Homework 6

Young Won Kim (yk41) and Minh Pham (mnp7)

Spring 2017

1 EM for mixtures of Bernoullis

A. The M step for ML estimation for a mixture of Bernoullis

$Q(\theta^t, \theta^{(t-1)}) = E[\sum_{i=1}^{m} logP(x^{(i)}, z^{(i)} \mid \theta)]$

$Q(\theta^t, \theta^{(t-1)}) = E[\sum_{i=1}^{m} log[\prod_{k=1}^{K} (\pi_k Ber(x^{(i)} \mid \mu_k))^{I(z^{(i)}=k)}]$

$Q(\theta^t, \theta^{(t-1)}) = \sum_{i=1}^{m} \sum_{k=1}^{K} E(I(z^{(i)} = k))[log\pi_k + logBer(x^{(i)} \mid \mu_k))$

$Q(\theta^t, \theta^{(t-1)}) = \sum_{i=1}^{m} \sum_{k=1}^{K} r_k^{(i)} log\pi_k + \sum_{i=1}^{m} \sum_{k=1}^{K} r_k^{(i)} logBer(x^{(i)} \mid \mu_k)$

$Q(\theta^t, \theta^{(t-1)}) = \sum_{i=1}^{m} \sum_{k=1}^{K} r_k^{(i)} log\pi_k + \sum_{i=1}^{m} \sum_{k=1}^{K} r_k^{(i)} log(\mu_k^{(x^{(i)}}(1-\mu_k)^{(1-x^{(i)})})$

$Q(\theta^t, \theta^{(t-1)}) = \sum_{i=1}^{m} \sum_{k=1}^{K} r_k^{(i)} log\pi_k + \sum_{i=1}^{m} \sum_{k=1}^{K} r_k^{(i)} (x^{(i)} log(\mu_k) + (1-x^{(i)}) log(1-\mu_k))$

$\frac{\partial Q}{\partial \mu_{kj}} = 0 + \sum_{i=1}^{m} \frac{r_k^{(i)} x_j^{(i)}}{\mu_{kj}} - \sum_{i=1}^{m} \frac{r_k^{(i)}(1-x_j^{(i)})}{(1-\mu_{kj})} = 0$

$\sum_{i=1}^{m} \frac{r_k^{(i)} x_j^{(i)}}{\mu_{kj}} = \sum_{i=1}^{m} \frac{r_k^{(i)}(1-x_j^{(i)})}{(1-\mu_{kj})}$

$(1-\mu_{kj}) \sum_{i=1}^{m} r_k^{(i)} x_j^{(i)} = \mu_{kj} \sum_{i=1}^{m} r_k^{(i)} (1-x_j^{(i)})$

$\sum_{i=1}^{m} r_k^{(i)} x_j^{(i)} = \mu_{kj} (\sum_{i=1}^{m} r_k^{(i)} (1-x_j^{(i)}) + \sum_{i=1}^{m} r_k^{(i)} x_j^{(i)})$

$\sum_{i=1}^{m} r_k^{(i)} x_j^{(i)} = \mu_{kj} \sum_{i=1}^{m} r_k^{(i)}$

$\mu_{kj} = \frac{\sum_{i=1}^{m} r_k^{(i)} x_j^{(i)}}{\sum_{i=1}^{m} r_k^{(i)}}$

B. The M step for the MAP estimation for a mixture of Bernoullis

Similarly,

$Q(\theta^t, \theta^{(t-1)}) = \sum_{i=1}^m \sum_{k=1}^K r_k^{(i)} log\pi_k + \sum_{i=1}^m \sum_{k=1}^K r_k^{(i)} logBer(x^{(i)} \mid \mu_{kj}) + log(\mu_{kj}^{(\alpha-1)}(1-\mu_k)^{(\beta-1)})$

$Q(\theta^t, \theta^{(t-1)}) = \sum_{i=1}^m \sum_{k=1}^K r_k^{(i)} log\pi_k + \sum_{i=1}^m \sum_{k=1}^K r_k^{(i)}(x^{(i)} log(\mu_{kj}) + (1-x^{(i)})log(1-\mu_{kj})) + ((\alpha-1)log(\mu_{kj}) + (\beta-1)log(1-\mu_{kj}))$

