Foundations of Data Science and Machine Learning – *Homework 3*

Isaac Martin

Last compiled February 24, 2023

EXERCISE 2. Let $\mathbf{X} \in \mathbb{R}^{n\times d}$ be a matrix whose n rows are the data points $\mathbf{x}_1, ..., \mathbf{x}_n \in \mathbb{R}^d$, and let $\mathcal{X} = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$. Consider the k-means optimization problem: find a partition $C_1, ..., C_k$ which minimizes, among all partitions of $[n]$ into k subsets,

$$\operatorname{cost}_{\mathcal{X}}(C_1, ..., C_k) := \sum_{j=1}^{k} \sum_{i\in C_j} \left\| \mathbf{x}_i - \frac{1}{|C_j|} \sum_{i\in C_j} \mathbf{x}_i \right\|_2^2 .$$

(a) Suppose that $\Phi : \mathbb{R}^d \longrightarrow \mathbb{R}^r$ with $r = \mathcal{O}(\log(n)/\varepsilon^2)$ is a random i.i.d. spherical Gaussian projection matrix and thus satisfies the JL lemma. Consider the projected points $\mathbf{y}_j = \Phi\mathbf{x}_j \in \mathbb{R}^r$ and suppose $\tilde{C}_1, ..., \tilde{C}_k$ are an optimal set of k-means clusters for the data points $\mathcal{Y} = \{\mathbf{y}_1, ..., \mathbf{y}_n\}$. That is,

$$\operatorname{cost}_{\mathcal{Y}}(\tilde{C}_1, ..., \tilde{C}_n) = \min_{C_\bullet} \operatorname{cost}_{\mathcal{Y}}(C_1, ..., C_k)$$

where the minimization is over all partitions of $\mathcal{Y}$ into k subsets. Show that the clusters $\tilde{C}_1, ..., \tilde{C}_k$ also represent a good clustering for the original dataset $\mathcal{X}$ in the sense that with high probability

$$\operatorname{cost}_{\mathcal{X}}(\tilde{C}_1, ..., \tilde{C}_k) \le (1+\varepsilon) \min_{C_1,...,C_k} \operatorname{cost}_{\mathcal{X}}(C_1, ..., C_k).$$

(b) Suppose we now project the points $\mathbf{x}_j$ to k dimensions using the SVD of $\mathbf{X}$. Let $\mathbf{V}_k \in \mathbb{R}^{d\times k}$ be the matrix whose columns are the first right singular vectors of $\mathbf{X}$. Suppose the $\tilde{C}_1, ..., \tilde{C}_k$ are the optimal k-means clusters for the points $\mathbf{V}_k^\top \mathbf{x}_1, ..., \mathbf{V}_k^\top \mathbf{x}_n$.
Show that the clusters $\mathbf{C}_1, ..., \mathbf{C}_k$ also represent a good clustering for the original dataset $\mathcal{X}$, in the sense that

$$\operatorname{cost}_{\mathcal{X}}(\tilde{C}_1, ..., \tilde{C}_k)) \le 2 \min_{C_1,...,C_k} \operatorname{cost}_{\mathcal{X}}(C_1, ..., C_k).$$

Proof:

(a) First consider a fixed $j \in \{1, ..., k\}$ and the following expression:

$$\sum_{i,\ell\in C_j} \|\mathbf{x}_i - \mathbf{x}_j\|_2^2 = \sum_{i\in C_j}\sum_{\ell\in C_j} \|\mathbf{x}_i - \mathbf{x}_j\|_2^2.$$

