HW Week 4

Code

German Tank Problem: Bayesian Approach

Perhaps the set is: [M98, M508, M727, M520, K85, K58, K13, K7, K74, K75, K64, F225, F292, F241, F453, F464, F165, F182, F334, F88]

We can assume that there are three factories: K,F,M.

Write out the likelihood of observing a single tank. If N is less than m (the observered highest serial number), than we cannot proceed with our computation. So, given $N \geqslant m$ the probability is $\frac{1}{N}$.

Derive the maximum likelihood formula for the total number of tanks given a dataset as above. For each factory, it is fair to assume the following probability mass function:

$$\Pr(N = n) = \begin{cases} 0 & \text{if } n < M \\ \frac{k-1}{k} \frac{\binom{m-1}{k-1}}{\binom{n}{k}} & \text{if } n \geq M, \end{cases}$$

Using the likelihood:

$$\frac{\binom{m-1}{k-1}}{\binom{n}{k}}$$

We get: Factory M:

$$\frac{\binom{726}{3}}{\binom{n}{4}} = 726$$

Factory K:

$$\frac{\binom{84}{6}}{\binom{n}{7}} = 84$$

Factory F:

$$\frac{\binom{463}{8}}{\binom{n}{9}} = 463$$

Use the formula that you derived in the previous question to estimate how many tanks there are given the dataset in the description.

In order to maximize/estimate the total number tanks, we modify n and we get: 726, 84,462 as shown above.

How strongly do you believe your estimate? I think that the method used is heavily realying on multiple assumptions, and therefore is not substantialy reliable. Firstly, we assume continuous linearity in the serial numbers. This is not promised. Moreover, we are making the assumption that the tanks observed tell us ALOT about the total number of tanks. There are too many unknowns in order to determine whether this assumption is even plausible. During what time period the tanks were seen? in how many different places observations were collected. There could be a situation where the observations were collected by a specific tool and this will lead to sample bias.*

Mixture of Gaussians:

Using your generic equations (derived from the pre-class work on expectation maximization) show what the equations would be for a mixture of Gaussian distributions.

A gaussian mixture model is defined by a sum of gaussians:

$$P(x) = \sum_i w_i(\mu_i, \Sigma_i)$$

with means μ and covariance matrices Σ.

When we generalize the Gaussian dist. to two dimmenssions the formula is:

$$f(x, \mu, \Sigma) = \frac{1}{(2\pi)^{d/2}\sqrt{(|\Sigma|)}} e^{(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}$$

(adopted from: http://www.ics.uci.edu/~smyth/courses/cs274/notes/EMnotes.pdf (http://www.ics.uci.edu/~smyth/courses/cs274/notes/EMnotes.pdf)) We are given a data set $D = x_1, \ldots, x_N$ where x_i is a d-dimensional vector measurement. Assume that the points are generated in an IID fashion from an underlying density $p(x)$. We further assume that $p(x)$ is defined as a finite mixture model with K components:

$$p(x|\Theta) = \sum_{k=1}^{K} \alpha_k p_k(x|z_k, \theta_k)$$

where:

• The $p_k(x|z_k, \theta_k)$ are mixture components, $1 \leq k \leq K$. Each is a density or distribution defined over $p(x)$, with parameters θ_k.

• z = (z1, . . . , zK) is a vector of K binary indicator variables that are mutually exclusive and exhaustive (i.e., one and only one of the zk's is equal to 1, and the others are 0). z is a K-ary random variable representing the identity of the mixture component that generated x. It is convenient for mixture models to represent z as a vector of K indicator variables.

• The αk = p(zk) are the mixture weights, representing the probability that a randomly selected x was generated by component k, where PK k=1 αk = 1.

