Homework 4 Part 1 - Solutions

Problem 1 (10 points)

Consider the k-Nearest Neighbors (kNN) classifier for m classes.

Define the class prior probability $P(C_j)$, data likelihood $p(\mathbf{x}|C_j)$, and show that the KNN's posterior probability is

$$P(C_j|\mathbf{x}) = \frac{k_j}{k}$$

where k_j is the number of neighbors from class C_j. In addition, explain how you can generate new samples.

Suppose that we place a sphere of volume $V(\mathbf{x})$ around $\mathbf{x}$ and capture k samples, k_j of which turn out to be labeled as C_j. The estimate for the joint probability $p(\mathbf{x}, C_j)$ is

$$P(\mathbf{x}, C_j) = \frac{k_j/N}{V(\mathbf{x})}$$

where N is the number of training samples.

The posterior probability can then be estimated,

$$\begin{aligned} P(C_j|\mathbf{x}) &= \frac{P(\mathbf{x}, C_j)}{P(\mathbf{x})} \\ &= \frac{P(\mathbf{x}, C_j)}{\sum_{i=1}^{M} P(\mathbf{x}, C_i)}, \quad M \equiv \text{ no. of classes} \\ &= \frac{\frac{k_j/N}{V(\mathbf{x})}}{\sum_{i=1}^{M} \frac{k_i/N}{V(\mathbf{x})}} \\ &= \frac{k_j/N}{\sum_{i=1}^{M} k_i/N} \\ &= \frac{k_j}{k} \end{aligned}$$

Problem 2 (15 points)

Consider the objective function for the Possibilistic C-Means (PCM)

$$J(\Theta, \mathbf{U}, \mathbf{H}) = \sum_{i=1}^{N} \sum_{k=1}^{K} u_{ik}^m d^2(\mathbf{x}_i, \theta_k) + \sum_{k=1}^{K} \eta_k \sum_{i=1}^{N} (1 - u_{ik})^m$$

where $m > 1$ is the *fuzzifier*, $u_{ik} \in [0, 1]$ is the membership value of sample $\mathbf{x}_i$ in cluster centroid θ_k, $d^2(\mathbf{x}_i, \theta_k)$ is the squared distance between sample $\mathbf{x}_i$ and cluster centroid θ_k, and $\eta_k > 0$ determines the relative significance of the two terms.

1. (5 points) **Explain the role of the second term in the objective function for PCM.**

Direct minimization of the first term alone will result in the trivial solution where $u_{ik} = 0$ for all samples. In order to avoid this situation, we introduce the second term in the objective function. As you can see, this term is a function of u_{ik}'s only. This term will be minimized for a membership value u_{ik} as close to 1 as possible. Therefore forcing the clustering algorithm to find an arrangement of the data cloud such that each data point is doubtlessly to belong to a certain cluster. This will forcibly "leave out" outliers.

1. (10 points) **Observe that, for each vector $\mathbf{x}_i$, the u_{ik}'s, $k = 1, \ldots, K$, are independent of each other, we can write $J(\Theta, \mathbf{U}, \mathbf{H})$ as**

$$J(\Theta, \mathbf{U}, \mathbf{H}) = \sum_{k=1}^{K} J_k$$

where

$$J_k = \sum_{i=1}^{N} u_{ik}^m d^2(\mathbf{x}_i, \theta_k) + \eta_k \sum_{i=1}^{N} (1 - u_{ik})^m$$

Each J_k corresponds to a different cluster and the minimization of $J(\Theta, \mathbf{U}, \mathbf{H})$ with respect to the u_{ik}'s can be carried out separately for each J_k.

Solve for J_k as a function of u_{ik}, η_k and m. For a fixed η_k, describe the placement of the centroids θ_k in the minimization of J. Discuss the implications of this behavior when the number of selected clusters K is *larger* then the number K_n of natural clusters in $\mathbf{X}$ (i.e. $K > K_n$).

