Machine learning in healthcare - HW3

Naama Rivlin

311576599

Question 1: Clustering

a. Is K-medoid more robust to noise (or outliers) than the K-means algorithm? Explain your answer.

Answer: Yes, the K-medoid algorithm is more robust to noise than the K-means algorithm. The mean is easily influenced by extreme values and therefore the K-means algorithm, which uses the mean point as the center of the clusters, is sensitive to outliers. K-medoids clustering is a variant of K-means, but instead of using the mean as center it uses an actual point in the cluster- the medoid. The medoid is the most centrally located data point in the cluster, that has the minimal sum of distances to the other points in the cluster. In other words, K-medoid algorithm minimizes a sum of general pairwise dissimilarities instead of a sum of squared Euclidean distances. As a result, even if an outlier is very far from the center point, it will not change the center point.

b. Prove that for the 1D case (x ∈ R1) of K-means, the centroid (μ) which minimizes the term $\sum_{i=1}^{m}(xi\ -\ \mu)^2$ is the mean of m examples.

Answer: By applying the differential operator $\frac{d}{d\mu}$ on $\sum_{i=1}^{m}(xi\ -\ \mu)^2$ and comparing it to 0:

$$\frac{d}{d\mu}\left(\sum_{i=1}^{m}(x_i-\mu)^2\right)=-2\sum_{i=1}^{m}(x_i-\mu)$$

$$-2\sum_{i=1}^{m}(x_i-\mu)=0$$

$$2\sum_{i=1}^{m}x_i=2m\mu$$

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

We get that the centroid μ which minimizes the term $\sum_{i=1}^{m}(xi\ -\ \mu)^2$ is the mean of m examples.

*The μ we found is necessarily the minima since the second derivative is a positive constant (2m).

Bonus: Prove that the centroid (practically, the medoid) which minimizes the term $\sum_{i=1}^{m}|(\mathrm{x_i}\ -\ \mu)|$ is the median of m examples given that μ belongs to the dataset.

Answer: We can divide x_i into 3 groups:
$\alpha: x_i > \mu,\ \beta: x_i < \mu,\ \gamma: x_i = \mu$

Now we can simplify the term $\sum_{i=1}^{m}|(\mathrm{x_i}\ -\ \mu)|$:

$$\sum_{i=1}^{m}|(\mathrm{x_i}\ -\ \mu)|=\sum_{i=1}^{\alpha}(x_i-\mu)+\sum_{i=1}^{\beta}(\mu-x_i)+(\mu-\mu)=\sum_{i=1}^{\alpha}x_i-\mu\alpha+\mu\beta-\sum_{i=1}^{\beta}x_i$$

μ that minimizes the term $\sum_{i=1}^{m}|(\mathrm{x_i}\ -\ \mu)|$ is the median if $\alpha=\beta$ when the term is minimized, because it means the minima is reached when the number of samples that are smaller than μ is equal to the number of samples that are bigger than μ.

applying the differential operator $\frac{d}{d\mu}$ on the simplified term and comparing it to 0:

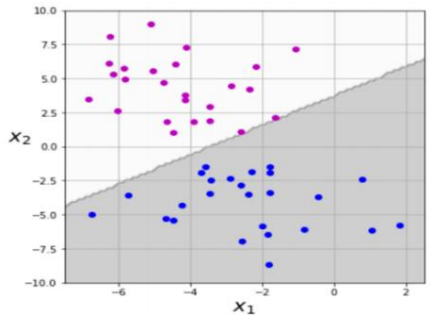
$$\frac{d}{d\mu}\left(\sum_{i=1}^{\alpha} x_i - \mu\alpha + \mu\beta - \sum_{i=1}^{\beta} x_i\right) = -\alpha + \beta = 0$$

$$\rightarrow a = b$$

→ $\alpha = \beta$ when the term is minimized → μ is indeed the median.

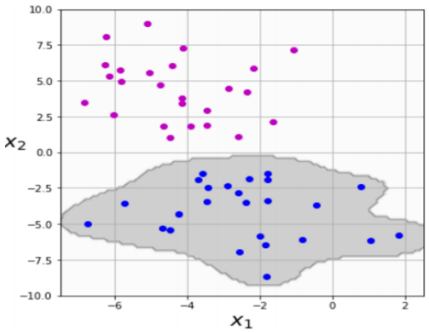
Question 2: SVM

A- 1: Linear kernel with $C = 0.01$



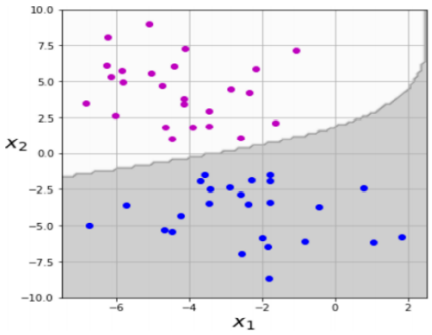
In linear SVM C controls the tradeoff between larger margins and correct classifications. Looking at plot A, it is difficult to recognize the margin, but one can observe 2 purple points that are almost on the decision boundary, while in the blue class there are no points this close to the decision boundary. This indicates that some misclassification is allowed in this model, which means C is relatively small (small C → small penalty → more misclassifications).

