1.1 Clustering with the $L_1$-norm

Consider the following distance function for vectors in $\mathbb{R}^d$, based on the $L_1$ norm:

$$D(\vec{u}, \vec{v}) = \sum_{\alpha=1}^{d} |u_{\alpha} - v_{\alpha}|.$$

In this problem, you will derive a clustering algorithm based on this distance function, which is somewhat less sensitive to outliers.

(a) Let $\{\vec{x}_n\}_{n=1}^{N}$ denote $N$ inputs and $\{\vec{\mu}_i\}_{i=1}^{k}$ denote $k$ prototypes in $\mathbb{R}^d$. Also, let $y_{in} \in \{0, 1\}$ denote a $k \times N$ binary assignment matrix whose columns sum to one. Consider the clustering cost function:

$$E(y, \mu) = \sum_{in} y_{in} D(\vec{\mu}_i, \vec{x}_n).$$

Show that the prototype $\vec{\mu}_i$ that minimizes this cost function for fixed $y_{in}$ can be interpreted as the elementwise median (not mean) of the inputs assigned to the $i$th cluster.

(b) Sketch an iterative algorithm, analogous to $k$-means clustering, to minimize this cost function.

1.2 Clustering of nonnegative vectors

Consider the following (asymmetric) distance function for vectors in the nonnegative orthant of $\mathbb{R}^d$ (that is, with only nonnegative elements):

$$D(\vec{u}, \vec{v}) = \sum_{\alpha=1}^{d} \left[ u_{\alpha} \log \left( \frac{u_{\alpha}}{v_{\alpha}} \right) - u_{\alpha} + v_{\alpha} \right].$$

In this problem, you will derive a clustering algorithm based on this distance function, which can be viewed as a generalized KL divergence.

(a) Let $\{\vec{x}_n\}_{n=1}^{N}$ denote $N$ inputs and $\{\vec{\mu}_i\}_{i=1}^{k}$ denote $k$ prototypes in $\mathbb{R}^d$. Also, let $y_{in} \in \{0, 1\}$ denote a $k \times N$ binary assignment matrix whose columns sum to one. Consider the clustering cost function:

$$E(y, \mu) = \sum_{in} y_{in} D(\vec{\mu}_i, \vec{x}_n).$$

Show that the prototype $\vec{\mu}_i$ that minimizes this cost function for fixed $y_{in}$ can be interpreted as the elementwise geometric mean of the inputs assigned to the $i$th cluster.

(b) Sketch an iterative algorithm, analogous to $k$-means clustering, to minimize this cost function.
1.3 Gaussian integrals

It is important to be able to work easily (and fearlessly) with multivariate Gaussian distributions. The following problems will give you some useful practice.

(a) Kullback-Leibler divergence

For continuous distributions \( P(x) \) and \( Q(x) \), the Kullback-Leibler divergence is defined as:

\[
\text{KL}(P, Q) = \int dx \ P(x) \log \frac{P(x)}{Q(x)}.
\]

Compute the Kullback-Leibler divergence between two multivariate Gaussian distributions \( P(x) \) and \( Q(x) \) with means \( \mu_1 \) and \( \mu_2 \) and covariance matrices \( \Sigma_1 \) and \( \Sigma_2 \). In particular, show that:

\[
\text{KL}(P, Q) = \frac{1}{2} \left\{ \log \frac{\mid \Sigma_2 \mid}{\mid \Sigma_1 \mid} - d + \text{Tr} \left[ \Sigma_2^{-1} \Sigma_1 \right] + (\mu_1 - \mu_2)^\top \Sigma_2^{-1} (\mu_1 - \mu_2) \right\}.
\]

(b*) Hellinger distance

For continuous distributions \( P(x) \) and \( Q(x) \), the squared Hellinger distance is defined as:

\[
H(P, Q) = \frac{1}{2} \int dx \left( \sqrt{P(x)} - \sqrt{Q(x)} \right)^2.
\]

Compute the squared Hellinger distance between two multivariate Gaussian distributions \( P(x) \) and \( Q(x) \) with means \( \mu_1 \) and \( \mu_2 \) and covariance matrices \( \Sigma_1 \) and \( \Sigma_2 \). In particular, show that:

\[
H(P, Q) = 1 - \frac{\mid \Sigma_1 \mid^{1/4} \mid \Sigma_2 \mid^{1/4}}{\mid \Sigma \mid^{1/2}} \exp \left\{ -\frac{1}{8} (\mu_1 - \mu_2)^\top \Sigma^{-1} (\mu_1 - \mu_2) \right\} \quad \text{where} \quad \Sigma = \frac{\Sigma_1 + \Sigma_2}{2}.
\]

