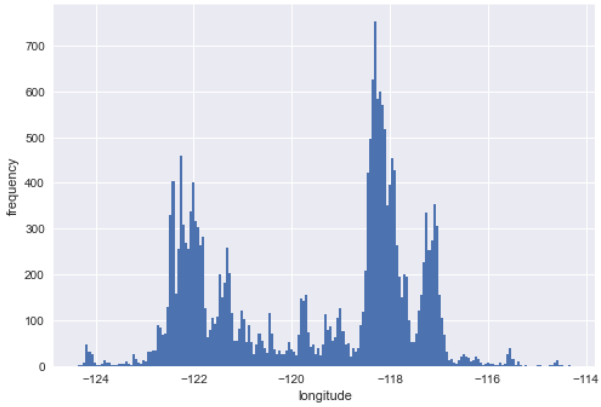
mix-proj

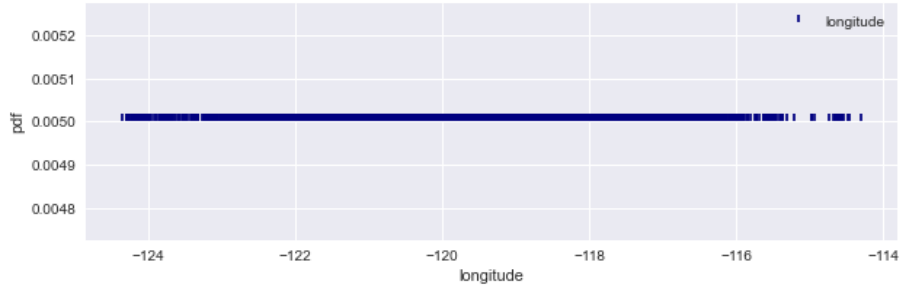
1. Overview

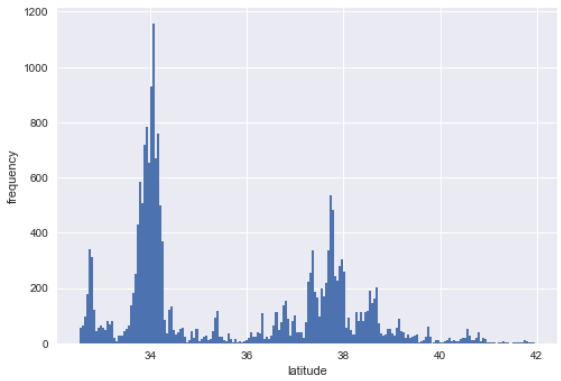
In my project, EM Algorithm and Gibbs Sampler were respectively used for data cluster analysis and comparison.

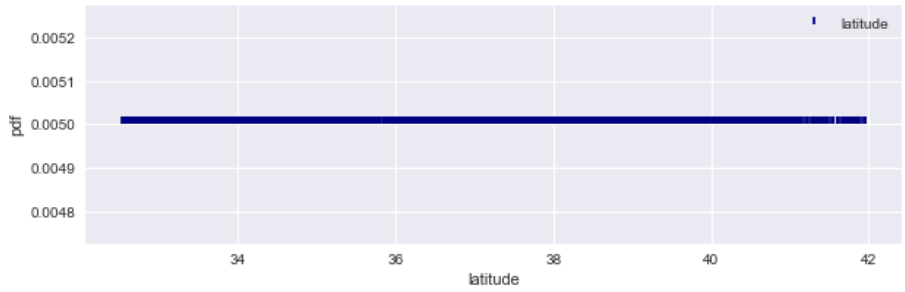
2. First look at the data

I use the data from http://www.rob-mcculloch.org/data/calhouse.csv. I tried to perform cluster analysis on longitude and latitude dimensions. Firstly,we can take a look at the distribution of the data.