B- 6: RBF kernel with $\gamma = 1$



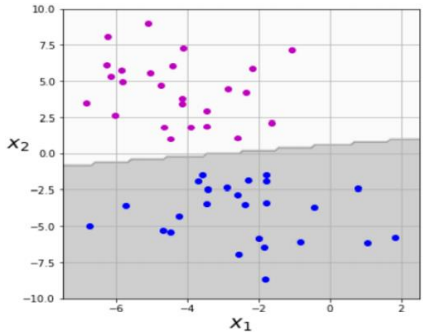
The closed shape of the decision region implies that an RBF kernel was used. In RBF SVM, gamma hyper-parameter decides how much curvature we have in a decision boundary, therefore controls under/overfitting. Higher γ → the decision boundary gets curvier and the model gets more overfitted. This plot shows the behavior of a model with a higher gamma than the other RBF model in this question.

C- 3: 2nd order polynomial kernel



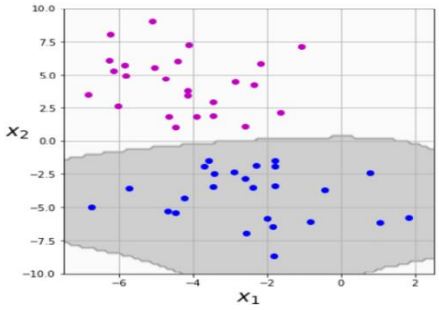
The decision boundary in this plot shows the form of a low degree polynomial.

D- 2: Linear kernel with $C = 1$



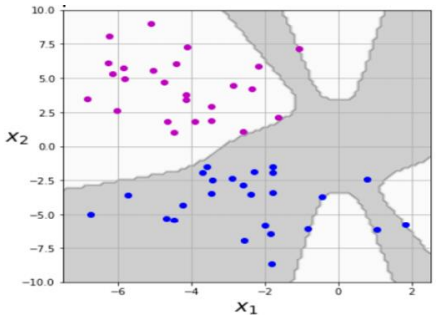
The bigger C is, the bigger the penalty for misclassification is. This plot shows a model that does not allow misclassifications, meaning it has a bigger C than the other linear SVM model in this question.

E- 5: RBF kernel with $\gamma = 0.2$



The plot has the characteristic form of a decision region that was determined with RBF kernel, and not overfitted as the plot B.

F- 4: 10th order polynomial kernel



Decision boundaries determined with polynomial kernels does not necessarily form a closed shape. The higher the polynomial's degree is, the more complex the boundaries are. This figure has complex decision boundaries that does not show a closed shape, which implies a high degree of polynomial kernel.

Question 3: Capability of generalization

a. What is the scientific term of the balance that Einstein meant to in machine learning aspect?

Answer: In machine learning, the scientific term is **generalization**, which lies on the balance between model simplicity and model complexity. On the one hand, a more complex model may be more explanatory, and on the other hand, the less complex a model is, the more likely it is that a good empirical result is valid. Generalization means finding that balance, and this is done by making sure we use valid data and by evaluating the complexity of the model and model's performance on training and testing data.

b. How does each of the terms $(2p, 2ln(\hat{L}))$ in AIC affect the terms of the balance you defined in (a)?

Answer: P is the number of parameters in the model and $\hat{L}$ is a measure of goodness of fit. **2p** increases as we increase the number of parameters (which increases the complexity of the model). This results in a bigger AIC. One can look at this as a kind of penalty for building a complicated model. $\mathbf{2ln}(\hat{\boldsymbol{L}})$ increases as our model gets better performance. This results in a smaller AIC because this term has a minus sign. One can look at this as a reward for building a model with good fit. The 2 terms together balance between model complexity and simplicity, since it is not "allowed" to complicate the model if it does not result in better performance (or as Ockham's razor states: " without necessity").

c. What are the two options that are likely to happen if this balance was violated?

Answer: Too much complexity together with poor fitting ability will result in an overfitted model that will have poor performance with new data. However, a model that is too simplified will be underfitted and will also have poor performance on any dataset.

d. What are we aiming for with the AIC? Should it be high or low? Explain.

Answer: AIC should be small. Small AIC indicates a good balance between how complex and how explanatory the model is. However, it cannot tell whether the model has good performance, but only whether the balance exists or not.