ML HW 3

By Iyar Hadad

1. a) Yes. K-medoid algorithm chooses the center point of each group from the k groups, as the value in the group which its average dissimilarity to all the other points in the group is minimal(=medoid). K-means is based on the mean value of each group. Mathematically, medoid is less sensitive to noises and outliers than mean, because the medoid, by its definition is similar to most of the values in its group, in contrast to mean, which could be very different from other points in case of outliers. Therefore, the K groups dividing will be more robust to exceptional values than K-means.

   b) $\mu = argmin_{\mu\in\mathbb{R}^1}\{\sum_{i=1}^{m}(x_i-\mu)^2\}$

   $$\frac{d}{d\mu}\sum_{i=1}^{m}(x_i-\mu)^2 = \frac{d}{d\mu}\sum_{i=1}^{m}(x_i^2-2x_i\mu+\mu^2) = \frac{d}{d\mu}[\sum_{i=1}^{m}x_i^2-2\mu\sum_{i=1}^{m}x_i+m\mu^2] = -2\sum_{i=1}^{m}x_i+2m\mu = 0$$

   $$\mu = \frac{\sum_{i=1}^{m}x_i}{m} = \bar{X}$$

   Let's check this is the argmin value:

   $$\frac{d}{d\mu}[-2\sum_{i=1}^{m}x_i+2m\mu] = 2m > 0 \quad \rightarrow \quad minimum\ point$$

   Bonus:

   $$\mu = argmin_{\mu\in\mathbb{R}^1}\{\sum_{i=1}^{m}|x_i-\mu|\}$$

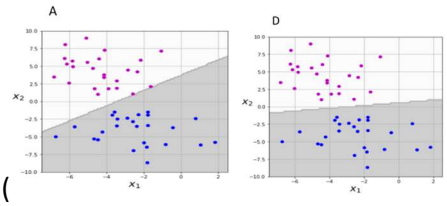
   Let's call n, the index of the first X which is $\geq \mu$, and $x_1 \leq x_2 \leq \cdots x_m$

   $$\frac{d}{d\mu}[\sum_{i=1}^{n-1}(\mu-x_i)+\sum_{i=n}^{m}(x_i-\mu)] = \frac{d}{d\mu}[\sum_{i=1}^{n-1}(\mu-x_i)+\sum_{j=1}^{m-n+1}(x_i-\mu)]$$

   $$= \frac{d}{d\mu}[(n-1)\mu-\sum_{i=1}^{n-1}x_i+\sum_{j=1}^{m-n+1}x_i-(m-n+1)\mu] = n-1-m+n-1 = 2n-m-2$$

   To minimize the expression, we demand $2n-m-2=0 \rightarrow m=2(n-1)$

   Which means, the index n of the first X which is $\geq \mu$ should be the median (according the m=f(n) we got).

2. In figures A, D we can visually see the classifier is linear (

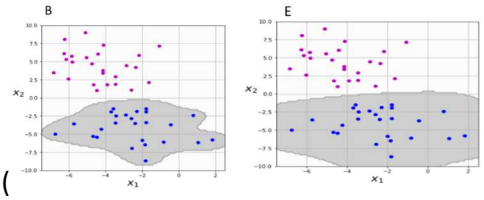
) because the planes are linear, which means their kernels are linear. In D, we can see bigger margins, from the plane, than A.

C represents the regularization constant of the SVM problem:

$$argmin_{w,b}\{C\sum_{i=1}^{m}\xi_i + \frac{1}{2}||w||^2\}$$

The bigger C is, the more we demand to minimize ξ which presents the softness of the margins. If we let soft margins likr in figure A, then we do not give a high weight to the expression $C\sum_{i=1}^{m}\xi_i$ in the minimalization problem, so we set a small C. Therefore:

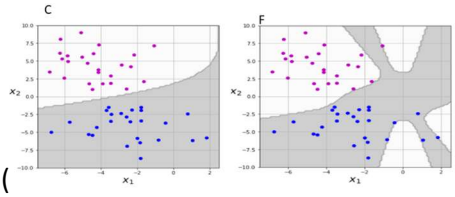
$C_D > C_A \rightarrow A: linear\ kernel\ C = 0.01, D: linear\ kernel\ C = 1$

In figures B, E we can visually see a gaussian separation ()

The gaussian kernel is defined by:

$$k(x,z) = \exp\left(-||x-z||^2\gamma\right)$$

The higher gamma is, the more we fit the training data. We can see B is much more fitted to the data, therefore: $\gamma_B > \gamma_E \rightarrow B: RBF\ kernel\ \gamma = 1, E: RBF\ kernel\ \gamma = 0.2.$

In figures C, F we can visually see a polynomial separation () when the separation plane in F has bigger order. Therefore:

$$order_F > order_C \rightarrow C: 2^{nd}\ order\ polynomial\ kernel, F: 10^{th}\ order\ polynomial\ kernel$$

3. a. Generalization. Which is the tradeoff by, the model could be effective and simple at the same time. We desire the model will be capable to make accurate predictions on new test data and avoid overfitting to the training dataset.

b. $2p$ represents overfitting of the model to the data, because the bigger p is, the more parameters the model learns, and by that it becomes more complex. We don't want overfitting, we want the model to be "As simple as possible".

$2\ln(\hat{L})$ represents the estimated probability to these parameters, if its value is high it means there is a good likelihood the model performs good classification. High probability that a point belongs to specific class 'Y' can tells that classifying that point as 'Y' is a good assumption. We want the model could make that decision, and for that it must contain a certain level of complexity, "not simpler" as Einstein said.

C. In science, we understand that if we build a model which perfectly fits to the training data it will not deal well with new points it did not see before. On the other way, if the model will be too much simple it could not tell anything we could say in confidence that it true, and we will stay with the meaningless assumption there is 50% it is true and 50% it's false.

d. The AIC should be low. It represents an error. We want low order of model (simple model, no underfitting) and therefore we want low $2p$, in the other hand we want high probability estimation, so we want high value of $2\ln(\hat{L})$. As much $2p$ is low, and $2\ln(\hat{L})$ is high, the lower AIC.