3. GMM and EM algorithm

3.1 1D mixture modelling and EM algorithm

I tried to do cluster analysis from a one-dimensional perspective, respectively at longitude and latitude. First, I need to give the initial value of the parameter. The initial values of the mean and variance can be calculated by estimation. After that, you can try the EM algorithm.

the E Step

If our observation XI comes from a mixing model with K mixing components, the marginal probability distribution of XI would look like this:

$$P(X_i = x) = \sum_{k=1}^{K} \pi_k P(X_i = x | Z_i = k)$$

where Zi ∈ {1, ··· K} is the latent variable representing the mixture component for Xi, P (Xi | Zi) is the mixture component, and πk is the mixture proportion representing the probability that Xi belongs to the kth mixture component.

The probability of each observation x_i can be calculated using the estimated parameters. For each cluster k = 1,2, estimates of the mean and variance can be used to calculate the probability density (PDF) of the data.

$$f(\mathbf{x}|\mu_k, \sigma_k^2) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp(-\frac{(\mathbf{x} - \mu_k)^2}{2\sigma_k^2})$$

The probability that a given x_i is located in a k cluster can be calculated.

$$\mathbf{b}_k = \frac{f(\mathbf{x}|\mu_k, \sigma_k^2)\phi_k}{\sum_{k=1}^{K} f(\mathbf{x}|\mu_k, \sigma_k^2)\phi_k}$$

Bayes' theorem can be used to find the posterior probability of the KTH Gaussian distribution to interpret the data, thus observing the possibility that x_i is generated by the KTH Gaussian. Since there is no other information to support a Gaussian, we can guess that an example with the same probability will come from every Gaussian. We can keep improving our priors in each iteration until they converge.

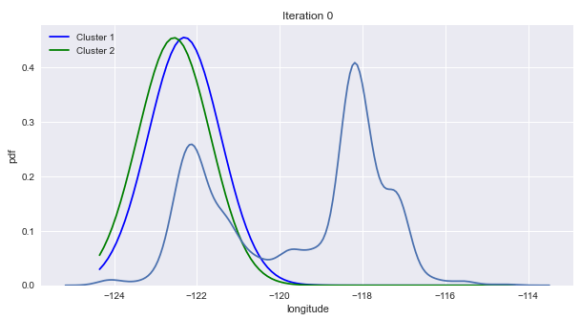
the M Step

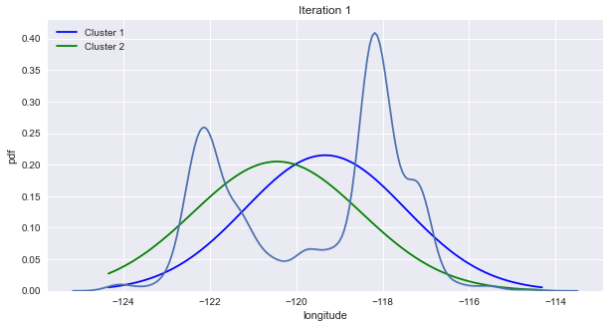
The following formula can be used to re-estimate the parameters.

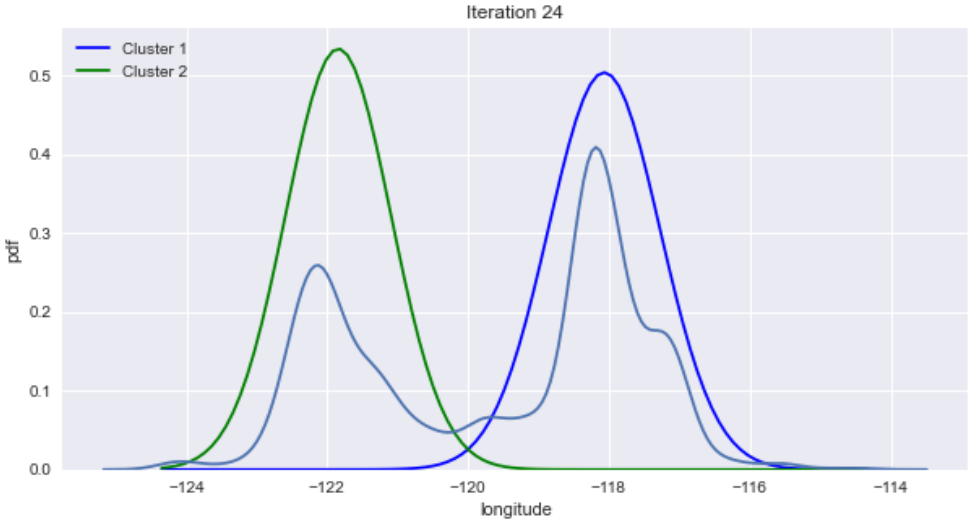
$$\mu_k = \frac{\sum \mathbf{b}_k \mathbf{x}}{\sum \mathbf{b}_k} \qquad \sigma_k^2 = \frac{\sum \mathbf{b}_k (\mathbf{x} - \mu_k)^2}{\sum \mathbf{b}_k} \qquad \phi_k = \frac{1}{N} \sum \mathbf{b}_k$$