$\frac{\partial Q}{\partial \mu_{kj}} = 0 + \sum_{i=1}^m \frac{r_k^{(i)} x_j^{(i)}}{\mu_{kj}} - \sum_{i=1}^m \frac{r_k^{(i)}(1-x_j^{(i)})}{(1-\mu_{kj})} + \frac{(\alpha-1)}{\mu_{kj}} - \frac{(\beta-1)}{(1-\mu_{kj})} = 0$

$(1-\mu_{kj})(\sum_{i=1}^m r_k^{(i)} x_j^{(i)} + \alpha - 1) = \mu_{kj}(\sum_{i=1}^m r_k^{(i)}(1-x_j^{(i)}) + \beta - 1)$

$(\sum_{i=1}^m r_k^{(i)} x_j^{(i)}) + \alpha - 1 = \mu_{kj}((\sum_{i=1}^m r_k^{(i)}(1-x_j^{(i)})) + \beta - 1 + (\sum_{i=1}^m r_k^{(i)} x_j^{(i)}) + \alpha - 1)$

$(\sum_{i=1}^m r_k^{(i)} x_j^{(i)}) + \alpha - 1 = \mu_{kj}((\sum_{i=1}^m r_k^{(i)}) + \alpha + \beta - 2)$

$\mu_{kj} = \frac{(\sum_{i=1}^m r_k^{(i)} x_j^{(i)}) + \alpha - 1}{(\sum_{i=1}^m r_k^{(i)}) + \alpha + \beta - 2}$

2 Principal Components Analysis

$f_u(x) = u(argmin_\alpha \|x - \alpha u\|^2)$

$f_u(x) = u(argmin_\alpha (x^T x - 2\alpha x^T u + \alpha^2 u^T u))$

Applying the minimum of a convex quadratic function $ax^2 + bx + c = 0$ is $x = \frac{-b}{2a}$, we have:

$f_u(x) = u(\frac{2x^T u}{2u^T u}) = ux^T u = u^T x u$

$argmin_{u:uu^T=1} \sum_{i=1}^m \|x^{(i)} - f_u(x^{(i)}\|^2$
$= argmin_{u:uu^T=1} \sum_{i=1}^m \|x^{(i)} - u^T x^{(i)} u\|^2$
$= argmin_{u:uu^T=1} \sum_{i=1}^m (x^{(i)} - u^T x^{(i)} u)^T (x^{(i)} - u^T x^{(i)} u)$
$= argmin_{u:uu^T=1} \sum_{i=1}^m (x^{(i)T} x^{(i)} - 2(u^T x^{(i)})^2 + u^T u(u^T x^{(i)})^2)$
$= argmin_{u:uu^T=1} \sum_{i=1}^m (x^{(i)T} x^{(i)} - 2(u^T x^{(i)})^2 + (u^T x^{(i)})^2)$
$= argmin_{u:uu^T=1} \sum_{i=1}^m -(u^T x^{(i)})^2$
$= argmax_{u:uu^T=1} u^T (\sum_{i=1}^m x^{(i)} x^{(i)T}) u$

This corresponds to the optimization problem that defines the first principal component.

3 K-means clustering

Problem 3.1: Finding closest centroids

Closest centroids for the first 3 examples: (should be [0 2 1]): [0 2 1]

Problem 3.2: Computing centroid means

Computing centroids means.
Centroids computed after initial finding of closest centroids:
[[2.428301113.15792418]
[5.813503312.63365645]
[7.119386873.6166844]]

The centroids should be
[2.4283013.157924], [5.8135032.633656], [7.1193873.616684]

k-means on example dataset

Figure 1: Expected output of k-means

Problem 3.3: Random initialization

Figure 2: Original and reconstructed image (when using k-means to compress the image).

