Optimal k-means vs. Bayes Decision Rule

Thomas Athey

April 3, 2019

1 Problem Definition

1.1 K Means

Say we have the dataset

$$X = \{x_1, x_2, \dots, x_n\} \,\epsilon\, (R^p)^n$$

that we seek to partition into k clusters. The k-means algorithm seeks to find the cluster centers:

$$C = \{c_1, c_2, \dots, c_k\} \,\epsilon\, (R^p)^k$$

and the cluster assignments

$$A = \{a_1, a_2, \dots, a_n\} \,\epsilon\, \{1, 2, \dots, k\}^n$$

that minimize the cost function:

$$J = \frac{1}{n}\sum_{i=1}^{n} ||x_i - c_{a_i}||_2^2$$

1.2 K Means Decision Rule

Given a set of centers C, it is obvious that the optimal decision rule ($D : R^p \rightarrow 1, 2, \dots, k$) assigns each point x_i to the closest centers i.e.

$$D(x_i) = a_i = \underset{j}{\operatorname{argmin}} ||x_i - c_j||_2^2 \tag{1}$$

This process of assigning datapoints to clusters can be thought of as a decision rule. This decision rule will involve linear boundaries between clusters since the relative proximity of two different cluster centers only changes along the axis that spans the two centers. Now, let's flip this problem around. Given a descision rule comprised of a set of linear decision boundaries (D), the ideal cluster centers are, obviously, the means of each cluster i.e.

$$c_j = \frac{\sum\limits_{i=1}^{n} x_i * I(a_i == j)}{\sum\limits_{i=1}^{n} I(a_i == j)}$$

where $I(x)$ is the indicator function of the boolean x.

Say that each X_i is an iid random variable that follows some distribution f_X. Given some decision rule with linear boundaries (D), and thus a set of cluster centers (C_D), the cluster assignment, $D(X)$ is a random variable according to 1.

The law of large numbers tells us that:

$$J_D = \lim_{n\to\infty} J = E[||X - c_{D(X)}||_2^2] = E_{D(X)}[E_{X|D(X)=a}[||X - c_a||_2^2]] \tag{2}$$

Finally, let us define the optimal large sample k-means decision rule, D_k^*, as:

$$D_k^*(x) = \operatorname*{argmin}_D J_D \tag{3}$$

where J_D is defined in 2.

1.3 Bayes Decision Rule of Mixture Distribution

We are studying clustering in the context of underlying mixture distributions. A typical form of a mixture distribution is a weighted sum of the same parameterized distribution, with each component having its own set of parameters. A k-mixture of a distribution f is defined as:

$$f_X(x) = \sum_{i=1}^{k} \tau_i * f(x;\theta_i) \tag{4}$$

where f may be, for example, the normal distribution, in which case $\theta_i = \{\mu_i, \Sigma_i\}$.

This distribution can be interpreted as the distribution of X marginalized over the hidden cluster assignment A i.e.:

$$Pr(A = a) = \tau_a$$

$$f_{X|A=a}(x) = f(x;\theta_a)$$

The Bayes decision rule, D_B^* chooses the value of A that minimizes the probability of error i.e.:

$$D_B^*(x) = \operatorname*{argmin}_a Pr(A \neq a|X = x)$$

1.4 Problem

We want to characterize the settings where $D_k^* = D_B^*$.

2 MGFs

2.1 MGF of Mixture

The moment generation function (MGF) of a random variable, X, is defined as:

$$M_X(t) = E_X[e^{t^T x}]$$

Due to linearity of expectation, the MGF of the mixture distribution is the weighted sum of the MGFs of the components. Say X is distributed according to 4, and the MGF associated with $f(x;\theta_i)$ is M_i, then:

$$M_X(t) = \sum_{i=1}^{k} \tau_i * M_i(t) \tag{5}$$

2.2 MGF of a Normal Mixture split in Half

In k-means, we partition the space of R^p into halfspaces and calculate the first two moments of X within these partitions. We calculate first moments to get cluster centers C, and second moments to get costs J. Here we present a series of equations that can be assmebled to find an explicit expression of the MGF of "half of a mixture" i.e. $I(u^T x > c) * f_X$.

$$q_i = \int_{u^T x = c}^{\infty} N(x;\mu_i,\Sigma_i)dx = \frac{1}{2} erfc(\frac{c - u^T \mu_i}{||D_i^{1/2} U_i^T u||\sqrt{2}}) \tag{6}$$

where $erfc(x)$ is the complementary error function and $\Sigma_i = U_i D_i U_i^T$ is the diagonalization of Σ.

