Statistical learning theory

Carlos Cotrini
Department of Computer Science
ETH Zürich
Switzerland

February 2, 2021

Chapter 1

Deterministic annealing

1.1 The problem of centroid-based clustering

In this chapter, we consider the problem of centroid-based clustering. In this problem, an observation is a sample $X = \{x_1, \ldots, x_N\} \subseteq \mathbb{R}^d$ of points and our objective is to compute a set $\{\theta_1, \ldots, \theta_K\} \subseteq \mathbb{R}^d$ of K centroids and a cluster assignment function $c : \mathbb{R}^d \to \{1, \ldots, K\}$, where $K \in \mathbb{N}$ is a hyperparameter. We let the hypothesis class $\mathcal{C}$ be all possible cluster assignment functions with fixed K.

As cost function, we choose the K-means cost function:

$$R(c, \theta, X) := \sum_{i \leq n} \left\| x_i - \theta_{c(i)} \right\|^2 . \tag{1.1}$$

Note that, in contrast to K-means, we neither require θ_k to be the mean of all points in X assigned to cluster k nor that $c(x)$ is the index of the centroid closest to x.

1.2 Chapter outline

In this chapter, we study how to apply simulated annealing to the problem of clustering, and derive a simplified version called deterministic annealing, proposed by Kenneth Rose [**?**]. We compare deterministic annealing against a popular ERM approach, K-means, and empirically show that deterministic annealing often provides solutions that generalize better than K-means. Furthermore, combining deterministic annealing with posterior agreement gives a method to estimate the number of clusters while executing deterministic annealing.

Deterministic annealing

Deterministic annealing originates from applying simulated annealing to clustering. We can summarize simulated annealing in this context in the following three steps.

1. Pick a first model c and an initial (high) temperature.
2. Use MCMC to draw a sample from $p(\cdot \mid X)$, the Gibbs distribution induced by $R(\cdot, \cdot, X)$, starting from c as the first model in the sample.
3. Decrease the temperature, let c be the last model in the MCMC sample, and go to Step 2.

We see in Section **??** that the Gibbs distribution induced by the cost function $R(\cdot, \cdot, X)$ is actually tractable. This means that we do not need to do MCMC sampling. We can instead directly computing the Gibbs distribution. The three steps become the following procedure, called *deterministic annealing*:

1. Pick an initial set of centroids and an initial (high) temperature.
2. Compute the Gibbs distribution $p(\cdot \mid X)$, for θ arbitrary.
3. Decrease the temperature and go to Step 2.

In Section **??**, we show that the centroids can then be computed as those that maximize the Gibbs distribution's entropy. However, this maximization is intractable, so we use the EM algorithm to efficiently compute an approximation. We combine all these insights to formulate the algorithm for deterministic annealing in Section 1.5.

In Section 1.6, we provide some experimental results comparing deterministic annealing against K-means.

In Section **??**, we describe an interesting behavior of deterministic annealing. When the temperature is high, there is only one cluster containing all points and whose centroid is just the sample mean. As the temperature decreases, this cluster decomposes into several clusters that continue to decompose as the temperature keeps decreasing. One could argue that at a sufficiently low temperature, each point becomes its own cluster.

Finally, in Section 1.8, we exploit this behavior and posterior agreement to determine the number of clusters while executing deterministic annealing. This yields an extra advantage of deterministic annealing over other standard clustering methods like K-means which demand to set the number of clusters in advance.

1.3 Tractability of the Gibbs distribution

We assume that our hypothesis class $\mathcal{C}$ is the set of all cluster assignments $c : \{1, \ldots, N\} \to \{1, \ldots, K\}$.ToDo: Observe that c's domain must be the set of numbers up to N and not an Euclidean domain. Otherwise, the Gibbs distribution is a continuous distribution! We define then the family of Gibbs distributions induced by the K-means cost function as all distributions over $\mathcal{C}$ of the form

$$p(\cdot \mid \theta, X) \propto \exp\left(-\frac{1}{T}R(c, \theta, X)\right). \tag{1.2}$$

Here, $T > 0$ is a hyper-parameter denoting the temperature and $\theta \in \mathbb{R}^{K \times d}$ are parameters denoting the cluster centroids.

The centroids are chosen to be hyper-parameters just for convenience. One could treat them as part of the hypothesis class, but this substantially complicates the analysis.

