Machine Learning in Healthcare 336546

HW3

Maya Fichmann Levital

300455433

14-01-2021

1. Clustering
   a. *Is K-medoid more robust to noise (or outliers) than the K-means algorithm? Explain your answer.*
   The K-medoid could be more robust to noise and outliers as compared to k-means because it minimizes a sum of general pairwise dissimilarities instead of a sum of squared Euclidean distances. If there is an outlier in the dataset it may affect the centroid chosen in the K-means whereas in the K-medoid the most central point is chosen (which is probably not an outlier). If there is white gaussian noise in the dataset it will be canceled by averaging the points in K-means. However, if the noise is distributed differently across samples the K-medoid will probably select the datapoint less effected by noise, as it is more similar to most of the points in the dataset, thereby making it more robust to it.

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

   The mean of m examples is:

$$mean\ of\ m\ examples = \sum_{i=1}^{m}\frac{x_i}{m}$$

   To find the minimum of the term $\sum_{i=1}^{m}(x_i-\mu)^2$ with respect to μ we will first derive it:

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

   We are looking for the minimal point, therefore we will compare the derivative to zero:

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

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

   The second derivative should be > 0:

$$\frac{d}{d\mu}\left(-2\left(\sum_{i=1}^{m}x_i - m\mu\right)\right) = 2m > 0$$

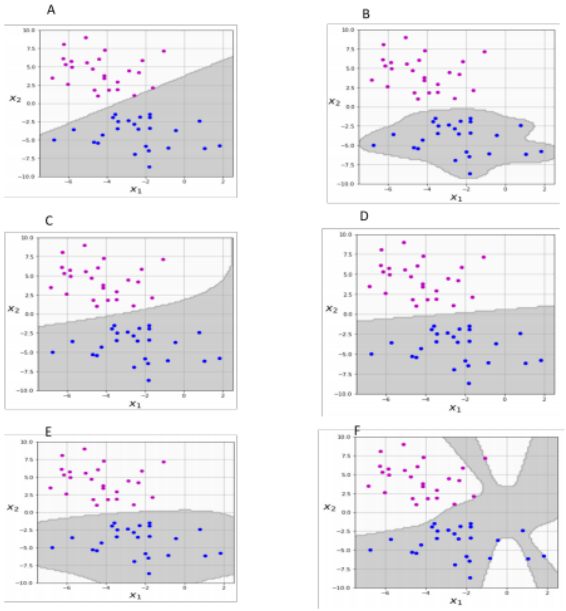
   c. Prove that the centroid (practically, the medoid) which minimizes the term $\sum_{i=1}^{m}|x_i-\mu|$ is the median of m examples given that μ belongs to the dataset.
   As before, we will derive the term and compare to zero:

$$\sum_{i=1}^{m}|x_i-\mu| = \sum_{i=1}^{m}\begin{cases} x_i-\mu\ for\ x_i-\mu \geq 0 \\ -(x_i-\mu)\ for\ x_i-\mu < 0 \end{cases}$$

$$\frac{d}{d\mu}\sum_{i=1}^{m}|x_i-\mu| = \sum_{i=1}^{m}\frac{d}{d\mu}|x_i-\mu| = -\sum_{i=1}^{m}\begin{cases} 1\ for\ x_i > \mu \\ -1\ for\ x_i < \mu \\ 0\ for\ x_i = \mu \end{cases}$$

Given that μ belongs to the dataset, for the point $x_i = \mu$ the term will go to zero. For all other points, the derivative will go to zero if and only if the number of points x_i larger than μ are the exactly the same as the number of points x_i smaller than μ. This is definition of median.

2. SVM:



The C parameter controls the SVM optimization margin. For large values of C, the optimization will choose a smaller-margin hyperplane if that hyperplane does a better job of getting all the training points classified correctly. Conversely, a very small value of C will cause the optimizer to look for a larger-margin separating hyperplane, even if that hyperplane misclassifies more points. In figure A the linear boundary is closer to the datapoints, meaning that there is less margin from the separating hyperplane to the datapoints. => C=1.

In figure D the linear boundary is more distant from the datapoints, meaning that the margin is larger. => C=0.01

Image C is a 2nd order polynomial as the separating boundary has a parabolic shape.

Image F is 10th order polynomial as the separating boundaries have the shape of a high order polynomial.

By using Gaussian RBF Kernel the boundary is radial, and this is what we can see on images B and E.

The γ parameter is the inverse of the standard deviation of the RBF kernel (Gaussian function), which is used as similarity measure between two points. A small gamma value defines a Gaussian function with a large variance. In this case, two points can be considered similar even if are far from each other. In the other hand, a large gamma value means defines a Gaussian function with a small variance and in this case, two points are considered similar just if they are close to each other.

Smaller values of γ would yield a less complex boundary, as in figure E whereas higher values of γ would yield a more complex boundary as in figure B.

1. D: Linear kernel with C = 0.01.
2. A: Linear kernel with C = 1.
3. C: 2nd order polynomial kernel.
4. F: 10th order polynomial kernel.
5. E: RBF kernel with γ = 0.2.
6. B: RBF kernel with γ = 1.

3. AIC:
   a. *What is the scientific term of the balance that Einstein meant to in machine learning aspect?*
      The scientific term of the balance that Einstein meant to in machine learning aspect is generalization. Generalization refers to how well the concepts learned by a machine learning model apply to specific examples not seen by the model when it was learning. A generalized model is balanced in terms of its complexity (the number of parameters used) and should not overfitted, nor underfitted. Overfitting refers to a model that is too complex, therefore models the training data too well. Overfitting happens when a model learns the detail and noise in the training data to the extent that it negatively impacts the performance of the model on new data. This means that the noise or random fluctuations in the training data is picked up and learned as concepts by the model. The problem is that these concepts do not apply to new data and negatively impact the model's ability to generalize. An overfitted model is not 'as simple as possible' as Einstein stated.
      Underfitting refers to a model that can neither model the training data nor generalize to new data. An underfit machine learning model is not a suitable model and will be obvious as it will have poor performance on the training data. An underfitted model is 'simpler' as Einstein stated.

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

      P in the number of parameters in the model. An overfitted model will have a large number of parameters, therefore increasing (penalizing) the AIC.

      Minimizing the $-2ln(\hat{L})$ is equivalent to maximizing the likelihood of the model on the training dataset. Therefore, underfitted model will have a larger value for $-ln(\hat{L})$, hence penalizing the AIC.

   c. *What are the two options that are likely to happen if this balance was violated?*
      If this balance is violated the model is either overfitting (as p is high) or underfitting as $-2ln(\hat{L})$ is high, meaning that the likelihood of the model is low on the training dataset.

   d. *What are we aiming for with the AIC? Should it be high or low? Explain.*
      When selecting a model we aim at minimizing AIC – penalizing for a model that is too complex and also for a model that has poor performance on the training set. It should be however noted that the AIC tends more towards selection of overfitting models as it penalizes less for highly parametrized models than Bayesian Information Criterion (BIC).