$$m_i(t) = \int_{u^T x = c}^{\infty} e^{t^T x} N(x;\mu_i,\Sigma_i)dx = \frac{1}{2} e^{\mu_i^T t + \frac{1}{2} t^T \Sigma_i t} erfc(\frac{c - u^T(\mu_i + \Sigma_i t)}{||D_i^{1/2} U_i^T u||\sqrt{2}}) \tag{7}$$

The derivation of both 6 and 7 involve the substitution $y = Ux$ where U is a unitary matrix whose first row is the $u^T/||u||$. 7 also uses completing the square within the exponent.

$$M_{X|u^T X > c}(t) = \frac{\sum_{i=1}^{k} \tau_i m_i(t)}{\sum_{i=1}^{k} \tau_i q_i} \tag{8}$$

where X is in the halfspace $u^T X > c$.

Now we can take derivatives of the MGF to get moments, here we find the first two moments. With $s_i = \mu_i^T t + \frac{1}{2}t^T \Sigma_i t$ and $z_i(t) = \frac{c - u^T(\mu_i + \Sigma_i t)}{||D_i^{1/2} U_i^T u|| \sqrt{2}}$:

$$\bigtriangledown_t m_i(t) = \frac{1}{2} e^{s_i} (erfc(z_i)(\mu_i + \Sigma_i t) + \sqrt{\frac{2}{\pi}} e^{-z_i^2} \frac{\Sigma_i u}{||D_i^{1/2} U_i^T u||}) \tag{9}$$

$$\bigtriangledown_t m_i(t)|_{t=0} = \frac{1}{2} (erfc(z_i(0))\mu_i + \sqrt{\frac{2}{\pi}} e^{-z_i(0)^2} \frac{\Sigma_i u}{||D_i^{1/2} U_i^T u||}) \tag{10}$$

And,

$$E[X|u^T X > c] = \frac{\sum\limits_{i=1}^{k} \tau_i \bigtriangledown_t m_i(t)}{\sum\limits_{i=1}^{k} \tau_i q_i} \tag{11}$$

And the second moment is:

$$H_t(m_i(t)) = \frac{1}{2} e^{s_i} [erfc(z_i)((\mu_i + \Sigma_i t)(\mu_i + \Sigma_i t)^T + \Sigma_i) + \\ \sqrt{\frac{2}{\pi}} \frac{e^{-z_i^2}}{||D_i^{1/2} U_i^T u||} (\Sigma_i u (\mu_i + \Sigma_i t)^T + (\mu_i + \Sigma_i t)(\Sigma_i u)^T + \frac{z_i \sqrt{2}}{||D_i^{1/2} U_i^T u||} \Sigma_i u u^T \Sigma_i)] \tag{12}$$

$$H_t(m_i(t))|_{t=0} = \frac{1}{2} [erfc(z_i(0))(\mu_i \mu_i^T + \Sigma_i) + \\ \sqrt{\frac{2}{\pi}} \frac{e^{-z_i(0)^2}}{||D_i^{1/2} U_i^T u||} (\Sigma_i u \mu_i^T + \mu_i u^T \Sigma_i + \frac{z_i(0)\sqrt{2}}{||D_i^{1/2} U_i^T u||} \Sigma_i u u^T \Sigma_i)] \tag{13}$$

And,

$$Var[X|u^T X > c] = \frac{\sum\limits_{i=1}^{k} \tau_i H_t(m_i(t))}{\sum\limits_{i=1}^{k} \tau_i q_i} - E[X|u^T X > c](E[X|u^T X > c])^T \tag{14}$$

The previous results applied to the half space defined by $u^T X > c$, finding the results for the other half space is easy, recall that m_i and q_i are integrals over a half space:

$$g(t; u, c) = \int_{u^T x = c}^{\infty} f(x, t) dx \tag{15}$$

Let's look at the same integral over the other half space:

$$\int_{-\infty}^{u^T x = c} f(x, t) dx = \int_{-u^T x = -c}^{\infty} f(x, t) dx = g(t; -u, -c) \tag{16}$$

Thus, we can find the MGF of the other half space by plugging in -u and -c. Further, we take derivatives with respect to t to find moments so, again, plugging in -u and -c will give the complementary moments. We denote these complementary values with a prime (') i.e. since $q_i = Pr(u^T X > c)$, we say $q_i' = Pr(u^T X < c)$

Lastly, note that in the k-means framework, the expected loss given that the decision boundary is defined by u and c is:

$$E[J|u,c] = (\sum_{i=1}^{k} \tau_i q_i) * tr(Var[X|u^T X > c]) + (\sum_{i=1}^{k} \tau_i q_i') * tr(Var[X|u^T X < c]) \tag{17}$$