Letting $\mu_j = \frac{1}{|C_j|}\sum_{i\in C_j} \mathbf{x}_i$ denote the centroid of $\{\mathbf{x}_i\}_{i\in C_j}$, we see that

$$\begin{aligned}\sum_{i\in C_j}\sum_{\ell\in C_j} \|\mathbf{x}_i - \mathbf{x}_j\|_2^2 &= \sum_{i\in C_j}\sum_{\ell\in C_j} \|\mathbf{x}_i - \mu_j + \mu_j - \mathbf{x}_j\|_2^2 \\ &= \sum_{i\in C_j}\left(\sum_{\ell\in C_j} \|\mu_j - \mathbf{x}_\ell\|_2^2 \;+\; 2(\mathbf{x}_i - \mu_j)\cdot \sum_{\ell\in C_j}(\mu_j - \mathbf{x}_\ell) \;+\; |C_j|\cdot\|\mathbf{x}_i - \mu_j\|_2^2 \right).\end{aligned}$$

The second equality above follows from the fact that

$$\sum_{i=1}^{n} \|\mathbf{a}_i - \mathbf{c} + \mathbf{c} - \mathbf{x}\|^2 = \sum_{i=1}^{n} \|\mathbf{a}_i - \mathbf{c}\| \; + \; 2(\mathbf{c} - \mathbf{x}) \cdot \sum_i (\mathbf{a}_i - \mathbf{c}) \; + \; n \cdot \|\mathbf{c} - \mathbf{x}\|^2,$$

which in turn can be derived by writing $\|\mathbf{a}_i - \mathbf{c} + \mathbf{c} - \mathbf{x}\|^2 = \langle \mathbf{a}_i - \mathbf{c} + \mathbf{c} - \mathbf{x}, \mathbf{a}_i - \mathbf{c} + \mathbf{c} - \mathbf{x} \rangle$, expanding by bilinearity and gathering up terms in a clever way. Using the fact that indexing over ℓ and i is equivalent together with the bilinearity properties of the inner product, we can continue our above chain of equalities to get that

$$\begin{aligned}
&\sum_{i,\ell \in C_j} \|\mathbf{x}_i - \mathbf{x}_j\|_2^2 \\
&\qquad = \sum_{i \in C_j} \left(\sum_{\ell \in C_j} \|\mu_j - \mathbf{x}_\ell\|_2^2 \; + \; 2(\mathbf{x}_i - \mu_j) \cdot \sum_{\ell \in C_j} (\mu_j - \mathbf{x}_\ell) \; + \; |C_j| \cdot \|\mathbf{x}_i - \mu_j\|_2^2 \right) \\
&\qquad = |C_j| \cdot \sum_{\ell \in C_j} \; + \; 2 \left(\sum_{\ell \in C_j} (\mu_j - \mathbf{x}_\ell) \right) \cdot \sum_{i \in C_j} (\mathbf{x}_i - \mu_j) \; + \; |C_j| \cdot \sum_{i \in C_j} \|\mathbf{x}_i - \mu_j\|_2^2 \\
&\qquad = 2|C_j| \cdot \sum_{i \in C_j} \|\mathbf{x}_i - \mu_j\|_2^2 \; - \; 2 \left\| \sum_{\ell \in C_j} (\mathbf{x}_i - \mu_j) \right\|_2^2 .
\end{aligned}$$

The term $\sum_{\ell \in C_j} (\mathbf{x}_i - \mu_j)$ is 0 because μ_j is the centroid of $\{\mathbf{x}_i\}_{i \in C_j}$, hence

$$\sum_{i,\ell \in C_j} \|\mathbf{x}_i - \mathbf{x}_j\|_2^2 = 2|C_j| \cdot \sum_{i \in C_j} \|\mathbf{x}_i - \mu_j\|_2^2 = 2|C_j| \cdot \sum_{i \in C_j} \left\| \mathbf{x}_i - \frac{1}{|C_j|} \sum_{\ell \in C_j} \mathbf{x}_\ell \right\|_2^2 .$$