4 Principal Components Analysis

Problem 4.1: Implementing PCA

Figure 3: Computed eigenvectors of the dataset

Problem 4.2: Projecting the data onto the principal components

The projection of the first example (should be about 1.496) [1.49631261]
Approximation of the first example (should be about [-1.058 -1.058]) [-1.05805279 -1.05805279]

Problem 4.3: Reconstructing an approximation of the data

Figure 4: The normalized and projected data after PCA
Figure 5: The first 25 principal components on the face dataset
Figure 6: Reconstructed face dataset from only the top 100 principal components

5 Anomaly detection

Problem 5.1: Estimating parameters of a Gaussian distribution

Figure 7: The Gaussian distribution contours of the distribution fit to the dataset

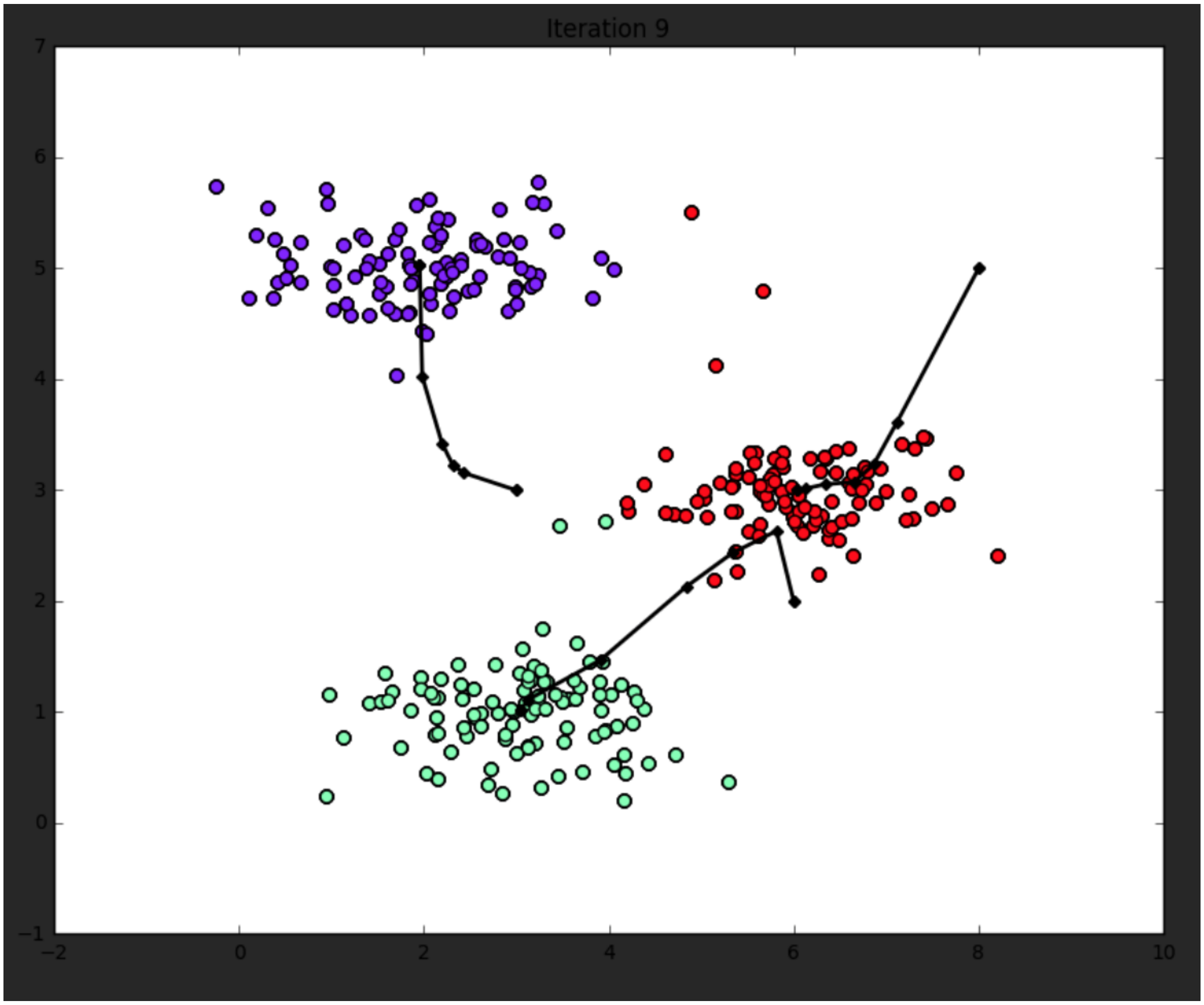

Figure 1: Expected output of k-means

Problem 5.2: Selecting the threshold epislon
Best threshold epsilon: 8.99085277927e-05
Best F1: 0.875
Figure 8: The classified anomalies