In each iteration, we continuously updated the mean (μ_k), variance (σ_2^2) and zoom parameter values of each class. This makes it look like a real data distribution. And then you repeat these steps until the convergence is stable.]

The following images are the results of our 0th, first and 24th iterations using the longitude data.







The mean value of the two longitude clusters is [-118.06738742, -121.84256838]. The variance of the two clusters is [0.62666746, 0.55809504]. The proportions of the two clusters are [0.60205467, 0.39794513].

Using the same method, we can get the mean of the two latitude clusters is [033.89702206, 37.9146538]. The variance of the two clusters is [0.3090631, 0.98696638]. The proportions of the two clusters are [0.56819393, 0.43180573].

3.2 2D mixture modelling and EM algorithm with 100 iterations

Similarly, in the EM algorithm under the Gaussian mixture model, let N (μ, σ2) represent the probability distribution function of a normal random variable. In this case, we have the conditional distribution $\hat{X}_i|Z_i = k \sim N(\mu_k, \sigma_k^2)$, so the marginal distribution of Xi as follows:

$$P(X_i = x) = \sum_{k=1}^{K} P(Z_i = k)P(X_i = x|Z_i = k) = \sum_{k=1}^{K} \pi_k N(x; \mu_k, \sigma_k^2)$$

the joint probability of observations X1,⋯ ,Xn is:

$$P(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k N(x_i; \mu_k, \sigma_k^2)$$

The unknown parameter is θ={μ1...μK, σ1...σk π1...πK}, so the likelihood value is:

$$L(\theta | X_1, \ldots, X_n) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k N(x_i; \mu_k, \sigma_k^2)$$

The log-likelihood is:

$$\ell(\theta) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} \pi_k N(x_i; \mu_k, \sigma_k^2) \right)$$

We can take the derivative of μk, and set the expression to 0, and we will get:

$$\sum_{i=1}^{n} \frac{1}{\sum_{k=1}^{K} \pi_k N(x_i; \mu_k, \sigma_k)} \pi_k N(x_i; \mu_k, \sigma_k) \frac{(x_i - \mu_k)}{\sigma_k^2} = 0$$

Then we can simply collect all the samples XI, set ZI = K, and then use the estimation from the previous section to estimate μk,
Firstly, choose initial values for μ,σ,π and use these in the E-step to evaluate the $\gamma_{Z_i}(k)$.

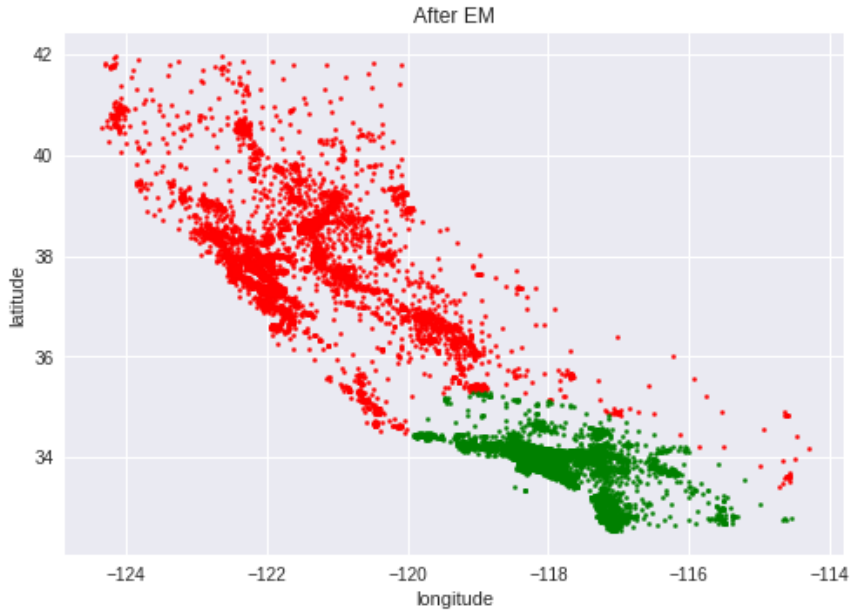
$$P(X, Z | \mu, \sigma, \pi) = \prod_{i=1}^{n} \prod_{k=1}^{K} \pi_k^{I(Z_i=k)} N(x_i | \mu_k, \sigma_k)^{I(Z_i=k)}$$