We saw in class that the solution for the membership values u_{ik} that minimize J is given as:

$$u_{ik} = \frac{1}{1 + \left(\frac{d^2(x_i, \theta_k)}{\eta_k}\right)^{\frac{1}{m-1}}}$$

Rearranging this equation, we find:

$$d^2(x_i, \theta_k) = \eta_k \left(\frac{1 - u_{ik}}{u_{ik}} \right)^{m-1}$$

Substituting in the J_k equation,

$$\begin{aligned} J_k &= \sum_{i=1}^{N} u_{ik}^m d^2(\mathbf{x}_i, \theta_k) + \eta_k \sum_{i=1}^{N} (1 - u_{ik})^m \\ &= \sum_{i=1}^{N} u_{ik}^m \eta_k \left(\frac{1 - u_{ik}}{u_{ik}} \right)^{m-1} + \eta_k \sum_{i=1}^{N} (1 - u_{ik})^m \\ &= \eta_k \sum_{i=1}^{N} u_{ik} (1 - u_{ik})^{m-1} + \eta_k \sum_{i=1}^{N} (1 - u_{ik})^m \\ &= \eta_k \sum_{i=1}^{N} u_{ik} (1 - u_{ik})^{m-1} + \eta_k \sum_{i=1}^{N} (1 - u_{ik})^{m-1} (1 - u_{ik}) \\ &= \eta_k \sum_{i=1}^{N} u_{ik} (1 - u_{ik})^{m-1} + \eta_k \sum_{i=1}^{N} (1 - u_{ik})^{m-1} - \eta_k \sum_{i=1}^{N} u_{ik} (1 - u_{ik})^{m-1} \\ &= \eta_k \sum_{i=1}^{N} (1 - u_{ik})^{m-1} \end{aligned}$$

For a fixed η_k, minimization of J_k requires maximization of uik's, which, in turn, requires minimization of $d^2(x_i, \theta_k)$. The last requirement implies that θ_k should be placed in a region dense in vectors of the data.

PCM has a **mode-seeking property** which implies that the **number of clusters in the dataset need not be known a priori**. Indeed, if we run PCM for K clusters while the dataset contains M natural clusters, with $K > M$, then, after proper initialization, some of the K clusters will coincide with others. It is hoped that the number of the non-coincident clusters will be equal to M. If, on the other hand, $K < M$, proper initialization will potentially lead to K different clusters. Of course, these are not all the natural clusters formed in the dataset, but at least they are some of them.

Problem 3 (10 points)

Suppose that you are interested in designing a clustering method that will cluster a dataset into spherical clusters with radius r_j centered at centroid θ_j, $j = 1, \ldots, K$.

Conditions:

- **The solution must find $r_j, j = 1, \ldots, K$ so that all K clusters have minimum volume.**

- **Allow a minimal number of points to lie outside the volume of its respective sphere/cluster.**
- **Introduce a penalty $\xi_i > 0, i = 1, \ldots, N$ for points lying outside sphere.**

Design an objective function for this clustering algorithm satisfying the conditions above. Define all hyperaparameters (if any), parameters, and solve for parameter solutions. Explain your reasoning.

(Several solutions will be considered, provided all terms are reasoned with and carried out correctly.)

Consider the scenario where points are clustered into a cluster of radius r (fixed radius). The goal is to find the centroid θ_j that produces a minimal-enclosure (hyper)sphere with the smallest radius r. In order to alleviate the negative effect of outliers, let's allow some points to lie outside the (hyper)sphere. For such points $\mathbf{x}_i$, we will penalize them with a penalty ξ_i. All other points will have $\xi_j = 0$.