Theorem 1. *A Gibbs distribution induced by the K-means cost function factorizes as follows:*

$$p(c \mid \theta, X) = \prod_{i \leq N} p(c(i) \mid \theta, X), \tag{1.3}$$

where

$$p(c(i) \mid \theta, X) \propto \exp\left(-\frac{1}{T}\left\|x_i - \theta_{c(i)}\right\|^2\right). \tag{1.4}$$

Proof. The Gibbs distribution is

$$p(c \mid \theta, X) = \frac{\exp\left(-\frac{1}{T}\sum_{i\leq N}\left\|x_i - \theta_{c(i)}\right\|^2\right)}{\sum_{c\in\mathcal{C}}\exp\left(-\frac{1}{T}\sum_{i\leq N}\left\|x_i - \theta_{c(i)}\right\|^2\right)}. \tag{1.5}$$

The numerator can be rewritten as follows:

$$\exp\left(-\frac{1}{T}\sum_{i\leq N}\left\|x_i - \theta_{c(i)}\right\|^2\right) = \prod_{i\leq N}\exp\left(-\frac{1}{T}\left\|x_i - \theta_{c(i)}\right\|^2\right). \tag{1.6}$$

We now apply the combinatorial trick that we studied in the lecture to rewrite the denominator. Unfortunately, I do not see a better way to explain this trick without the diagrams from the lecture.

$$\sum_{c\in\mathcal{C}}\exp\left(-\frac{1}{T}\sum_{i\leq N}\left\|x_i - \theta_{c(i)}\right\|^2\right) = \ldots = \prod_{i\leq N}\sum_{k\leq K}\exp\left(-\frac{1}{T}\left\|x_i = \theta_i\right\|^2\right). \tag{1.7}$$

Putting these results together yields that

$$\begin{aligned} p(c \mid \theta, X) &= \frac{\prod_{i\leq N}\exp\left(-\frac{1}{T}\left\|x_i - \theta_{c(i)}\right\|^2\right)}{\prod_{i\leq N}\sum_{k\leq K}\exp\left(-\frac{1}{T}\left\|x_i - \theta_k\right\|^2\right)} && (1.8)\\ &= \prod_{i\leq N}\frac{\exp\left(-\frac{1}{T}\left\|x_i - \theta_{c(i)}\right\|^2\right)}{\sum_{k\leq K}\exp\left(-\frac{1}{T}\left\|x_i - \theta_k\right\|^2\right)} && (1.9)\\ &= \prod_{i\leq N} p(c(i) \mid \theta, X). && (1.10) \end{aligned}$$

□

Theorem 1 shows that we can compute $p(c \mid \theta, X)$ in just $O(NK)$-time. You just have to compute $\exp\left(-\frac{1}{T}\left\|x_i - \theta_k\right\|^2\right)$, which is $O(1)$-time, for $i \leq N$ and $k \leq K$. Hence, computing $p(c(i) \mid \theta, X)$ takes $O(K)$-time, for $i \leq N$, yielding a total computing time of $O(NK)$.

1.4 Computing the centroids

The calculations from the previous section do not tell us the values of the centroids. Remember that we follow the maximum-entropy principle, so we compute the centroids that maximize the entropy of $p(\cdot \mid \theta, X)$. In this section, we use C to denote a random cluster assignment whose distribution is $p(\cdot \mid \theta, X)$.

We assume here that $\mathbb{E}_{C \sum p(\cdot|\theta,X)} = \mu$, for some fixed $\mu \in \mathbb{R}^+$ and that the temperature hyper-parameter of $p(\cdot \mid \theta, X)$ depends exclusively on μ.

Lemma 1. *If, for a set θ^* of centroids,*

$$\frac{\partial H\left[p(\cdot \mid \theta, X)\right]}{\partial \theta}\Big|_{\theta^*} = 0, \tag{1.11}$$

then

$$\mathbb{E}_{C \sim p(\cdot|\theta,X)}\left[\frac{\partial}{\partial \theta} R(C, \theta, X)\Big|_{\theta^*}\right] = 0. \tag{1.12}$$

Proof. We leave the proof details as an exercise and provide just a vague proof. For convenience, all expectations are with respect to a random cluster assignment $C \sim p(\cdot \mid \theta, X)$.

First, show that

$$H\left[p(\cdot \mid \theta, X)\right] = \frac{1}{T}\mathbb{E}\left[R(C, \theta, X)\right] + \mathbb{E}\left[\log \sum_{c \in \mathcal{C}} \exp\left(-\frac{1}{T} R(c, \theta, X)\right)\right]. \tag{1.13}$$

Recall that $\mathbb{E}\left[R(C, \theta, X)\right]$ is fixed to be constant, by construction. Hence, the first term on the right-hand side of the equation above is constant with respect to θ. Moreover, the second term does not depend on the random variable C. We get then that

$$H[p(\cdot \mid \theta, X)] = \text{const} + \log \sum_{c \in \mathcal{C}} \exp\left(-\frac{1}{T} R(c, \theta, X)\right). \tag{1.14}$$

Take the derivative with respect to on both sides and show that

$$\frac{\partial H[p\left(\cdot \mid \theta, X\right)]}{\partial \theta} = \mathbb{E}_{C \sim p(\cdot|\theta^*,X)}\left[\frac{\partial}{\partial \theta} R(C, \theta, X)\right]. \tag{1.15}$$

