336546 - Machine Learning in Healthcare:

Sivan Geva 313581589

1. Clustering

a. K-Medoids is more robust as compared to K-Means. As in K-Medoids we find k data points as centers to minimize the sum of dissimilarities of data objects whereas, K-Means used sum of squared Euclidean distances for data objects. Calculating Manhattan distance reduces noise and outliers by choosing a point near other clustered points, and not a point in the center influenced toward the outliners.

b. Say the mean of a vector of m observations is:

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

μ will be the centroid of k-mean clustering.

Prove: the μ that minimizes $\sum_{i=1}^{m}(x_i - \mu)^2$ is $\mu = \bar{x}$

Proof:

$$\begin{aligned}\sum_{i=1}^{m}(x_i-\mu)^2 &= \sum_{i=1}^{m}(x_i-\bar{x}+\bar{x}-\mu)^2 = \sum_{i=1}^{m}(x_i-\bar{x}+\bar{x}-\mu)^2 \\ &= \sum_{i=1}^{m}(x_i-\bar{x})^2 + 2\sum_{i=1}^{m}(x_i-\bar{x})(\bar{x}-\mu) + \sum_{i=1}^{m}(\bar{x}-\mu)^2 \\ &= \sum_{i=1}^{m}(x_i-\bar{x})^2 + 2(\bar{x}-\mu)\underbrace{\sum_{i=1}^{m}(x_i-\bar{x})}_{=0\ (definition\ of\ mean)} + m(\bar{x}-\mu)^2 \\ &= \sum_{i=1}^{m}(x_i-\bar{x})^2 + m(\bar{x}-\mu)^2\end{aligned}$$

$$\frac{\partial(\sum_{i=1}^{m}(x_i-\mu)^2)}{\partial\mu} = \frac{\partial(\sum_{i=1}^{m}(x_i-\bar{x})^2 + m(\bar{x}-\mu)^2)}{\partial\mu} = -2m(\bar{x}-\mu) \underbrace{=}_{want} 0$$

$$\Rightarrow \bar{x} - \mu$$

Check that it's minimum:

$$\left.\frac{\partial^2(\sum_{i=1}^{m}(x_i-\mu)^2)}{(\partial\mu)^2}\right|_{\mu=\bar{x}} = 2m > 0 \Rightarrow minimum$$

Bonus:

Recall: median of $\bar{x} \in R^1$is $x_{\frac{n}{2}+1}$ if n id odd, and any number between $\left[x_{\frac{n}{2}}, x_{\frac{n}{2}+1}\right]$ if n is even.

Prove: the μ that minimizes $\sum_{i=1}^{k}|x_i - \mu|$ is the median.

Since $\frac{1}{k}\sum_{i=1}^{k}|x_i - x_1| = \bar{x} - x_1 < \bar{x} - \mu = \frac{1}{k}\sum_{i=1}^{k}|x_i - \mu|$ for $\mu < x_1$, and $\frac{1}{k}\sum_{i=1}^{k}|x_i - x_k| = x_k - \bar{x} < \mu - \bar{x} = \frac{1}{k}\sum_{i=1}^{k}|x_i - \mu|$ for $\mu > x_k$ than we consider $\mu \in [x_1, x_k]$.

Proof: we will use induction.

For k=2:

$$\sum_{i=1}^{2}|x_i - \mu| = (\mu - x_1) + (x_2 - \mu) = x_2 - x_1$$

The proposition is true

For any k=n: assume the proposition is true.

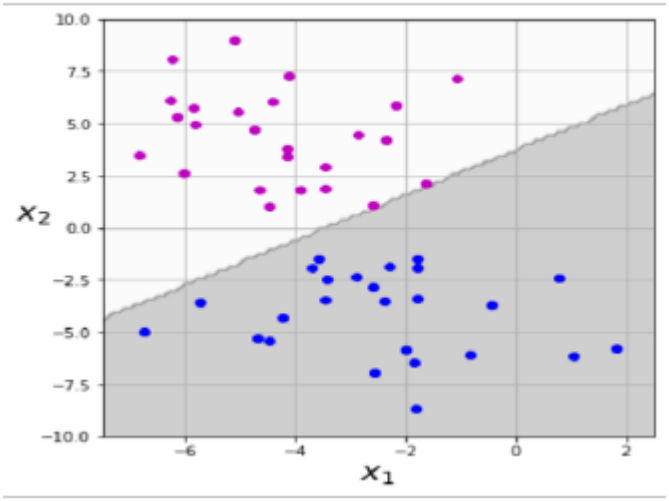
For k=n+1:

$$\sum_{i=1}^{n+1}|x_i - \mu| = (\mu - x_1) + \sum_{i=2}^{k}|x_i - \mu| + (x_{k+1} - \mu) = \sum_{i=2}^{k}|x_i - \mu| + (x_{k+1} - x_1)$$