Taking sums over all $j \in \{1, ..., k\}$, we then get that

$$\begin{aligned}
\mathrm{cost}_{\mathcal{X}}(C_1, ..., C_k) &= \sum_{j=1}^{k} \sum_{i \in C_j} \left\| \mathbf{x}_i - \frac{1}{|C_j|} \sum_{i \in C_j} \mathbf{x}_i \right\|_2^2 \\
&= \sum_{j=1}^{k} \frac{1}{2|C_j|} \sum_{i,\ell \in C_j} \|\mathbf{x}_i - \mathbf{x}_\ell\|_2^2,
\end{aligned}$$

as suggested by the hint. This form of the cost function integrates more favorably with the properties of the Johnson-Lindenstrauss theorem, since for $r > C \cdot \frac{\log(n/\delta)}{\varepsilon^2}$, we get that

$$\left\|\mathbf{y}_i - \mathbf{y}_j\right\|_2^2 = \|\Phi(\mathbf{x}_i - \mathbf{x}_\ell)\|_2^2 \leq (1 + \varepsilon) \|\mathbf{x}_i - \mathbf{x}_\ell\|_2^2$$

occurs with probability at least $1 - \delta$ and therefore

$$\begin{aligned}
\mathrm{cost}_{\mathcal{Y}}(C_1, ..., C_k) &= \sum_{j=1}^{k} \frac{1}{2|C_j|} \sum_{i,\ell \in C_j} \|\mathbf{y}_i - \mathbf{y}_\ell\|_2^2 \\
&\leq (1 + \varepsilon) \sum_{j=1}^{k} \frac{1}{2|C_j|} \sum_{i,\ell \in C_j} \|\mathbf{x}_i - \mathbf{x}_\ell\|_2^2 = (1 + \varepsilon) \, \mathrm{cost}_{\mathcal{X}}(C_1, ..., C_k)
\end{aligned}$$

also occurs with probability at least $1-\delta$ for any partition $C_1 \sqcup ... \sqcup C_k = \{1..n\}$. Combining this with the other bound from the JL theorem we have that

$$(1-\varepsilon)\,\mathrm{cost}_{\mathcal{X}}(C_1,...,C)k \le \mathrm{cost}_{\mathcal{Y}}(C_1,...,C_k) \le (1+\varepsilon)\,\mathrm{cost}_{\mathcal{X}}(C_1,...,C_k)$$

for all partitions $C_\bullet$ of $\mathcal{X}$. Since this holds for all partitions of $\mathcal{X}$ it also holds for the partition $\tilde{C}_\bullet$ which minimizes $\mathrm{cost}_{\mathcal{Y}}$, hence

$$(1-\varepsilon)\,\mathrm{cost}_{\mathcal{X}}(\tilde{C}_1,...,\tilde{C}_k) \le \mathrm{cost}_{\mathcal{Y}}(\tilde{C}_1,...,\tilde{C}_k) \le \mathrm{cost}_{\mathcal{Y}}(C_1,...,C_k) \le (1+\varepsilon)\,\mathrm{cost}_{\mathcal{X}}(C_1,...,C_k).$$

We then have that

$$\mathrm{cost}_{\mathcal{X}}(\tilde{C}_1,...,\tilde{C}_k) \le \frac{1+\varepsilon}{1-\varepsilon}\,\mathrm{cost}_{\mathcal{Y}}(C_1,...,C_k).$$

This is not quite what we want. However, we chose $r = \mathcal{O}(\log(n)/\varepsilon^2)$, which only means that $r = C\log(n)/\varepsilon^2$ for some C. Rescale C and choose a new $\varepsilon' \in (0,1)$ so that

$$r = 9C \cdot \frac{\log(n)}{\varepsilon'^2} \implies \varepsilon = 3\sqrt{\frac{C\log(n)}{r}} = 3\varepsilon.$$

If our original ε was less than $1/3$, then

$$\begin{aligned} 1-3\varepsilon > 0 &\iff 0 < \varepsilon(1-3\varepsilon) \\ &\iff 1+\varepsilon < 1+3\varepsilon-\varepsilon-3\varepsilon^2 \\ &\iff \frac{1+\varepsilon}{1-\varepsilon} < 1+3\varepsilon = 1+\varepsilon'. \end{aligned}$$

