CSE 291: Unsupervised learning

Homework 3

This homework is due on Thursday May 4 at 11.59pm.

- All homeworks must be typewritten and uploaded to Gradescope.
- No late homeworks will be accepted.

1. *Mean and median.* One of the most basic tasks in statistics is to summarize a set of observations $x_1, \ldots, x_n \in \mathbb{R}$ by a single number. Two popular choices for this summary statistic are the *median* and the *mean*.

   (a) Let P be a probability distribution on $\mathbb{R}$. A median of P is any number v such that $P((-\infty, v]) \geq 1/2$ and $P([v, \infty)) \geq 1/2$. For finitely many points $x_1, \ldots, x_n$, this translates to the following definition: a median is any v such that at least half the points are $\leq v$ and at least half the points are $\geq v$.

   Show that any median of $x_1, \ldots, x_n$ is a value v that minimizes
   $$L(v) = \sum_{i=1}^{n} |x_i - v|.$$
   You may assume for simplicity that n is odd. *Hint:* show that for any v greater than the median, the function decreases if you move v slightly to the left; while for any v less than the median, the function decreases if you move v slightly to the right. Therefore, such values of v cannot possibly be minimizers of $L(\cdot)$.

   (b) Show that the mean is the value v that minimizes
   $$L(v) = \sum_{i=1}^{n} (x_i - v)^2.$$
   One way to do this is by calculus. For this problem, you should use a different argument: Show by algebraic manipulation that if $\mu = \text{mean}(x_1, \ldots, x_n)$, then for any v,
   $$\sum_i (x_i - v)^2 \ = \ \sum_i (x_i - \mu)^2 + n(\mu - v)^2.$$

   (c) Generalize your proof in (c) to higher-dimensional points, $x_1, \ldots, x_n \in \mathbb{R}^d$. This is an interesting relation: it exactly captures the squared distortion induced by using a location parameter other than the mean.

2. *Hierarchical k-means?* In this problem, we'll see that Ward's method can be viewed as a greedy bottom-up heuristic for the k-means problem.

For any finite set of points $C \subset \mathbb{R}^d$, let mean(C) denote the average,

$$\text{mean}(C) = \frac{1}{|C|} \sum_{x \in C} x$$

and let cost(C) be the k-means cost if C is treated as a single cluster:

$$\text{cost}(C) = \sum_{x \in C} \|x - \text{mean}(C)\|^2.$$

(a) Let $S_1, S_2 \subset \mathbb{R}^d$ consist of m_1 and m_2 points, respectively, with means μ_1 and μ_2. You may assume that these points are all distinct. Let $S = S_1 \cup S_2$; thus S has size $m = m_1 + m_2$. If μ denotes the mean of S, give an expression for μ in terms of μ_1, μ_2, m_1, m_2.

(b) Show that

$$\text{cost}(S) - (\text{cost}(S_1) + \text{cost}(S_2)) = m_1 \|\mu_1 - \mu\|^2 + m_2 \|\mu_2 - \mu\|^2.$$

You may find it useful to use the following relation from an earlier problem: for any subset $C \subset \mathbb{R}^d$ and any point $z \in \mathbb{R}^d$,

$$\sum_{x \in C} \|x - z\|^2 = \sum_{x \in C} \|x - \text{mean}(C)\|^2 + |C| \|z - \text{mean}(C)\|^2.$$

(c) Simplifying further, show that

$$\text{cost}(S) - (\text{cost}(S_1) + \text{cost}(S_2)) = \frac{m_1 m_2}{m_1 + m_2} \|\mu_1 - \mu_2\|^2.$$

Explain how this relates to Ward's method.

(d) With the result from (c) in mind, can you suggest a greedy *top-down* hierarchical clustering algorithm using the k-means cost function?

3. M is a 2×2 real-valued symmetric matrix with eigenvalues $\lambda_1 = 2, \lambda_2 = -1$ and corresponding eigenvectors

$$u_1 = \frac{1}{\sqrt{5}} \begin{pmatrix} 2 \\ 1 \end{pmatrix}, \quad u_2 = \frac{1}{\sqrt{5}} \begin{pmatrix} -1 \\ 2 \end{pmatrix}.$$

(a) What is M?

(b) What are the eigenvalues of the matrix $M + 2I$?