This equality yields the desired result. □

For an event A, we now denote by $\mathbf{1}A$ the indicator function. That is,

$$\mathbf{1}A(\omega) \begin{cases} 1 & \text{if } \omega \in A \text{ and} \\ 0 & \text{otherwise.} \end{cases} \tag{1.16}$$

Lemma 2. *Let C denote a random cluster assignment and $C(i)$, for $i \leq N$, the cluster that C assigns to x_i. If $R(\cdot, \cdot, X)$ is the K-means cost function, then*

$$\frac{\partial}{\partial \theta_k} R(C, \theta, X) = 2\theta_k \sum_{i \leq N} \mathbf{1}\left\{C(i) = k\right\} - 2 \sum_{i \leq N} x_i \mathbf{1}\left\{C(i) = k\right\}. \tag{1.17}$$

Proof. The proof is a straightforward use of vector calculus and is left as an exercise. □

Theorem 2. *The set θ^* of centroids that maximize the entropy of $p(\cdot \mid \theta, X)$ must satisfy the following conditions.*

$$\theta_k^* = \frac{\sum_{i \leq N} x_i \mathbf{P}\left(C(i) = k \mid \theta^*, X\right)}{\sum_{i \leq N} \mathbf{P}\left(C(i) = k \mid \theta^*, X\right)}, \text{ for } k \leq K, \tag{1.18}$$

where

$$\mathbf{P}\left(C(i) = k \mid \theta^*, X\right) \propto \exp\left(-\frac{1}{T}\left\|x_i - \theta_k^*\right\|^2\right). \tag{1.19}$$

Proof. If θ^* maximizes the entropy of $p(\cdot \mid \theta, X)$, then, by Lemma 1,

$$\mathbb{E}_{C \sim p(\cdot \mid \theta^*, X)}\left[\frac{\partial}{\partial \theta} R(C, \theta, X)\big|_{\theta^*}\right] = 0. \tag{1.20}$$

Plugging Equation 1.17 in the equation above yields that

$$\mathbb{E}_{C \sim p(\cdot \mid \theta^*, X)}\left[\theta_k^* \sum_{i \leq N} \mathbf{1}\left\{C(i) = k\right\} - \sum_{i \leq N} x_i \mathbf{1}\left\{C(i) = k\right\}\right] = 0. \tag{1.21}$$

Apply now linearity of the expectation and the fact that $\mathbb{E}\left[\mathbf{1}A\right] = \mathbf{P}\left(A\right)$ to get that

$$\theta_k^* \sum_{i \leq N} \mathbf{P}\left(C(i) = k \mid \theta^*, X\right) - \sum_{i \leq N} x_i \mathbf{P}\left(C(i) = k \mid \theta^*, X\right) = 0. \tag{1.22}$$

The desired condition follows from this equation. □

Unfortunately, Equation **??** does not give us a closed formula to compute θ^*. We can still attempt to estimate θ^* iteratively, with the following procedure.

1. Set $t \leftarrow 0$ and θ_t^* to an arbitrary value.

2. Set $t \leftarrow t+1$ and

$$\theta_{t+1}^* \leftarrow \frac{\sum_{i \leq N} \mathbf{P}\left(C(i) = k \mid \theta_t^*, X\right) x_i}{\mathbf{P}\left(C(i) = k \mid \theta_t^*, X\right)}. \tag{1.23}$$

This procedure converges to a local maximum of $H[p(\cdot \mid \theta, X)]$.

Exercise 1. *Demonstrate that the procedure above also results from applying the EM-algorithm to the following maximization problem:*

$$\max_\theta \log \sum_{c \in \mathcal{C}} p(X, c \mid \theta), \tag{1.24}$$

where $p(X, c \mid \theta) := p(c \mid \theta, X)p(X)$ and $p(X)$ is the pdf of the phenomenon where X comes from. Use this to demonstrate why the procedure above converges.

1.5 Deterministic annealing

We now collect all insights from this chapter and use them to propose an improved version of simulated annealing, called *deterministic annealing*. Algorithm **??** provides the details. The idea of this procedure is the following.

1. We set a high value for the temperature and start with arbitrary centroids.

2. We alternate between (i) computing the maximum-entropy distribution for a random cluster assignment while fixing the centroids and (ii) computing the centroids that maximize the entropy of that distribution. This alternation is repeated until convergence of the centroids.

3. We reduce the temperature and repeat Step 2. To avoid having repeated centroids, we add a small amount of noise.

We remark some differences between deterministic annealing and simulated annealing:

- There is no MCMC sampling. This is because the Gibbs distribution induced by the K-means cost is tractable. So we can just compute it directly.
- The centroids are treated as parameters of the Gibbs distribution and not as part of the hypothesis class. This is done mainly for convenience. A Gibbs distribution that also treats the centroids as random variables makes the whole procedure intractable.

1.6 Experimental results

1.7 Phase transitions

1.8 Deterministic annealing and posterior agreement