Course: Machine Learning in HealthCare 336546 Faculty of Bio-Medical Engineering, Technion
Submitted by: Ori Rajchert
Date: 15/01/2021

HW3 - machine learning in health-care, 336546

Question 1:

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

K-medoid and K-means are algorithms for data partition that help us to minimize the distance between points in our clusters and the center of these clusters.

K-medoid is more robust to noise and outliers as compared to k-means because it minimizes a sum of pairwise dissimilarities instead of a sum of squared Euclidean distances which is used in the K-means algorithm.

As we learnt, the median is more robust to outliers than the arithmetic mean. Essentially, this explanation will also relevant for the medoid. It is a more robust estimation of a representative point than the mean as used in k-means.

Medoids are more robust to outliers than centroids, but they need more computation for high dimensional data.
In addition, we can use k-medoids with any measures. K-means however, may fail to converge - it really must only be used with distances that are consistent with the mean.

References:

- https://stackoverflow.com/questions/21619794/what-makes-the-distance-measure-in-k-medoid-better-than-k-means
- https://stats.stackexchange.com/questions/476286/when-to-use-k-medoids-instead-of-k-means
- https://stats.stackexchange.com/questions/156210/an-example-where-the-output-of-the-k-medoid-algorithm-is-different-than-the-outp

Course: Machine Learning in HealthCare Faculty of Bio-Medical Engineering, Technion
336546
Submitted by: Ori Rajchert
Date: 15/01/2021

b. Prove that the centroid (practically, the medioid) which minimizes the term $\sum_{i=1}^{m}(x_i-\mu)^2$ is the mean of m examples.

$$S(\mu)=\sum_{i=1}^{m}(x_i-\mu)^2=\sum_{i=1}^{m}\left(x_i^2-2\mu\cdot x_i+\mu^2\right)=$$
$$=\sum_{i=1}^{m}x_i^2-\sum_{i=1}^{m}2\mu x_i+\sum_{i=1}^{m}\mu^2=$$
$$=\sum_{i=1}^{m}x_i^2-2\mu\sum_{i=1}^{m}x_i+m\cdot\mu^2$$
$$S'(a)=-2\cdot\sum_{i=1}^{m}x_i+2\mu\cdot m$$
$$S'(a)=0$$
$$\Rightarrow -2\cdot\sum_{i=1}^{m}x_i+2\mu\cdot m=0$$
$$\sum_{i=1}^{m}x_i=\mu\cdot m$$
$$\mu=\frac{\sum_{i=1}^{m}x_i}{m}=mean\ \ definition$$

In order to make sure that it's a minimum point I applied a second derivative function:

Since $S''(\mu)=2\cdot m>0$, the value gives a minimum.

Bonus:

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

First way, inspired by Analysis of Biological Signals course, 336208, tutorial 3, slide 24:

$$as\ we\ know:\frac{\mathrm{d}|x|}{\mathrm{d}x}=sign(x)\ \ or\ \ subgradient$$
$$\frac{\partial\left(\sum_{i=1}^{m}|x_i-\mu|\right)}{\partial\mu}=-\sum_{i=1}^{m}sign(x_i-\mu)$$

This equals to zero only when the number of positive items equals the number of negative which happens when $\mu=median\{x_1,x_2,...x_m\}$.

Course: Machine Learning in HealthCare Faculty of Bio-Medical Engineering, Technion
336546
Submitted by: Ori Rajchert
Date: 15/01/2021

Second way, inspired by statistical course, Homework 1:

Assumption:

n is even

$$median := C \in \left(x_{\frac{m}{2}}, x_{\frac{m}{2}+1} \right) \Rightarrow x_{\frac{m}{2}} < C < x_{\frac{m}{2}+1}$$

We would like to prove that: $for\, a \in \mathbb{R} \Rightarrow this\ exists : \sum_{i=1}^{m} |x_i - c| \le \sum_{i=1}^{m} |x_i - a|$

$$and\ therefore : \arg\min_{\mu} \sum_{i=1}^{m} |x_i - \mu| = C$$

We will try to prove that: $\sum_{i=1}^{m} |x_i - a| - \sum_{i=1}^{m} |x_i - c| \ge 0$

Without limitation of generality we assume that $a < C$ (the case that $a > C$ is symmetrically the same). We define 3 groups:

$$group\, A = \{ i : x_i < a \}$$
$$group\, B = \{ i : a < x_i < C \}$$
$$group\, c = \{ i : x_i > C \}$$