(c) What are the eigenvalues of the matrix $M^2 = MM$?

4. *Linear and affine subspaces.*

(a) The points $(1, 1, 1)$ and $(-1, -1, 1)$ lie in a two-dimensional subspace of $\mathbb{R}^3$. What is the projection of $(2, 4, 5)$ into this subspace? Your answer should be a vector in $\mathbb{R}^3$.

(b) The points $(1, 1, 1)$ and $(-1, -1, 1)$ lie in a one-dimensional affine subspace of $\mathbb{R}^3$. Can you give a simple description of this subspace?

5. *Singular values versus eigenvalues.* Recall from class that any $p \times q$ matrix M (with $p \leq q$, say) can be written in the form:

$$M = \underbrace{\begin{pmatrix} \uparrow & & \uparrow \\ u_1 & \cdots & u_p \\ \downarrow & & \downarrow \end{pmatrix}}_{p \times p \text{ matrix } U} \underbrace{\begin{pmatrix} \sigma_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_p \end{pmatrix}}_{p \times p \text{ matrix } \Lambda} \underbrace{\begin{pmatrix} \longleftarrow & v_1 & \longrightarrow \\ & \vdots & \\ \longleftarrow & v_p & \longrightarrow \end{pmatrix}}_{p \times q \text{ matrix } V^T}$$

where $u_1, \ldots, u_p$ are orthonormal vectors in $\mathbb{R}^p$, $v_1, \ldots, v_p$ are orthonormal vectors in $\mathbb{R}^q$, and $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p \geq 0$ are known as *singular values*. In this problem, we will try to understand these quantities by relating them to eigenvalues and eigenvectors of suitably defined matrices.

(a) What is Mv_i (for $1 \leq i \leq p$)? Express the answer as simply as possible, in terms of the singular values and vectors of M.

(b) What is $M^T u_i$?

(c) What is $M^T M v_i$? And what is $MM^T u_i$?

(d) Notice that MM^T is a symmetric $p \times p$ matrix and therefore has p real eigenvalues. What are its eigenvalues and eigenvectors?

(e) How do the eigenvalues and eigenvectors of $M^T M$ relate to those of MM^T?

(f) Suppose M has rank k. How would this be reflected in the singular values σ_i?

6. A particular 4×5 matrix M has the following singular value decomposition:

$$M = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} 4 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

Find the best rank-2 approximation to M.

7. *Choosing representations for nearest neighbor.* In this problem, we will study how different representations of images can affect the performance of nearest neighbor methods. We will use the CIFAR-10 data set, which has 50,000 training images and 10,000 test images, with ten different classes (`airplane`, `automobile`, `bird`, `cat`, `deer`, `dog`, `frog`, `horse`, `ship`, `truck`). The images are in color, of size 32×32.

We will compare several image representations:

- The raw pixel representation
- Histogram-of-gradients (HoG) features
- The representation obtained by passing the image through a pre-trained convolutional net (VGG) and using one of the last layers (`last-fc`, meaning "last fully-connected layer")
- The representation obtained by passing the image through a pre-trained convolutional net (VGG) and using one of the earlier layers (`last-conv`, meaning "last convolutional layer")
- The representation obtained by using a convolutional net with the same architecture but with *random weights* (and again, with two variants, `last-fc` and `last-conv`)

In each case, the idea is study the classification performance (on the test set) using 1-nearest neighbor on the training data with Euclidean (ℓ_2) distance.

Download `cifar-representations.zip` from the course website. The directory contains a Jupyter notebook, some helper functions, and some data. In the notebook, we have provided code that will extract HOG and neural net features from the CIFAR data; look through it to get a sense of how it works.

(a) What is the dimensionality of each of the representations (raw pixel, HoG, VGG-last-fc, VGG-last-conv)?

(b) Report test accuracies for 1-nearest neighbor classification using the various representations (raw pixel, HoG, VGG-last-fc, VGG-last-conv, random-VGG-last-fc, random-VGG-last-conv).

(c) For the raw pixel representation:

- Show the first five images in the test set whose label is *correctly* predicted by 1-NN, and show the nearest neighbor (in the training set) of each of these images.
- Show the first five images in the test set whose label is *incorrectly* predicted by 1-NN, and show the nearest neighbor (in the training set) of each of the images.

Repeat for the HoG and VGG-last-fc representations.