In this way, we can write the objective function,

$$\begin{aligned}
\arg_{r,\theta} \min \;\; & r^2 + C\sum_{i=1}^{N} \xi_i \\
\text{where} \;\; & u_{ij} \in \{0,1\}, \; \forall i = 1,2,\ldots,N \; \text{ and } \; j = 1,2,\ldots,K \\
\text{subject to} \;\; & \|\mathbf{x}_i - \theta\|^2 \leq r^2 + \xi_i, \; i = 1,2,\ldots,N \\
& \xi_i \geq 0, \; i = 1,2,\ldots,N
\end{aligned}$$

The hyperparameter C controls the importance it is given on how many samples are allowed to lie outside the (hyper)sphere. As $C \to \infty$ we will recover a solution with no points lying outside the (hyper)sphere. As $C \to 0$, more points will be allowed to be outside the (hyper)sphere.

The Lagrangian of the above constrained problem is given by

$$\mathcal{L}(r,\theta,\mu,\lambda) = r^2 + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\mu_i\xi_i - \sum_{i=1}^{N}\lambda_i\left(r^2 + \xi_i - \|\mathbf{x}_i - \theta\|^2\right)$$

Taking the derivatives of the Lagrangian and equating to zero, we find:

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial r} &= 2r - \sum_{i=1}^{N}(-2r)\lambda_i = 0 \iff \sum_{i=1}^{N}\lambda_i = 1 \\
\frac{\partial \mathcal{L}}{\partial \theta} &= \sum_{i=1}^{N}\lambda_i(x_i - \theta) = 0 \iff \sum_{i=1}^{N}\lambda_i x_i = \theta\sum_{i=1}^{N}\lambda_i \iff \theta = \sum_{i=1}^{N}\lambda_i x_i \\
\frac{\partial \mathcal{L}}{\partial \mu_i} &= C - \mu_i - \lambda_i = 0 \iff \lambda_i = C - \mu_i \\
\frac{\partial \mathcal{L}}{\partial \lambda_i} &= r^2 + \xi_i - \|\mathbf{x}_i - \theta\|^2 = 0 \iff \|\mathbf{x}_i - \theta\|^2 = r^2 + \xi_i
\end{aligned}
$$

Substituting in the Lagrangian, the dual Lagrangian form results in

$$
\begin{aligned}
\arg_\lambda \max & \left(\sum_{i=1}^{N}\lambda_i x_i^T x_i - \sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_j x_i^T x_j\right) \\
\text{subject to } & 0 \le \lambda_i \le C,\ i = 1, 2, \ldots, N \\
& \sum_{i=1}^{N}\lambda_i = 1
\end{aligned}
$$

and the KKT conditions are

$$
\begin{aligned}
\mu_i\xi_i &= 0 \\
\lambda_i[\|x_i - \theta\|^2 - r^2 - \xi_i] &= 0 \\
\theta &= \sum_{i=1}^{N}\lambda_i x_i \\
\lambda_i &= C - \mu_i,\ i = 1, 2, \ldots, N
\end{aligned}
$$

From these conditions the following remarks are easily deduced.

- Only points with $\lambda_i \neq 0$ contribute to the definition of the center of the optimal sphere. These points are known as support vectors.
- Points with $\xi_i > 0$ correspond to $\mu_i = 0$, which leads to $\lambda_i = C$ and, these points lie outside the sphere. Let's refer to these points as bounded support vectors.
- Points with $0 < \lambda_i < C$ have corresponding $\mu_i > 0$ leading to $\xi_i = 0$ and, these points lie on the sphere.
- Points with $\lambda_i = 0$ correspond to $\xi_i = 0$. All points lying inside the sphere satisfy, necessarily, these two conditions.

Instead of working directly in the feature space governed by $x \in \mathbf{X}$, we can consider a higher-dimensional mapping $x \in \mathbf{X} \longrightarrow \phi(x) \in H$ and use the kernel trick properties that appear in SVMs.