Thus, for ε', we have

$$\mathrm{cost}_{\mathcal{X}}(\tilde{C}_1,...,\tilde{C}_k) \le \frac{1+\varepsilon}{1-\varepsilon}\,\mathrm{cost}_{\mathcal{Y}}(C_1,...,C_k) \le (1+\varepsilon')\,\mathrm{cost}_{\mathcal{Y}}(C_1,...,C_k).$$

This is perfectly fine, since the change $\varepsilon \longrightarrow \varepsilon'$ corresponds to scaling r by r, and hence we still have that $r = \mathcal{O}(\log(n)/\varepsilon^2)$ for our original choice of ε. This gives us the desired result.

(b) We first prove that $\mathrm{cost}_{\mathcal{X}}(C_1,...,C_k) = \|\mathbf{X} - MM^\top\mathbf{X}\|_F^2$ where $M \in \mathbb{R}^{n\times k}$is defined by $M_{ij} = \frac{1}{\sqrt{|C_j|}}$ if $i \in C_j$ and 0 otherwise. Notice that each row of M has only one nonzero element at index (i,j) where $i \in C_j$, and hence

$$[MM^\top]_{ij} = [M]_{i,\bullet} \cdot [M]_{j,\bullet} = \begin{cases} \frac{1}{|C_{\ell_i}} & i,j \in C_{\ell_i} \\ 0 & \text{else} \end{cases}$$

for some $\ell_i = 1,...,k$. In particular, each diagonal element $[MM^\top]_{ii}$ is nonzero and is equal to $1/|C_{\ell_i}|$ where $i \in C_{\ell_i}$. Thus, when we multiply a row $[MM^\top]_{i,\bullet}$ by a column $[\mathbf{X}]_{\bullet,j}$ of $\mathbf{X}$, the result is a sum

$$[MM^\top\mathbf{X}]_{ij} = \frac{1}{|C_{\ell_i}}\sum_{a\in C_{\ell_i}} \mathbf{x}_a^j$$

where C_{ℓ_i} is the partition containing i and $\mathbf{x}_a^j$ is the jth term in the data point $\mathbf{x}_a$. That is, $[MM^\top \mathbf{X}]_{ij}$ is the sum of the jth components of all data points belonging to C_{ℓ_i} scaled by $1/|C_{\ell_i}|$. The ith row of $MM^T\mathbf{X}$ is therefore the centroid of the ℓ_ith cluster $\{\mathbf{x}_j\}_{j\in C_{\ell_i}}$. Denoting by μ_{ℓ_i} the ℓ_ith centroid, we see that

$$\begin{aligned}\|\mathbf{X} - MM^\top \mathbf{X}\|_F^2 &= \sum_{i=1}^n \|[\mathbf{X} - MM^\top \mathbf{X}]_{i,\bullet}\|^2 \\ &= \sum_{i=1}^n \|\mathbf{x}_i - \mu_{\ell_i}\|^2 \\ &= \sum_{j=1}^k \sum_{i\in C_j} \|\mathbf{x}_i - \mu_i\|^2 = \mathrm{cost}_{\mathcal{X}}(C_1, ..., C_k).\end{aligned}$$

This gives us yet another expression for the k-means cost function.