High dimensional dataset
Best threshold epsilon: 1.37722889076e-18
Best F1: 0.615384615385
117 anomalies

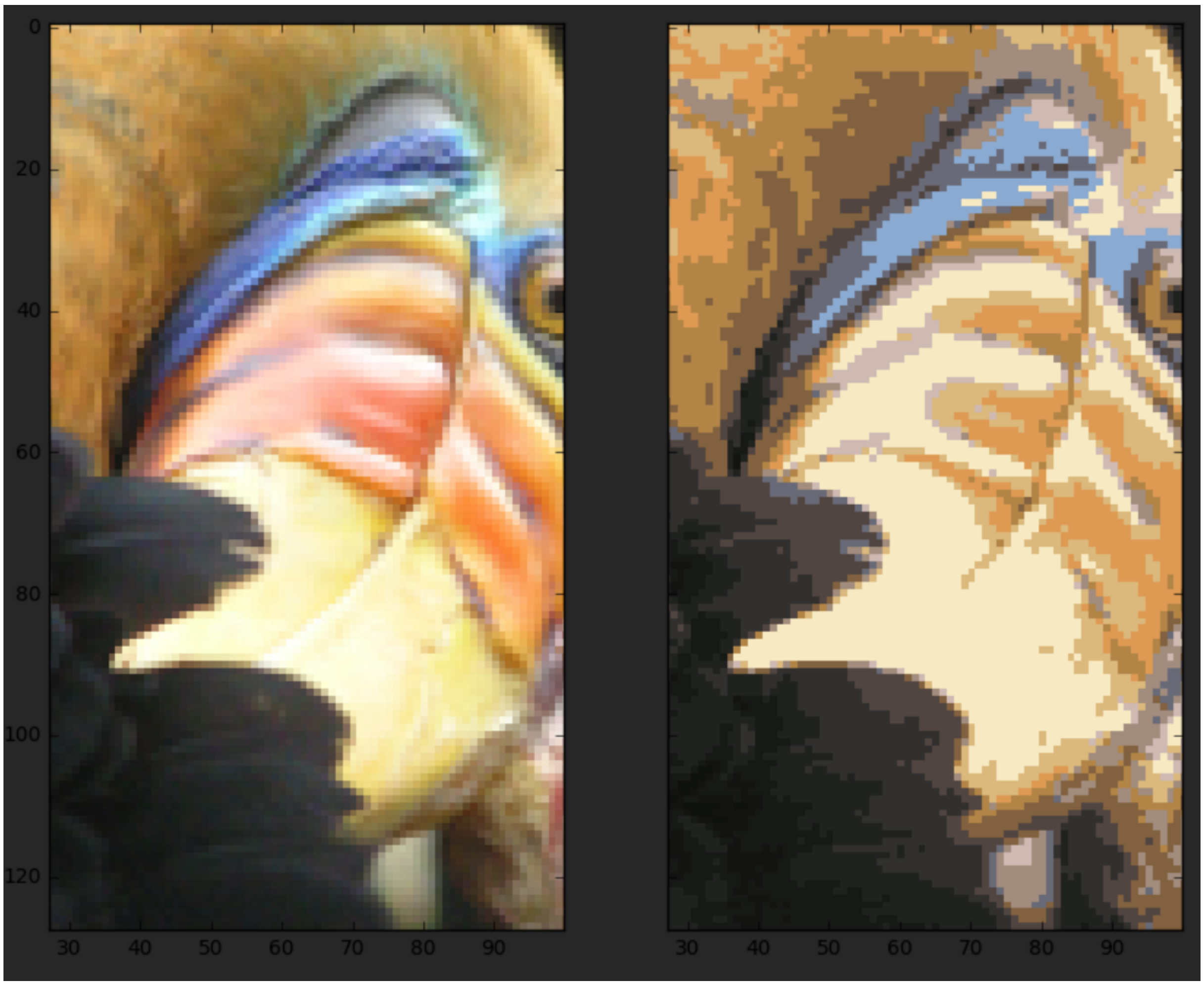

Figure 2: Original and reconstructed image (when using k-means to compress the image)