The number, according to the assumption, that minimize $\sum_{i=2}^{k}|x_i - \mu|$ is the median, and since for $k \geq 2$ the set of medians of $x_1, \dots, x_{k+1}$is the same set of medians of $x_2, \dots, x_k$, the validity of the proposition is true for all n.

2. SVM

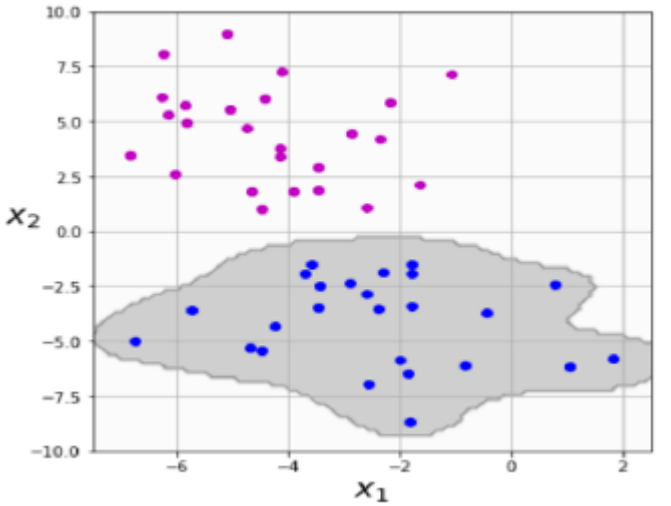
A



Linear kernel with C = 0.01.

The separation is linear. The margines are bigger than those in plot D- thus the C hyperparameter is smaller. The two purple observations are close to the hyperplane margine meaning the model allows misclassification or data to be between margins created by the support vectors.

B



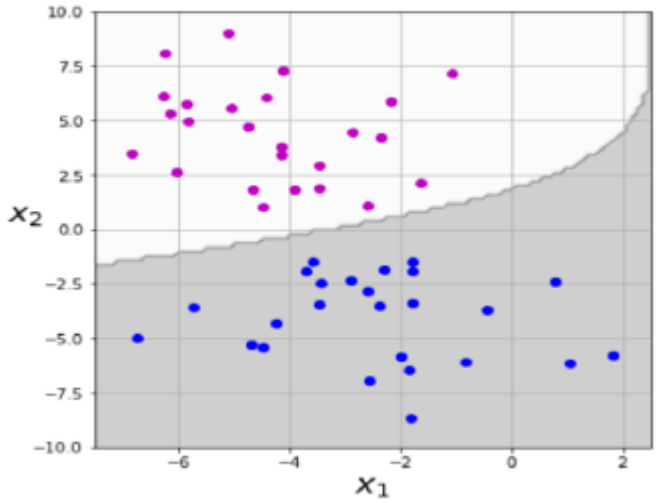
Matches:

RBF kernel with γ = 1

The classifier's shape correlates with RBF kernel.

The model is more over-fitted than figure E thus gamma hyperparameter is bigger.