$$\log(P(X, Z | \mu, \sigma, \pi)) = \sum_{i=1}^{n} \sum_{k=1}^{K} I(Z_i = k) \left(\log(\pi_k) + \log(N(x_i | \mu_k, \sigma_k)) \right)$$

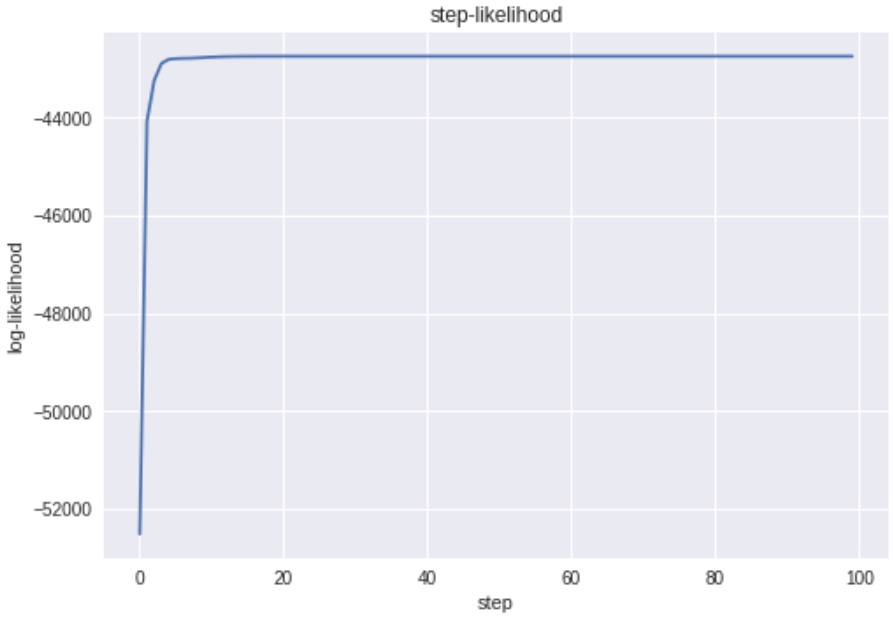
Then, with $\gamma_{Z_i}(k)$ fixed, maximize the expected complete log-likelihood above with respect to μk, σk and πk. This leads to the closed form solutions we derived in the previous section. Formula is as follows:

$$E_{Z|X}[\log(P(X, Z | \mu, \sigma, \pi))] = \sum_{i=1}^{n} \sum_{k=1}^{K} \gamma_{Z_i}(k) \left(\log(\pi_k) + \log(N(x_i | \mu_k, \sigma_k)) \right)$$

In the longitude and latitude clustering analysis, we obtained the classification as shown in the figure after 100 iterations of EM algorithm.



And then there's the Log Likelihood convergence image of EM.



Our result is

```
result:
k: [0.4596379526967196, 0.5403620473032804]
mu: [array([-121.49936234,   37.74430448]), array([-117.92831605,   33.83499395])]
sigma: [array([[ 1.37044727, -0.81963015],
       [-0.81963015,  1.39304297]]), array([[ 0.40105933, -0.20872926],
       [-0.20872926,  0.23316952]])]
```

4. GMM and Gibbs sampler

The main idea behind Gibbs sampling (and all of MCMC) is to approximate a distribution with a set of samples. For example, in the mixture model,

$$p(\mu, z \mid x) \approx \frac{1}{B} \sum_{b=1}^{B} \delta_{(\mu^{(b)}, z^{(b)})}(\mu, z),$$

In the Gibbs sampler, we maintain a value for each latent variable. In each iteration, sample from each latent variable conditional on the other latent variables and the observations.