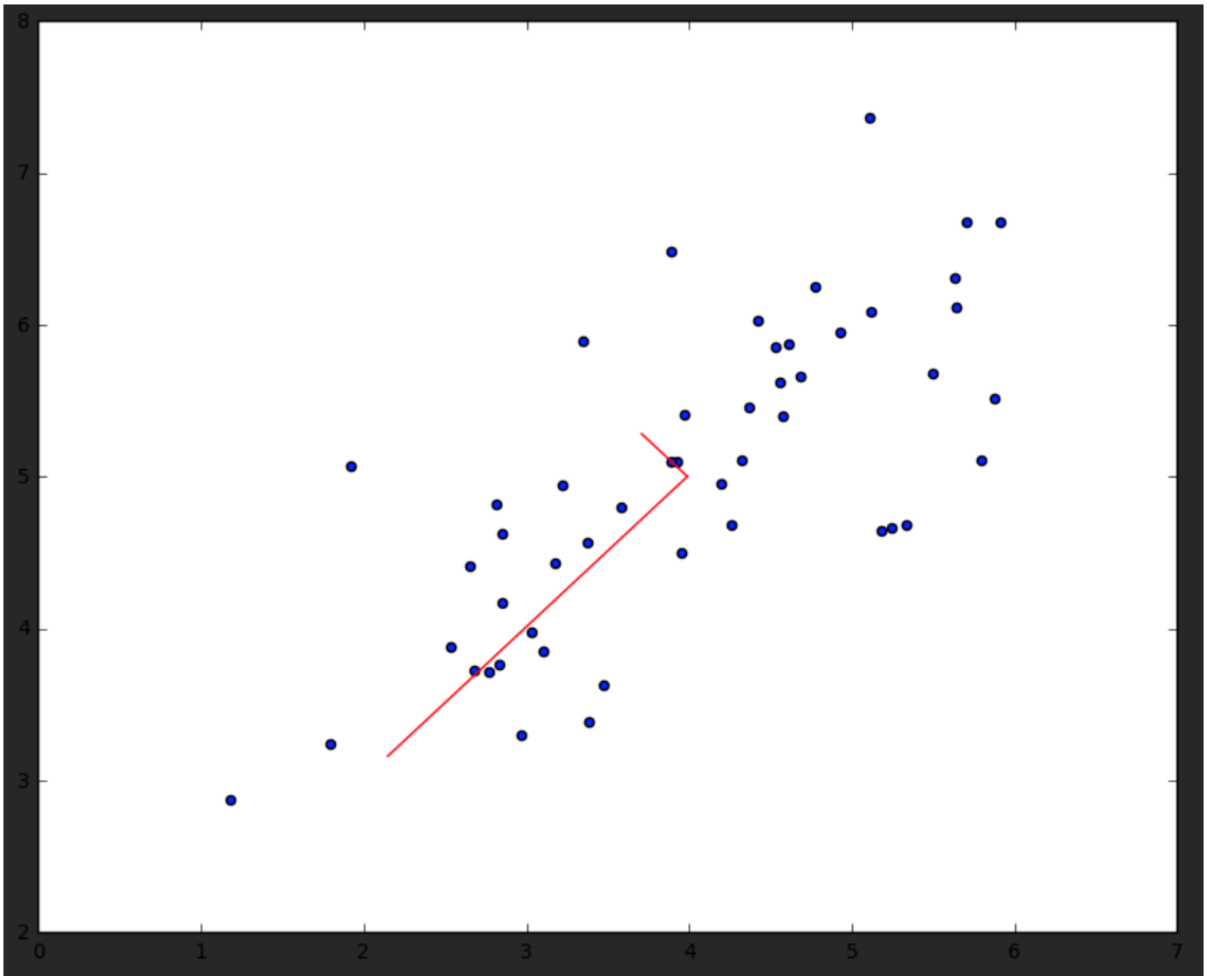

Figure 3: Computed eigenvectors of the dataset

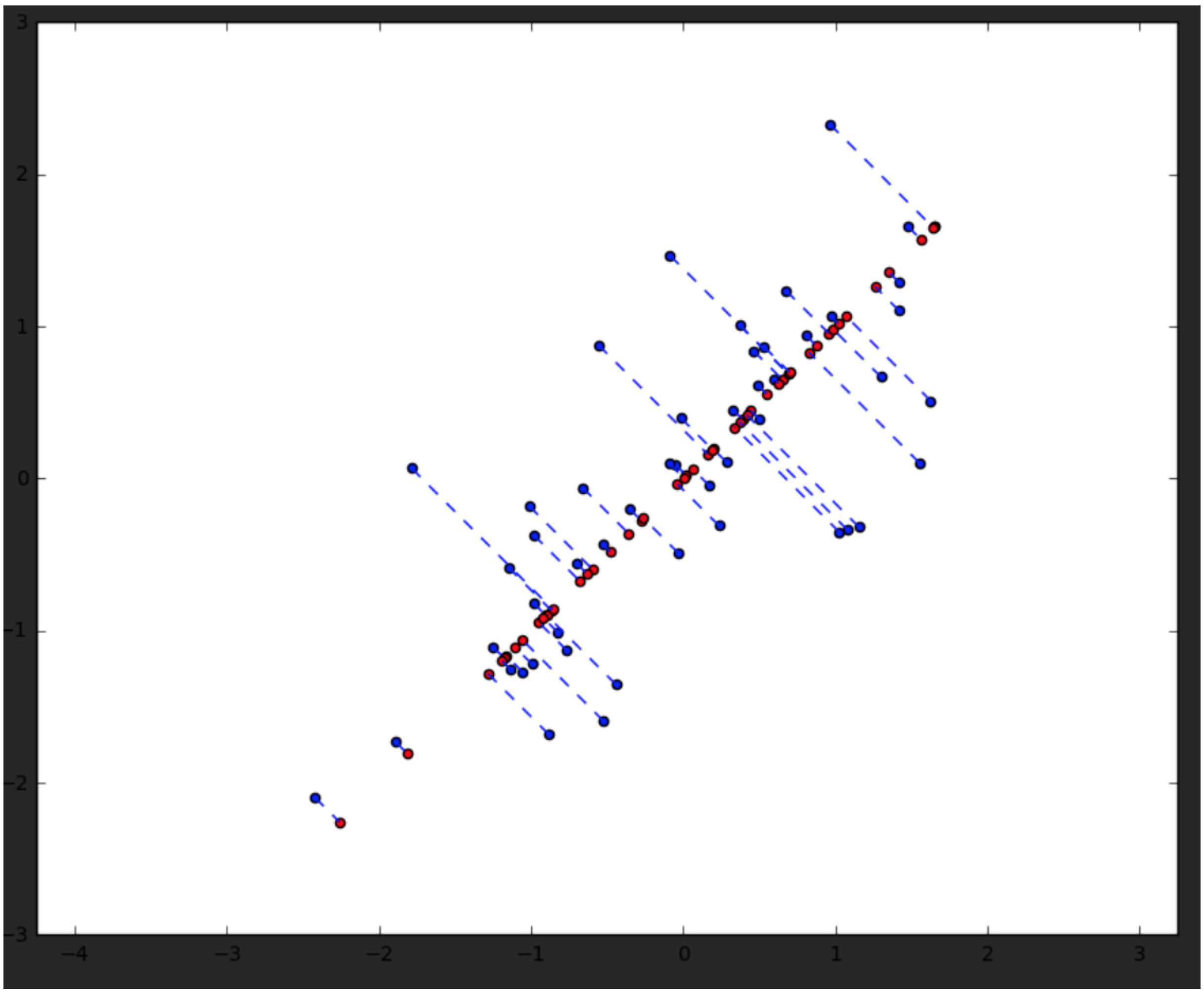