For group A:

$$\forall i \in A \{ x_i < a < C \} \Rightarrow |x_i - a| - |x - C| = -(x_i - a) - (-(x_i - C)) = a - C$$

For group B:

$$\forall i \in B : a \le x_i \le C \Rightarrow |x_i - a| - |x - C| = (x_i - a) - (-(x_i - C)) = 2x_i - a - C \ge 2a - a - C = a - C$$

For group c:

$$\forall i \in c : a \le C \le x_i \Rightarrow |x_i - a| - |x - C| = x_i - a - x_i + C = C - a$$

$$\Rightarrow \sum_{i=1}^{m} (|x_i - a| - |x_i - C|) = \left(\sum_{i \in A} (|x_i - a| - |x_i - C|) + \sum_{i \in B} (|x_i - a| - |x_i - C|) + \sum_{i \in c} (|x_i - a| - |x_i - C|) \right)$$

Because we assumed that C is the median :

$$The\ amount\ of\ samples(numbers)\ in\ group\ A + B = \frac{m}{2} = \#(A + B)$$

$$The\ amount\ of\ samples(numbers)\ in\ group\ c = \frac{m}{2} = \#(c)$$

Course: Machine Learning in HealthCare 336546 Faculty of Bio-Medical Engineering, Technion
Submitted by: Ori Rajchert
Date: 15/01/2021

$$\sum_{i=1}^{m}\left(|x_i-a|-|x_i-C|\right)=\left(\sum_{i\in A}\left(|x_i-a|-|x_i-C|\right)+\sum_{i\in B}\left(|x_i-a|-|x_i-C|\right)+\sum_{i\in c}\left(|x_i-a|-|x_i-C|\right)\right)=$$

$$=\left(\sum_{i\in A}(a-C)+\sum_{i\in A}(a-C)+\sum_{i\in A}(C-a)\right)=\left[(C-a)\cdot\#(c)\right]-\left[(C-a)\cdot\left(\#(A)+\#(B)\right)\right]=$$

$$=(C-a)\cdot\left[\#(c)-\left(\#(A)+\#(B)\right)\right]=(C-a)\cdot\left(\frac{m}{2}-\frac{m}{2}\right)=0$$

To sum up:

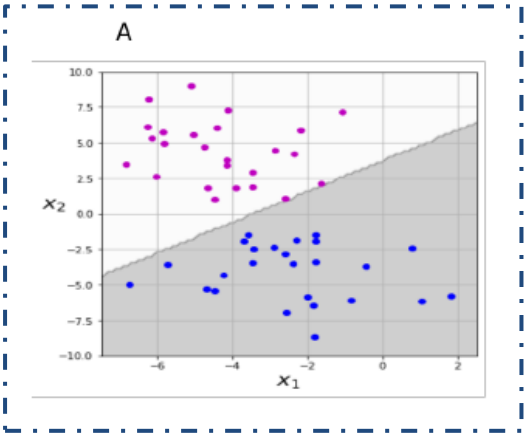
$$\sum_{i=1}^{m}\left(|x_i-a|-|x_i-C|\right)\geq 0\Rightarrow\sum_{i=1}^{m}\left(|x_i-C|\right)\leq\sum_{i=1}^{m}\left(|x_i-a|\right)$$

$$\Rightarrow\boxed{\arg\min_{\mu}\sum_{i=1}^{m}|x_i-\mu|=C=median}$$

Question 2:

The C parameter C is a regularization parameter for SVMs. It controls how much you want to punish your model for each misclassified point for a given curve:

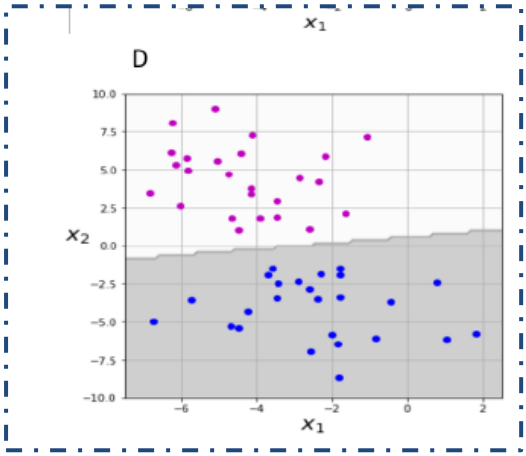
Large values of C	Small Values of C
Large effect of noisy points. A plane with very few misclassifications will be given precedence.	Low effect of noisy points. Planes that separate the points well will be found, even if there are some misclassifications



A=1 – Linear kernel with C=0.01

If the hyper parameter C is low (penalty is low) we have soft margins SVM model – which means that we can tolerate misclassification in the space between our margins and on the margins' boundaries as well.
In the classification in picture A we can see that there are misclassifications. It means that we have soft margin and that the model's penalty is low.