Firstly, we input data x and a number of components K and initialize the Mixture locations μ. Maintain mixture locations μ and mixture assignments z. Then we need to repeat following step:

For each i∈{1,⋯,n},

$$\begin{aligned} p(z_i \mid \boldsymbol{\mu}, x_i) &\propto p(z_i) p(x_i \mid \mu_{z_i}) \\ &= \pi_{z_i} \phi(x_i; \mu_{z_i}, \sigma^2). \end{aligned}$$

For each k∈{1,⋯,K},

$$\mu_k \mid \mathbf{z}, \mathbf{x} \sim \mathcal{N}(\hat{\mu}_k, \hat{\lambda}_k)$$

where

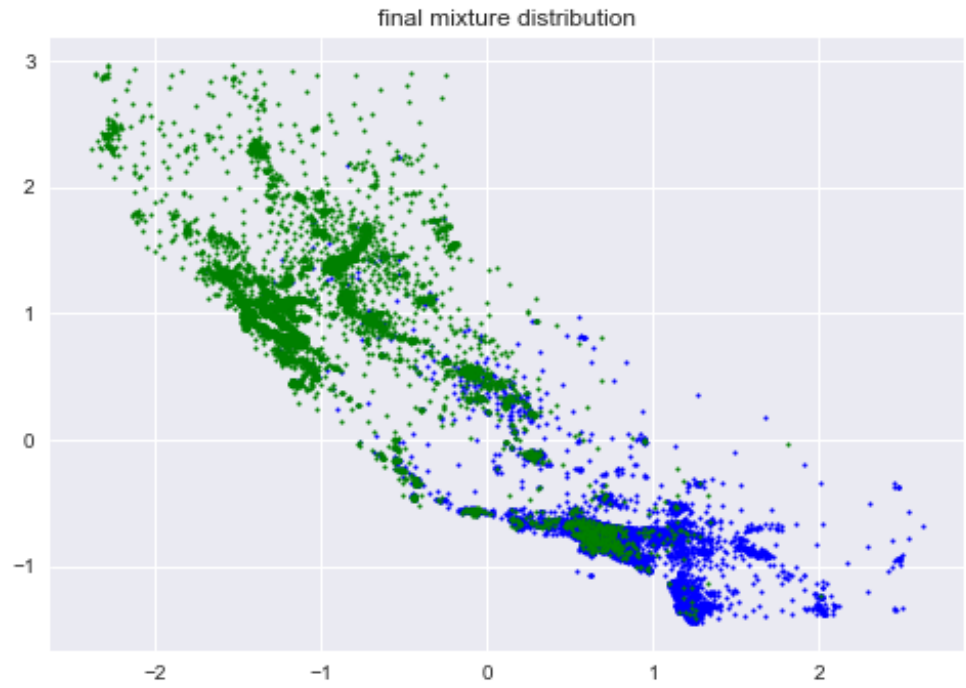
$$\begin{aligned} \hat{\mu}_k &= \left(\frac{n_k/\sigma^2}{n_k/\sigma^2 + 1/\lambda^2} \right) \bar{x}_k \\ \hat{\lambda} &= \left(n_k/\sigma^2 + 1/\lambda^2 \right)^{-1}, \end{aligned}$$