1.4 Relation between EM algorithm and \( k \)-means clustering

Consider a Gaussian mixture model (GMM) with hidden variable \( z \in \{1, 2, \ldots, k\} \) and observed variable \( x \in \mathbb{R}^d \). The mixture component distributions of the GMM are given by:

\[
P(x|z=i) = \frac{1}{\sqrt{(2\pi)^d|\Sigma_i|}} \exp \left\{ -\frac{1}{2} (x - \mu_i)^\top \Sigma_i^{-1} (x - \mu_i) \right\}.
\]

Show that if \( \Sigma_i = \sigma^2 I \) for all \( i \), where \( \sigma^2 \) is a scalar and \( I \) is the identity matrix, then:

\[
\lim_{\sigma^2 \to 0} P(z=i|x) = \begin{cases} 1 & \text{if } i = \arg \min_j \|x - \mu_j\| \\ 0 & \text{otherwise} \end{cases}
\]
1.5 Matrix lemmas

Let $\Psi \in \mathbb{R}^{D \times D}$ denote a diagonal square matrix and $\Lambda \in \mathbb{R}^{D \times d}$ a tall rectangular matrix with $d \leq D$. Prove the matrix inverse and matrix determinant lemmas stated in class:

\[
\begin{align*}
(\Psi + \Lambda \Lambda^\top)^{-1} &= \Psi^{-1} - \Psi^{-1} \Lambda (I + \Lambda^\top \Psi^{-1} \Lambda)^{-1} \Lambda^\top \Psi^{-1} \\
\det(\Psi + \Lambda \Lambda^\top) &= \det(\Psi) \det(I + \Lambda^\top \Psi^{-1} \Lambda)
\end{align*}
\]

Your proofs may appeal to standard results from linear algebra (e.g., that the determinant of a matrix is equal to the product of its eigenvalues).

1.6 Factor analysis

In factor analysis of zero mean data, the latent and observed variables are assumed to have the multivariate Gaussian distributions:

\[
P(z) = \frac{1}{(2\pi)^{d/2}} \exp \left\{ -\frac{1}{2} z^\top z \right\},
\]

\[
P(x|z) = \frac{1}{\sqrt{\det(\Psi)}} \exp \left\{ -\frac{1}{2} (x - \Lambda z)^\top \Psi^{-1} (x - \Lambda z) \right\}.
\]

Starting from the above, derive the form of the marginal distribution $P(x)$. In particular, show that $P(x)$ is a multivariate Gaussian distribution with $E[x] = 0$ and $E[xx^\top] = \Psi + \Lambda \Lambda^\top$.

1.7 Maximum entropy projections

Consider a multidimensional random variable $\vec{x} \in \mathbb{R}^d$ and a linear projection $y = \vec{x} \cdot \hat{u}$ of this random variable onto the real line, where $\hat{u}$ is a unit vector with $\hat{u} \cdot \hat{u} = 1$. The scalar $y \in \mathbb{R}$ is a random variable in its own right whose uncertainty is measured by its entropy $H[y] = -E[\log P(y)]$.

(a) Suppose that $\vec{x} \in \mathbb{R}^d$ is Gaussian distributed with zero mean and covariance matrix $\Sigma$. Compute the mean $\mu = E[y]$ and variance $\sigma^2 = E[(y - \mu)^2]$ of the linear projection $y = \vec{x} \cdot \hat{u}$ as a function of the direction $\hat{u}$ and the covariance matrix $\Sigma$.

(b) If $\vec{x} \in \mathbb{R}^d$ is Gaussian distributed, then $y = \vec{x} \cdot \hat{u}$ will also be Gaussian distributed since it is a linear function of $\vec{x}$. Write down the Gaussian distribution for $y$.

(c) Compute the entropy $H[y] = -\int dy P(y) \log P(y)$ of a Gaussian random variable $y$ with variance $\sigma^2$.

(d) Use your answer for the variance $\sigma^2$ in part (a) to express the entropy $H[y]$ of the linear projection $y = \vec{x} \cdot \hat{u}$ as a function of the direction $\hat{u}$ and the covariance matrix $\Sigma$.

(e) Derive an eigenvalue equation for the direction $\hat{u}$ that maximizes this entropy. How is this direction related to the first principal component of a data set with zero mean and covariance matrix $\Sigma$?