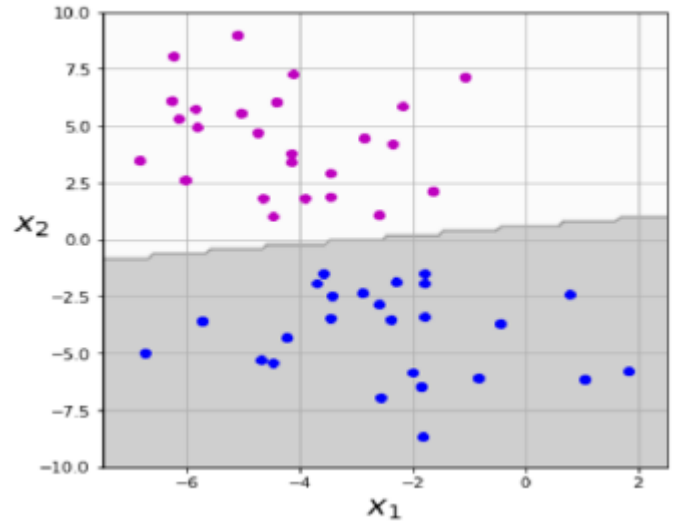
C



matches:

2nd order polynomial kernel. The polynomial kernel separates the data by non-linear margin from small order, 2.

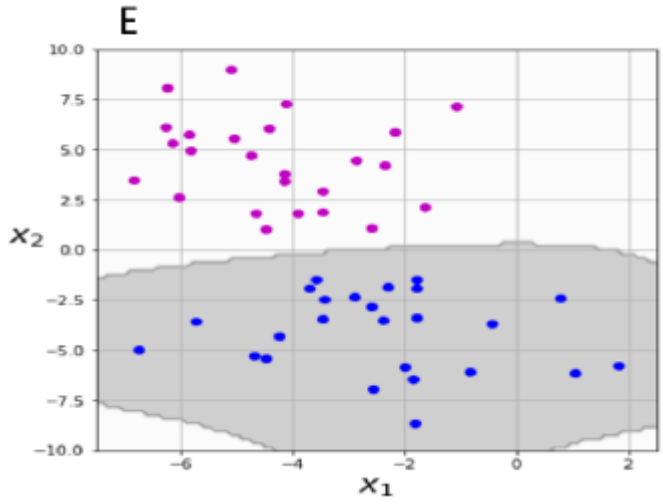
D



matches:

Linear kernel with C = 1.

The separation is linear. The margins are smaller than those in plot D- thus the C is higher.

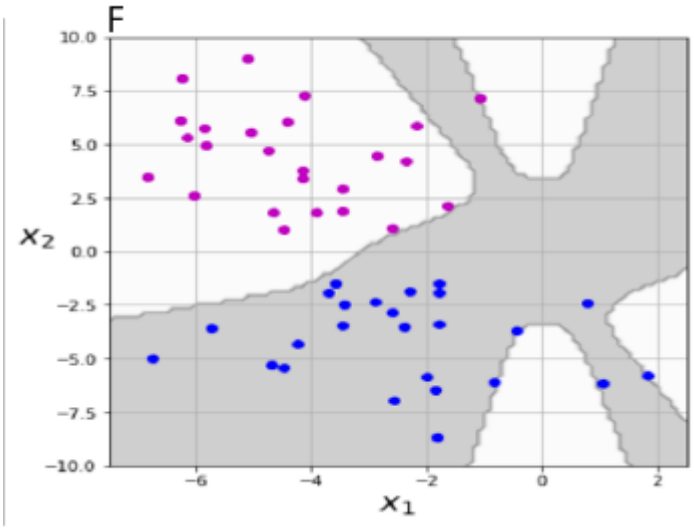


Matches:

RBF kernel with γ = 0.2

The classifier's shape correlates with RBF kernel.

The model is less over-fitted than figure B thus gamma hyperparameter is smaller.



matches:

10th order polynomial kernel

The higher the order of the polynomial kernel, the higher the over- fitting will be. We see that graph F is more over-fiitted than graph C, with smaller degree polynomial kernel.

3. <u>Capability of generalization:</u>

a. In machine learning, the scientific term of the balance Albert Einstein meant in "Everything should be made as simple as possible but not simpler" is generalization. We want our model to be able to fit to as much new data as possible (not over-fitted) but without harming its complexity.

b. The 2p in AIC measures the complexity of the model, while $ln(\hat{L})$ assesse the goodness-to-fit of the model.

c. The balance is violated between over fitted model with high complexity (high number of feature) to under fitted model with low complexity (low number of

features). We want a good fitted model with low complexity to avoid overfitting, and gain generalization.

d. Comparing two different model's AIC value with the same data set, we will choose the model with minimum AIC value. That will tell us that our model has a goodness-to-fit (higher likelihood function) but low complexity (low number of parameters).