Course: Machine Learning in HealthCare 336546 Faculty of Bio-Medical Engineering, Technion
Submitted by: Ori Rajchert
Date: 15/01/2021



D = 2 – Linear kernel with C=1

The bigger the hyper parameter C is, the penalty is higher. In this case our model would be much less tolerant to misclassifications. This model have hard margins.

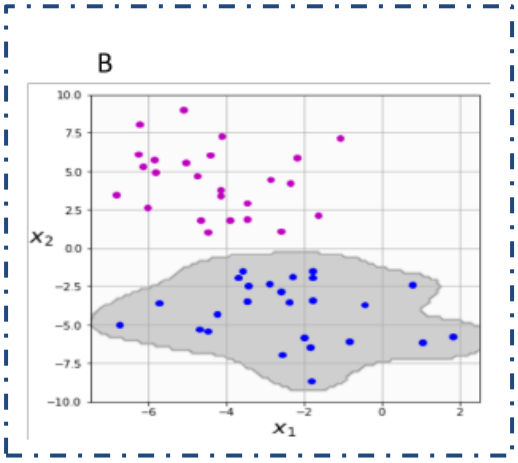
Hyper parameter $\gamma : \gamma = \frac{1}{\sigma^2}$

From the tutorials we have an illustration of the Gaussian RBF Gamma:

Image 1, from tutorial in ML course

- $High\ \gamma \Rightarrow low\ \sigma \Rightarrow low\ \text{var} \Rightarrow$ Less scattering of the classification which means separated Gaussian peaks resulting more complex separation plane.
- $Low\ \gamma \Rightarrow High\ \sigma \Rightarrow High\ \text{var} \Rightarrow$ higher scattering of the classification which means overlapping Gaussians resulting unified separation plane.

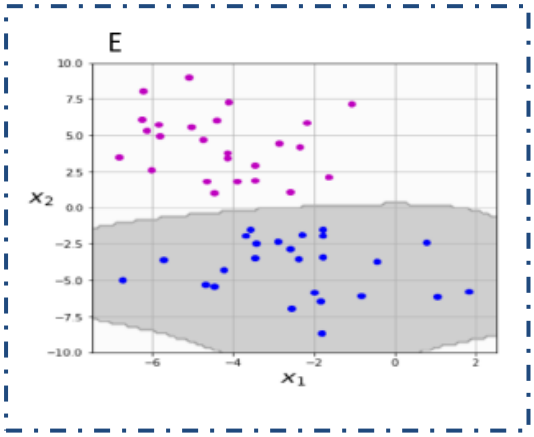
Therefore: B=6, E=5.



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

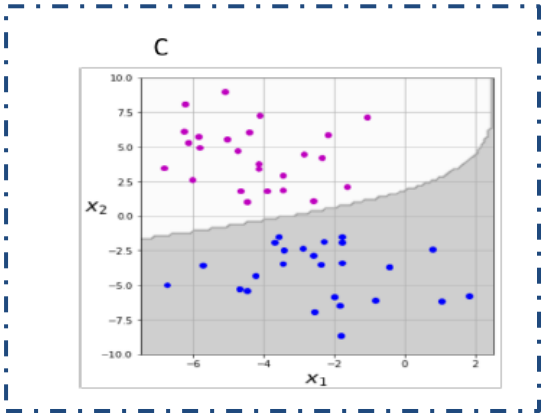
For high values of gamma, the points need to be very close to each other in order to be considered in the same group (or class). Large gamma values are likely to end up in overfitting.

The bigger the value of γ is, the smaller the radius of the classification margin is.

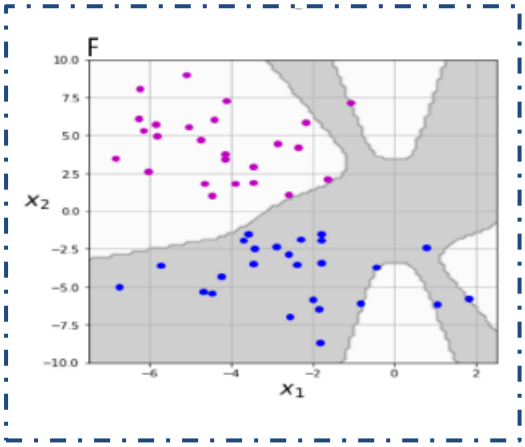


E = 5 - RBF kernel with $\gamma = 0.2$.