Figure 4: The normalized and projected data after PCA

Figure 5: The first 25 principal components on the face dataset

Figure 6: Reconstructed face dataset from only the top 100 principal components

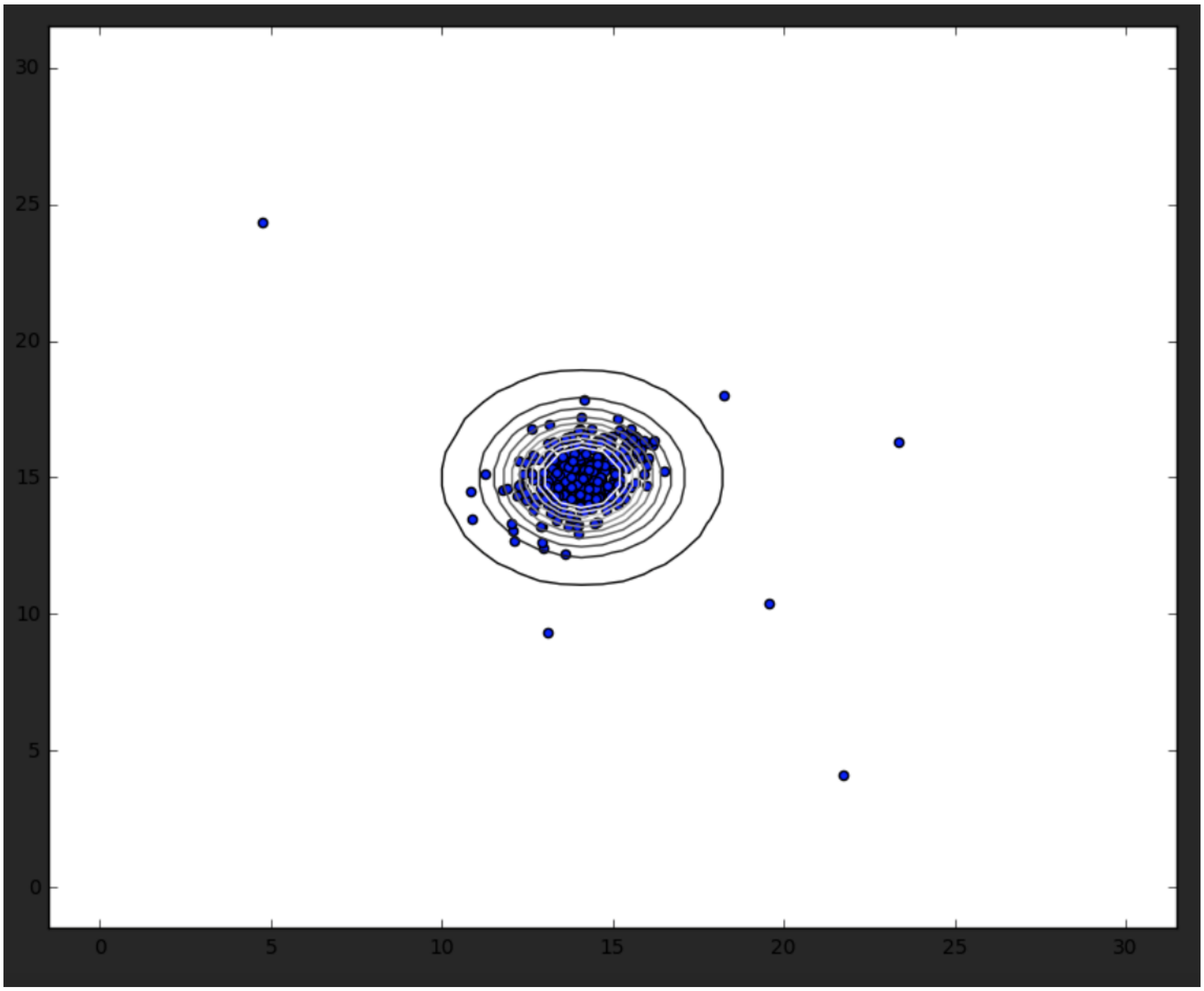

Figure 7: The Gaussian distribution contours of the distribution fit to the dataset