The probability of the data for a Mixture Gaussian is:

$$p(x1 : N|c1 : N) = \prod_n Nor(x_n|\mu_{cn}, \sigma 2 \cdot I)$$

Maximizing the likelihood of the data with regards to the model parameters:

$$\theta^{new} = argmax_\theta \cdot Q(\theta, \theta^{old})$$

where

$$Q(\theta, \theta^{old}) = \sum_Z p(Z|X, \theta^{old}) \cdot ln(p(X, Z|\theta))$$

An adapted code with Gaussian Mixture and plotting can be found on: https://gist.github.com/oba2311/8ce5273e5d7dadf52bf70a929d4441b0 (https://gist.github.com/oba2311/8ce5273e5d7dadf52bf70a929d4441b0)

The number of clusters is crucial for GMM. If the data is not normal, running a method to find hte perfect K (optimized number of clusters) will most liekly arrive at wrong number of clusters, a number that is most likely to represent that data as somewhat guassian. Moreover, if the sizes of the clusters (even in normally distributeed dataset) is highly uneven, the error will grow.

Stretch Goal: The frequentist approach for the German Tanks question:

A frquentist would use the following formula:

$$N \approx m + \frac{m}{k} - 1 = 727 + \frac{727}{20} - 1 = 762.35$$

(there is also a frequentist version that uses the median (noted usually "m") and assumes $2m - 1$ for n)

$$\begin{aligned} N &\approx \mu \pm \sigma = 88.5 \pm 50.22, \\ \mu &= (m-1)\frac{k-1}{k-2}, \\ \sigma &= \sqrt{\frac{(k-1)(m-1)(m-k+1)}{(k-3)(k-2)^2}} \end{aligned}$$

We get:

$$\begin{aligned} N &\approx 766 \pm 42 = 802, 724), \\ \mu &= 766, \\ \sigma &= 42 \end{aligned}$$

We make the assumption that the factories are independent, i.e number of tanks produced in factofy F · number of tanks in factory m is equal to number of tanks in factory m (iid assumption). Therefore, the maximum likelihood for the three factories is an addition of the maximum likelihood per each one added.

Stretch Goal: another method that was actually used (according to wikipedia):

Given randomly chosen k units out of discrete uniform distribution with 1…N units, the minimum-variance unbiased estimator

$$\frac{k+1}{k}m - 1,$$

where m is the sample maximum.

So: k=20,m=727.

$$\frac{21}{20}(726) = 762.3$$

with variance of:

$$\mathrm{Var}(\hat{N}) = \frac{1}{k}\frac{(N-k)(N+1)}{(k+2)} \approx \frac{N^2}{k^2} \text{ for small samples } k \ll N,$$

Which is:

$$\frac{(762.3)^2}{(20)^2} = 1452.753$$

and SD of

$$38.114$$

.

*#Sampling - Design effective sampling methods and evaluate the interpretation of results accordingly. (C) EA In this case, there is not enough information about the sampling method in order to determine whether sample bias could take place. One has to be aware of this caveat when assessing the MLE estimation.

Resources: http://hameddaily.blogspot.kr/2015/03/when-not-to-use-gaussian-mixtures-model.html (http://hameddaily.blogspot.kr/2015/03/when-not-to-use-gaussian-mixtures-model.html) http://yulearning.blogspot.kr/2014/11/einsteins-most-famous-equation-is-emc2.html (http://yulearning.blogspot.kr/2014/11/einsteins-most-famous-equation-is-emc2.html) http://www.ics.uci.edu/~smyth/courses/cs274/notes/EMnotes.pdf (http://www.ics.uci.edu/~smyth/courses/cs274/notes/EMnotes.pdf) https://www.r-bloggers.com/bayesian-tanks/ (https://www.r-bloggers.com/bayesian-tanks/) , https://en.wikipedia.org/wiki/German_tank_problem#Minimum-variance_unbiased_estimator (https://en.wikipedia.org/wiki/German_tank_problem#Minimum-variance_unbiased_estimator), https://www.wired.com/2010/10/how-the-allies-used-math-against-german-tanks/ (https://www.wired.com/2010/10/how-the-allies-used-math-against-german-tanks/)