Low values of gamma indicate a large similarity radius which results in more points being grouped together.



C = 3 = 2^{nd} order polynomial kernel – the data is classified by a polynomial (parabola/hyperbola) model. The data is not over fitted.



F = 4 = 10^{th} order polynomial kernel. The data classified by a polynomial model. However, it is over fitted and the margins are large.
In general the degree parameter controls the flexibility of the decision boundary. Higher degree kernels yield a more flexible decision boundary.

References:

- https://towardsdatascience.com/hyperparameter-tuning-for-support-vector-machines-c-and-gamma-parameters-6a5097416167
- https://queirozf.com/entries/choosing-c-hyperparameter-for-svm-classifiers-examples-with-scikit-learn

Question 3:

a. What is the scientific term of the balance that Einstein meant to in machine Learning aspect?
"Everything should be made as simple as possible but not simpler".

To be technical, the only way of knowing for sure what Einstein meant by this is to ask him, and he's dead. So we're out of luck on this one. 😊

Nevertheless, the scientific term in machine learning that fits to Einstein's saying is generalization. This is a term that used to describe a model's ability to react to new data.
The meaning of this is that after being trained on a training set, a model can handle new data and make accurate predictions. A model's ability to generalize is central to the success of a model.
If a model has been trained too well on the training data, it will be unable to generalize. It will make inaccurate predictions when given new data, making the model useless even though it is able to make accurate predictions for the training data. This is called overfitting.
Under-fitting happens when a model has not been trained enough on the data. In the case of under-fitting, it makes the model just as useless and it is not capable of making accurate predictions, even with the training data.

The figure demonstrates the three concepts discussed above.

- On the left graph, the blue line represents a model that is under-fitting. The model notes that there is some trend in the data, but it is not specific enough to capture relevant information. It is unable to make accurate predictions for training or new data (according to Einstein's idiom – this model implements the "too much simpler" method).
- In the middle graph, the blue line represents a model that is balanced. This model presents the data trend and accurately models it.
  This middle model will be able to generalize successfully.
- On the right graph, the blue line represents a model that is overfitting. The model presents the data trend and accurately models the training data, but it is too specific. It will fail to make accurate predictions with new data because it learned the training data too well.

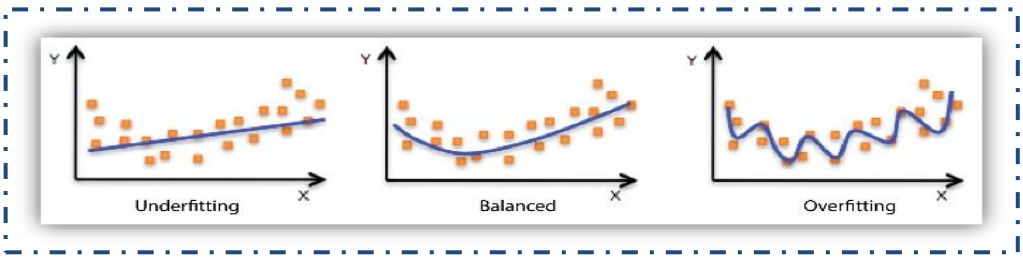


Image 2. from wp.wwu.edu/machinelearning/2017/01/22/generalization-and-

References:

- https://wp.wwu.edu/machinelearning/2017/01/22/generalization-and-overfitting/

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

Our model consists of deterministic function and noise. We try to fit the best predicted model. In case of overfitting our model will be highly affected by the noise. Therefore, the model will correspond very closely or even exactly to a particular set of data and may fail to properly fit additional new data or predict future observations reliably (overfitting).
The model should be accurate, but also not be complicated. In AIC statistical estimator, there is a tradeoff between complexity and accuracy (performance of the training dataset).
We use classification accuracy to measure the performance of our model.
AIC can be viewed as a measure that combines fit and complexity.