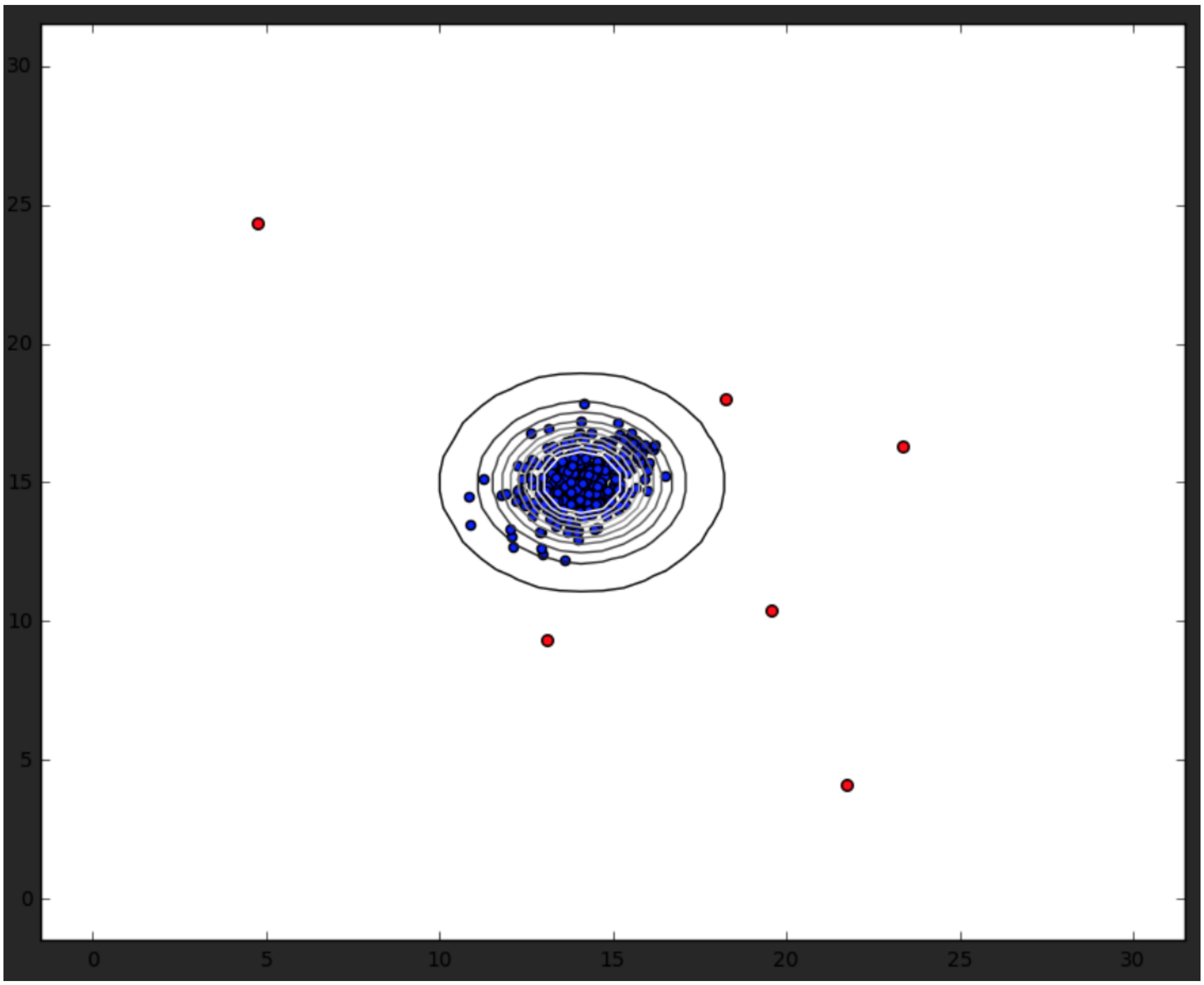

Figure 8: The classified anomalies