We now turn to the problem in earnest. Let $\{v_1, ..., v_k, \tilde{v}_{k+1}, ..., \tilde{v}_d\}$ be the extension of $\{v_1, ..., v_k\}$ to a complete orthonormal basis on $\mathbb{R}^d$. Let W be the matrix whose columns are $\tilde{v}_{k+1}, ..., \tilde{v}_d$. Then $V_k \oplus W$ is a $d \times d$ orthogonal matrix, preserves the Frobenius norm, and hence

$$\begin{aligned}\mathrm{cost}_{\mathcal{X}}(C_1, ..., C_k) &= \|\mathbf{X} - MM^\top \mathbf{X}\|_2^2 = \|(\mathbf{X} - MM^\top \mathbf{X})(V_k \oplus W)\|_2^2 \\ &= \|(\mathbf{X}V_k - MM^\top \mathbf{X}V_k) \oplus (\mathbf{X}W - MM^\top \mathbf{X}W)\|_2^2 \\ &= \|(\mathbf{X}V_k - MM^\top \mathbf{X}V_k)\|_2^2 + \|(\mathbf{X}W - MM^\top \mathbf{X}W)\|_2^2\end{aligned}$$

where M is the matrix defined earlier corresponding to the clustering $C_\bullet$. Let

□

Exercise 3. Find the mapping $\varphi(\mathbf{x})$ that gives rise to the polynomial kernel

$$K(\mathbf{x}, \mathbf{y}) = (x_1x_2 + y_1y_2)^2.$$

Proof: Consider the map $\varphi : \mathbb{R}^2 \longrightarrow \mathbb{R}^3$ defined $\varphi(x_1, x_2) = (x_1^2, x_2^2, \sqrt{2}x_1x_2)$. Interestingly, this is similar to the map one considers from a polynomial ring $R[x_1, x_2]$ to its 2nd Veronese subring $R[x_1^2, x_1x_2, x_2^2]$. We then have that

$$\begin{aligned}\varphi(\mathbf{x}) \cdot \varphi(\mathbf{y})^T &= (x_1^2, x_2^2, \sqrt{2}x_1x_2) \cdot (y_1^2, y_2^2, \sqrt{2}y_1y_2) \\ &= x_1^2y_1^2 + x_2^2y_2^2 + 2x_1y_1x_2y_2 \\ &= (x_1y_1 + x_2y_2)^2,\end{aligned}$$

hence φ gives rise to the desired kernel. □

Exercise 4. Consider a Support Vector Machine with "soft margin" constrains which allows for misclassification: Given training data $(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)$ where $\mathbf{x}_j \in \mathbb{R}^d$ and $y_j \in \{-1, +1\}$, the SVM is

$$\min_{\mathbf{w},b,\xi} \ \lambda\|\mathbf{w}\|_2^2 + \frac{1}{n}\sum_{j=1}^n \xi \quad \text{s.t. } y_j \cdot (\langle \mathbf{w}, \mathbf{x}_j\rangle - b) \geq 1 - \xi_j, \quad \xi_j \geq 0.$$

(a) Discuss the relationship between the parameter λ and the allowable misclassification error.

(b) Where does a data point lie relative to where the margin is when $\xi_j = 0$? Is this data point classified correctly?

(c) Where does a data point lie relative to where the margin is when $0 < \xi_j \leq 1$? Is this data point classified correctly?

(d) Where does a data point lie relative to where the margin is when $\xi_j > 1$? Is this data point classified correctly?

Proof:

(a)

(b) When $\xi_j = 0$ we have that $y_j \cdot (\langle \mathbf{w}, \mathbf{w}_j \rangle - b) \geq 1$. This means that the data point is classified correctly *and* is outside the margin.

(c) When $0 < \xi_j \leq 1$, then ξ_j contributes to penalizing the cost function. Because this is still an optimal solution, this means that we cannot make ξ_j smaller, and thus

$$1 - \xi_j \leq y_j \cdot (\langle \mathbf{w}, \mathbf{x}_j \rangle - b \rangle) < 1.$$

(d) If the optimal solution to the cost function includes a value $\xi_j > 1$, then as before, it means we cannot make ξ_j any smaller without adding error. This means that $0 > y_j \cdot (\langle \mathbf{w}, \mathbf{x}_j \rangle - b)$, and hence $\mathbf{x}_j$ is misclassified.

$\square$