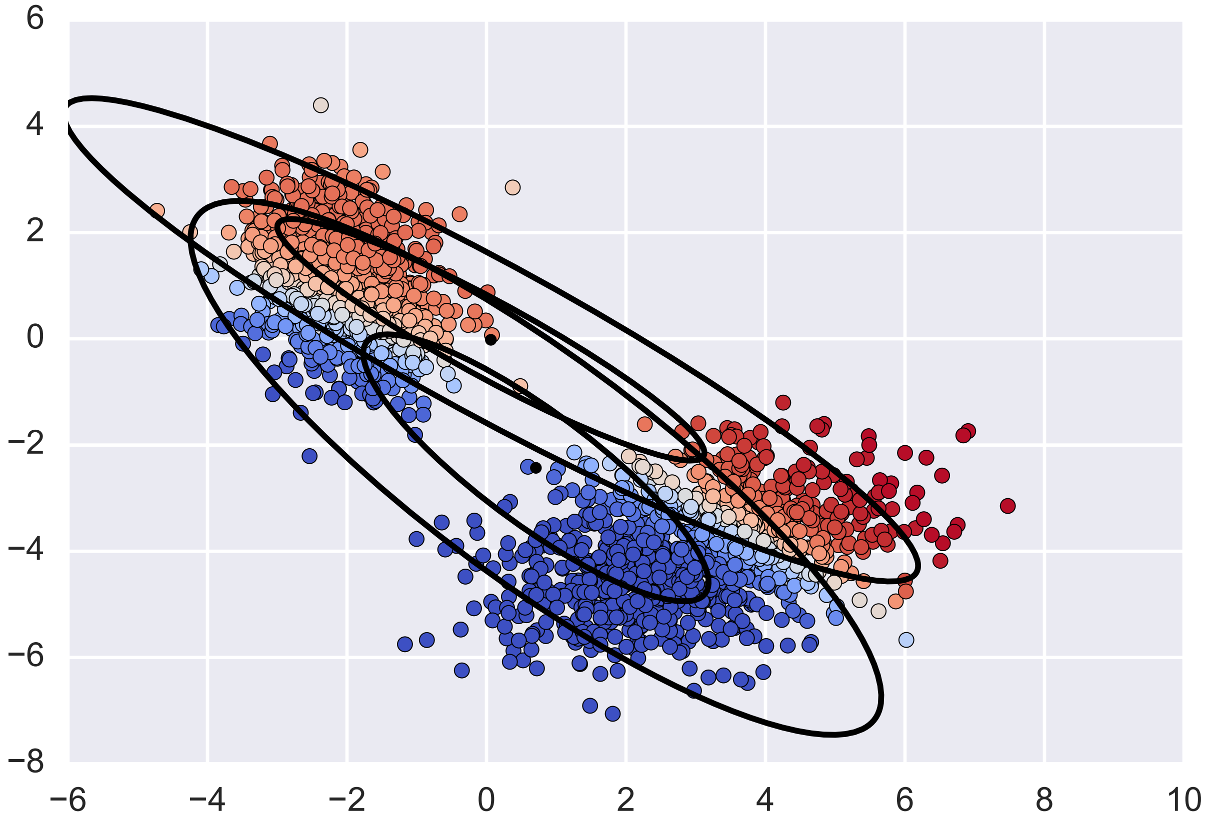
Possibility of local optima



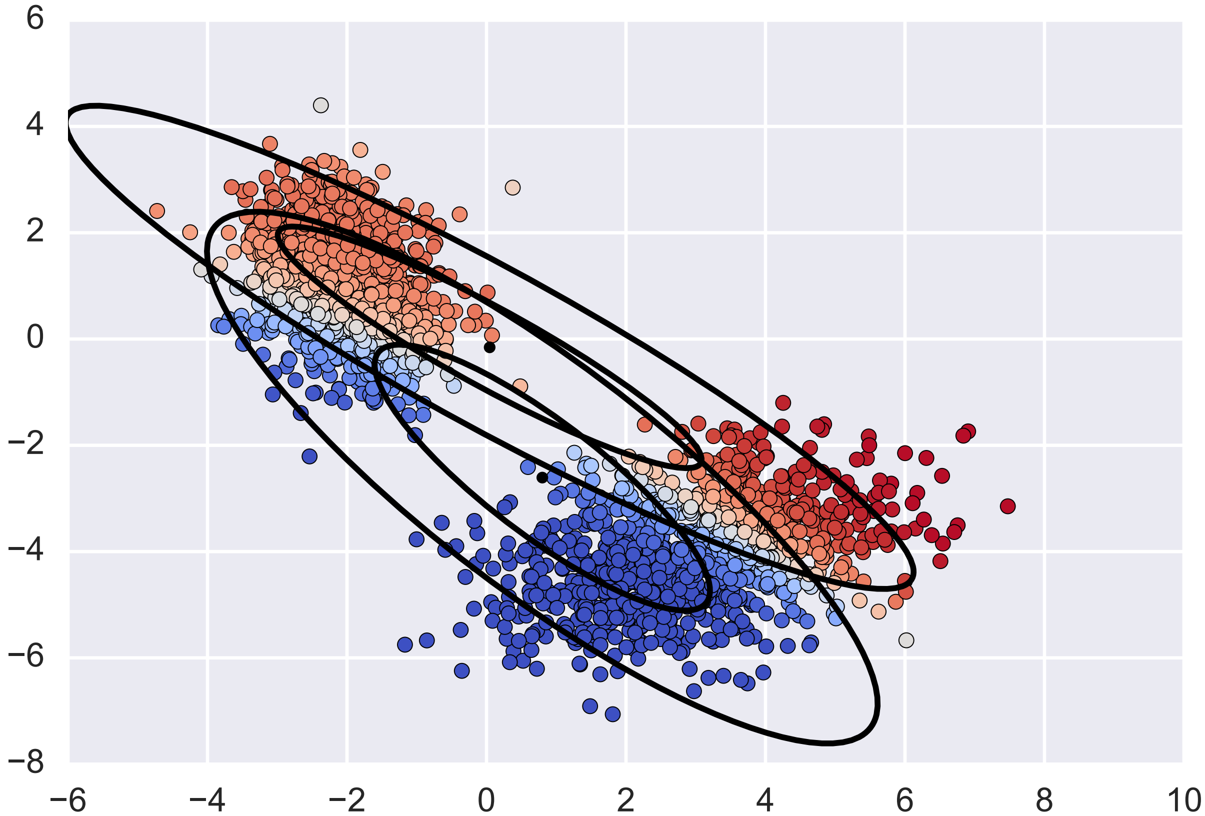
Possibility of local optima



Possibility of local optima



Possibility of local optima



EM and k-means

As you may have noticed, EM for mixture of Gaussians and k-means seem to be doing very similar things

Primary differences: EM is computing “distances” based upon the inverse covariance matrix, allows for “soft” assignments instead of hard assignments