and

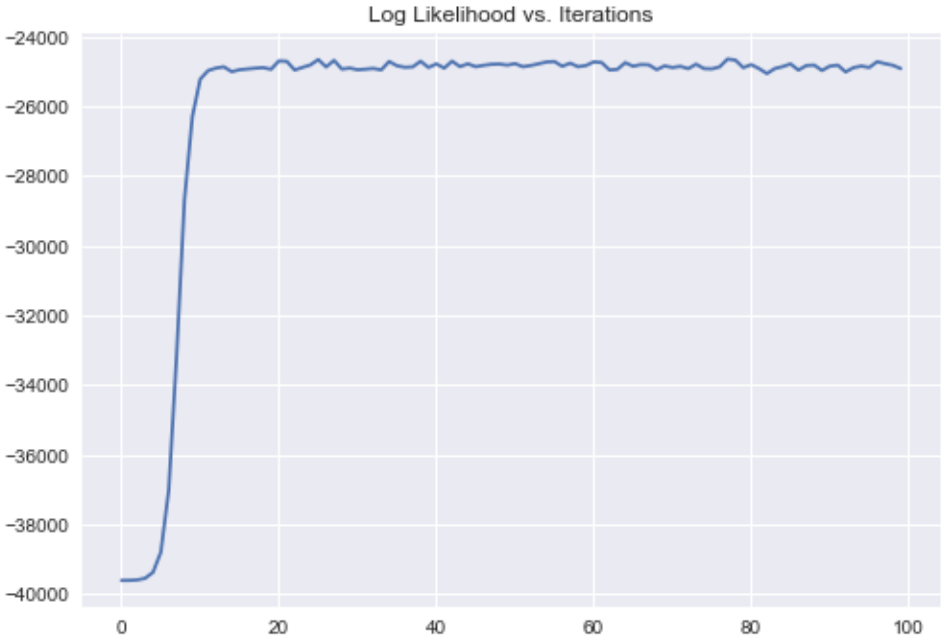
$$\begin{aligned} n_k &= \sum_{i=1}^{n} z_i^k \\ \bar{x}_k &= \frac{\sum_{i=1}^{n} z_i^k x_i}{n_k}. \end{aligned}$$

The reason is that we have defined a Markov chain whose state space are the latent variables and whose stationary distribution is the posterior we care about. After a long time, the samples of μ and Z are from the later samples. And if we do this multiple times, we can get sample B from a posterior.

In the longitude and latitude clustering analysis, we obtained the classification as shown in the figure after 100 iterations of Gibbs Sampler.



And then there's the Log Likelihood convergence image of Gibbs Sampler.



Our result is

```
[1 1 1 ... 1 1 1]
[[ 0.73561834 -0.73673703]
 [-0.96532037  0.9667682 ]]
```

```
ProbCluster0,ProbCluster1
```

```
(0.5659399224806202, 0.43406007751937986)
```

5. **Compare EM algorithm with Gibbs Sampler**

   The percentages of the two clusters obtained by EM algorithm after 100 iterations are 0.46 and 0.54. The mean values of the two clusters' longitude and latitude are about(-

121.49936234, 37.74430448), (-117.92831605, 33.83499395). The calculations took more than an hour, significantly longer than the Gibbs Sampler.

The proportions of the two clusters obtained by Gibbs Sampler after 100 iterations are 0.43 and 0.57. The scaled mean values of the two clustering methods, longitude and latitude, are (0.73561834, -0.73673703) and (-0.96532037, 0.9667682), which require inverse scaling to obtain accurate results, which are troublesome in coding.

The log Likelihood of EM converges between -44000 and -42000.The loglikelihood of Gibbs Sampler converges between -26000 and -24000.In this respect, Gibbs Sampler is the better performer.

EM and Gibbs Sampler are two different ways to predict the hidden variables of a statistical model. EM is essentially an algorithm for finding MLE, which is a point estimate. The EM likelihood function for multimodal cannot guarantee the global optimum. EM algorithm is non-convex and can easily fall into local optimum. And the code is slower to run, the computer memory requirements are higher.

Gibbs Sampler is actually the construction of a Markov process, the advantage is simple derivation. It is not about convergence for a particular point in the parameter space θ, but about whether the probability distribution covering θ converges to the true distribution P*(θ).In other words, its convergence is discussed in a space containing all probability distributions over θ, where each point represents a distribution over θ. If the Markov Chain behind Gibbs Sampler is not ergodic, there will be a problem. If Ergodicity is satisfied, theoretically it must converge. Gibbs Sampler seems to test data more rigorously, requiring scaling of data in large volumes.