Problem 4 (10 points)

Let $\mathbf{e}_i, i = 1, \ldots, D$, be any orthonormal basis in the D-dimensional space. Consider a D-dimensional random vector $\mathbf{x}$, which is approximated by

$$\hat{\mathbf{x}} = \sum_{i=1}^{m} y_i \mathbf{e}_i + \sum_{i=m}^{D} c_i \mathbf{e}_i$$

where c_i are non-random constants. Show that the minimum mean square error $\mathbb{E}\left[\|\mathbf{x} - \hat{\mathbf{x}}\|^2\right]$ is achieved if

1. $c_i = \mathbb{E}[y_i], i = m, \ldots, D$**,**
2. **the orthonormal basis consists of the eigenvectors of Σ_x (covariance of $\mathbf{X}$); and**
3. $\mathbf{e}_i, i = m, \ldots, D$**, correspond to the eigenvectors associated with the $D - m$ smallest eigenvalues.**

In this problem, we defined $y_i = e_i^T \mathbf{x}_i$ as the linear projection onto i-th eigenvector.

If the data is projected onto all eigenvectors, then its reconstruction is lossless. In this case, $\mathbf{x} = \sum_{i=1}^{D} y_i \mathbf{e}_i$.

Now, consider the expected square error $\mathbb{E}\left[\|\mathbf{x} - \hat{\mathbf{x}}\|^2\right]$. We find:

$$\begin{aligned}
\mathbb{E}\left[\|\mathbf{x} - \hat{\mathbf{x}}\|^2\right] &= E\left[\left\|\sum_{i=m}^{D} (y_i - c_i) e_i\right\|\right] \\
&= E\left[\sum_{i=m}^{D} \sum_{j=m}^{D} (y_i - c_i)(y_j - c_j) e_i^T e_j\right] \\
&= E\left[\sum_{i=m}^{D} (y_i - c_i)^2\right] \\
&= \sum_{i=m}^{D} \left(E[y_i^2] - 2E[y_i]c_i + c_i^2\right)
\end{aligned}$$

If we want to pick c_i's that make this as small as possible, we can take the derivative with respect to c_i set the result equal to zero and solve for c_i we find

$$\frac{\partial \mathbb{E}\left[\|\mathbf{x} - \hat{\mathbf{x}}\|^2\right]}{\partial c_i} = 0 \iff -2E[y_i] + 2c_i = 0$$

This gives $c_i = E[y_i]$, for $i = m, \ldots, D$.

We now want to ask for an approximation to $\mathbf{x}$ given by

$$\hat{\mathbf{x}} = \sum_{i=1}^{m} y_i \mathbf{e}_i + \sum_{i=m}^{D} E[y_i]\mathbf{e}_i,$$

how do we pick the orthonormal basis vectors e_i. We do that by minimizing the square norm of the error ϵ defined as $\epsilon = x - \hat{x}$.

$$\begin{aligned}
E[||\epsilon||^2] &= E\left[\sum_{i=m}^{D}(y_i - E[y_i])^2\right] \\
&= E\left[\sum_{i=m}^{D}(e_i^T x - e_i^T E[x])^2\right] \\
&= E\left[\sum_{i=m}^{D}(e_i^T (x - E[x]))^2\right] \\
&= E\left[\sum_{i=m}^{D} e_i^T (x - E[x])(x - E[x])^T e_i\right] \\
&= \sum_{i=m}^{D} e_i^T E[(x - E[x])(x - E[x])^T] e_i \\
&= \sum_{i=m}^{D} e_i^T \Sigma_x e_i
\end{aligned}$$

Thus to pick the orthonormal basis that minimizes $E[||\epsilon||^2]$ we minimize $\sum_{i=m}^{D} e_i^T \Sigma_x e_i$ subject to the constraint that $e_i^T e_i = 1$. Introducing Lagrange multipliers, we will find that the e_i's are the eigenvectors of Σ_x.

Finally, to make the expression for $E[||\epsilon||^2]$ as small as possible we we order these eigenvectors so that they are ranked in decreasing order of their eigenvalues, therefore the vectors $e_m, e_{m+1}, \ldots, e_D$ will be the eigenvectors of Σ_x corresponding to the $D - m$ smallest eigenvalues. ■

On-Time (5 points)

Submit your assignment before the deadline.