3 Simplest Case

Let us try to solve our initial problem ($D_k^* = D_B^*$) where we are in 1-D, and k=2 with $\tau_1 = \tau_2 = 1/2$, $\mu_1 = -m$, $\mu_2 = m$, and $\Sigma_1 = \Sigma_2 = 1$ and our decision boundary is at c. In this case:

3.1 Region $X > c$

$$z_1 = \frac{c+m}{\sqrt{2}}$$

$$z_2 = \frac{c-m}{\sqrt{2}}$$

$$q_1 = \frac{1}{2} erfc(z_1)$$

$$q_2 = \frac{1}{2} erfc(z_2)$$

$$\nabla_t m_1|_{t=0} = \frac{1}{2}(-erfc(z_1) * m + \sqrt{\frac{2}{\pi}} e^{-z_1^2})$$

$$\nabla_t m_2|_{t=0} = \frac{1}{2}(erfc(z_2) * m + \sqrt{\frac{2}{\pi}} e^{-z_2^2})$$

$$E[X|X > c] = \frac{m * (erfc(z_2) - erfc(z_1)) + \sqrt{\frac{2}{\pi}} * (e^{-z_1^2} + e^{-z_2^2})}{2(q_1 + q_2)}$$

$$H_t(m_1(t))|_{t=0} = \frac{1}{2}[erfc(z_1)(m^2+1) + \sqrt{\frac{2}{\pi}} e^{-z_1^2}(-2m + z_1\sqrt{2})]$$

$$H_t(m_2(t))|_{t=0} = \frac{1}{2}[erfc(z_2)(m^2+1) + \sqrt{\frac{2}{\pi}} e^{-z_2^2}(2m + z_2\sqrt{2})]$$

$$Var(X|X > 0) = \frac{H_t(m_1(t))|_{t=0} + H_t(m_2(t))|_{t=0}}{q_1 + q_2} - E[X|X > c]^2$$

3.2 Region $X < c$

All we need to do is define

$$z_1' = \frac{-c+m}{\sqrt{2}} = -z_2$$

$$z_2' = \frac{-c-m}{\sqrt{2}} = -z_1$$

Then the rest of the equations are identical to those in section 3.1 after we replace z_1 with z_1' and z_2 with z_2'.

3.3 K-means Cost

Here we use equation 2 with the decision rule:

$$D(x) = \begin{cases} 1 & x < c \\ 2 & x \geq c \end{cases}$$

$$\begin{aligned} J_D = Pr(X < c) * Var(X|X < c) + Pr(X > c) * Var(X|X > c) = \\ \frac{1}{2}(q_1' + q_2')Var(X|X < c) + \frac{1}{2}(q_1 + q_2)Var(X|X > c) = \\ m^2 + 1 - \frac{1}{2}((q_1' + q_2')E[X|X < c]^2 + (q_1 + q_2)E[X|X > c]^2) \end{aligned} \tag{18}$$

So, we seek,

$$c^* = \operatorname*{argmin}_c J_D$$

We should expect (and surely hope) that $c^* = 0$, the Bayes optimal decision rule. Below is a graph of the cost function at varying c, giving evidence that the optimal k means rule indeed recovers the Bayes optimal rule.

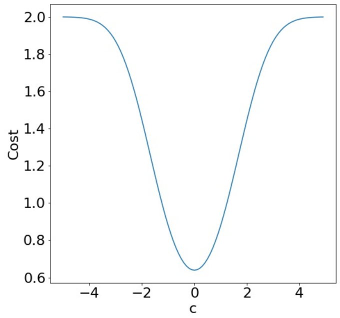

Figure 1: Simple Case with m=1: Cost function at different decision rules.

Returning to the cost function:

$$c^* = \underset{c}{\operatorname{argmin}}\, J_D = \underset{c}{\operatorname{argmax}}\, (q_1' + q_2')E[X|X < c]^2 + (q_1 + q_2)E[X|X > c]^2$$

This is a sum of two terms that each depend on c. When we revisit the equations from Section 3.1, we notice the two terms are identical, except for the sign change in c.

$$f(c) = (q_1' + q_2')E[X|X < c]^2 \Rightarrow f(-c) = (q_1 + q_2)E[X|X > c]^2$$

i.e.

$$c^* = \underset{c}{\operatorname{argmax}}\, f(c) + f(-c)$$

where $f(c)$ is a differentiable function. Clearly, $f(c) + f(-c)$ has a local extremum at $c = 0$, so we just need to consider whether this local extremum is a global extremum.