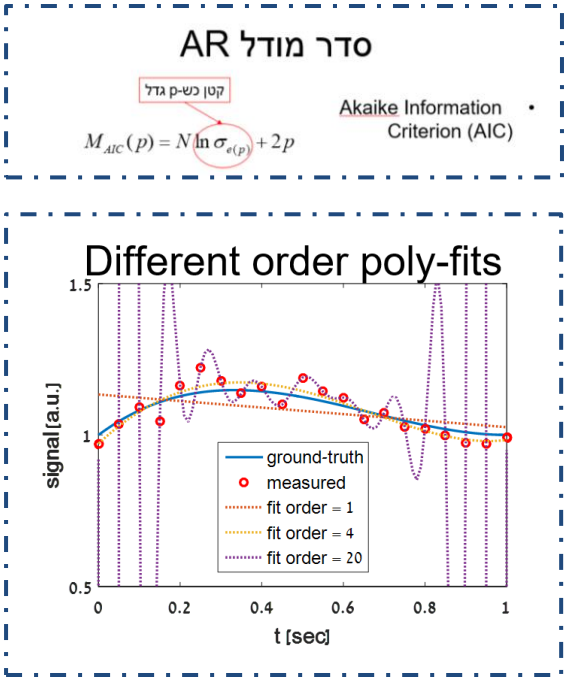


Image 3, 4: from Analysis of Biological Signals course, lecture 3, slides 13, 14

- Under fitting and over fitting illustration and AIC equations are taken from Analysis of Biological Signals course, 336208, lecture 3, slides 13, 14.

<u>Influence of the first term (likelihood):</u>

AIC uses a model's maximum likelihood estimation (log-likelihood) as a measure of fit. Log-likelihood is a measure of how likely one is to see their observed data, given a model. The model with the maximum likelihood is the one that "fits" the data the best. $\hat{L}$ = likelihood = reflects the goodness of fit for MLE.

We use classification accuracy to measure the performance of our model.
The higher the value the more accurate the model is.
In case that the model is very accurate and the p (the second term which represents the complexity) is high as well, the result will be overfitting.

Influence of the second term (P):
P = penalty = stands for the model complexity.
Penalty is an increasing function of the number of estimated parameters. The penalty discourages overfitting, which is desired because increasing the number of parameters in the model almost always improves the goodness of the fit.
The higher the penalty (p), the model will be more complex.
If the model is too simple (the value of p is low), the result will be under fitting.

In order to have a balanced model (which generalizes the samples correctly) we prefer the model to be accurate as possible yet the model should not be too complex.

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

Overfitting and under-fitting (bias, variance tradeoff).

- In case the model is very accurate and the penalty (P) is high as well, the result will be overfitting. In supervised learning, overfitting happens when our model captures the noise along with the underlying pattern in data. It happens when we train our model a lot over noisy dataset. These models have low bias and high variance. These models are very complex.
- If the model is too simple (the value of p is low) and the accurate value is low as well, the result will be under fitting. In supervised learning, under-fitting happens when a model unable to capture the underlying pattern of the data. These models usually have high bias and low variance.

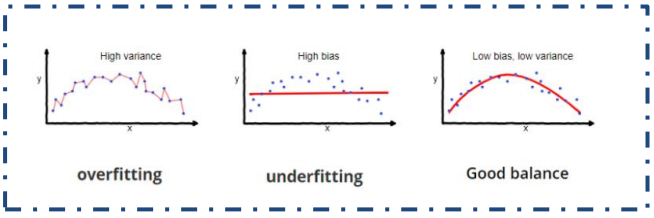


Image 5: from towardsdatascience.com/understanding-the-bias-variance-tradeoff-165e6942b229

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

AIC is a single number score that can be used to determine which of multiple models is most likely to be the best model for a given dataset.
It estimates models relatively, meaning that AIC scores are only useful in comparison with other AIC scores for the same dataset.
Moreover, the AIC score gives you a way to measure the goodness-of-fit of your model, while at the same time penalizing the model for over-fitting the data.

AIC is aimed to provide an estimator of out-of-sample prediction error. Therefore it will indicate the relative quality of statistical model for a given dataset.
A lower AIC score indicates superior goodness-of-fit and a lesser tendency to over-fit.
The model with the lower AIC score is expected to strike a superior balance between its ability to fit the data set and its ability to avoid over-fitting the data set. **This is the reason we prefer lower AIC score.**

References:

- https://towardsdatascience.com/introduction-to-aic-akaike-information-criterion-9c9ba1c96ced
- https://www.sciencedirect.com/topics/social-sciences/akaike-information-criterion
- https://towardsdatascience.com/the-akaike-information-criterion-c20c8fd832f2
- https://en.wikipedia.org/wiki/Akaike_information_criterion
- https://towardsdatascience.com/understanding-the-bias-variance-tradeoff-165e6942b229