Solution KNN,SVM

June 20, 2024

1 solution KNN

Euclidean Distance

The Euclidean distance between two points (x_1, y_1) and (x_2, y_2) is given by:

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Let's calculate the Euclidean distance from the test point $(0.5, 0)$ to each of the points in the plot.

1. Distance to $(0, 0)$:

$$d = \sqrt{(0.5 - 0)^2 + (0 - 0)^2} = \sqrt{0.25} = 0.5$$

2. Distance to $(0, 1)$:

$$d = \sqrt{(0.5 - 0)^2 + (0 - 1)^2} = \sqrt{0.25 + 1} = \sqrt{1.25} \approx 1.118$$

3. Distance to $(0, -1)$:

$$d = \sqrt{(0.5 - 0)^2 + (0 - (-1))^2} = \sqrt{0.25 + 1} = \sqrt{1.25} \approx 1.118$$

4. Distance to $(2, 0)$:

$$d = \sqrt{(0.5 - 2)^2 + (0 - 0)^2} = \sqrt{2.25} \approx 1.5$$

5. Distance to $(3, 0)$:

$$d = \sqrt{(0.5 - 3)^2 + (0 - 0)^2} = \sqrt{6.25} = 2.5$$

6. Distance to $(1, 1)$:

$$d = \sqrt{(0.5 - 1)^2 + (0 - 1)^2} = \sqrt{0.25 + 1} = \sqrt{1.25} \approx 1.118$$

7. Distance to $(1, -1)$:

$$d = \sqrt{(0.5-1)^2 + (0-(-1))^2} = \sqrt{0.25+1} = \sqrt{1.25} \approx 1.118$$

So the three closest points are $(0,0)$, $(0,1)$, and $(0,-1)$ with distances of 0.5, 1.118, and 1.118 respectively. Among these points, $(0,0)$ and $(0,1)$ belong to Class 2 (dots), and $(0,-1)$ belongs to Class 2 as well. Since 2 out of the 3 nearest neighbors are from Class 2, the test point $(0.5,0)$ will be classified as **Class 2**.

Manhattan Distance

The Manhattan distance between two points (x_1, y_1) and (x_2, y_2) is given by:

$$d = |x_2 - x_1| + |y_2 - y_1|$$

Let's calculate the Manhattan distance from the test point $(0.5,0)$ to each of the points in the plot.

1. Distance to $(0,0)$:
$$d = |0.5-0| + |0-0| = 0.5$$
2. Distance to $(0,1)$:
$$d = |0.5-0| + |0-1| = 0.5+1 = 1.5$$
3. Distance to $(0,-1)$:
$$d = |0.5-0| + |0-(-1)| = 0.5+1 = 1.5$$
4. Distance to $(2,0)$:
$$d = |0.5-2| + |0-0| = 1.5$$
5. Distance to $(3,0)$:
$$d = |0.5-3| + |0-0| = 2.5$$
6. Distance to $(1,1)$:
$$d = |0.5-1| + |0-1| = 0.5+1 = 1.5$$
7. Distance to $(1,-1)$:
$$d = |0.5-1| + |0-(-1)| = 0.5+1 = 1.5$$

So the three closest points using Manhattan distance are $(0,0)$, and any two from $(0,1)$, $(0,-1)$, $(2,0)$, $(1,1)$, or $(1,-1)$ with distances of 0.5 and 1.5. From these, the point $(0,0)$ belongs to Class 2, and the others are mixed, but Class 2 still dominates the majority. Therefore, the test point $(0.5,0)$ will be classified as **Class 2** by both Euclidean and Manhattan distance methods.

2 SVM

1. Explain what is meant by support vectors and show an example.

Support vectors are the data points that are closest to the decision surface (or hyperplane) in a Support Vector Machine (SVM) model. These points are crucial because they define the position and orientation of the hyperplane. Removing these points would change the position of the hyperplane, which is not the case for the other points.

Example: Consider a simple binary classification problem with two features. Imagine we have two classes of points: red and blue. The SVM algorithm finds the optimal hyperplane that maximizes the margin between these two classes. The support vectors are the red and blue points that lie closest to this hyperplane.

2. For what types of data is the SVM classification inappropriate in your opinion?

SVM might not be suitable for the following types of data:

- **High noise data:** If the data has a lot of noise and overlaps, SVM might struggle to find a clear margin of separation.
- **Large datasets:** SVMs can be computationally intensive, making them less practical for very large datasets.
- **Non-linear data without a proper kernel:** If the data is not linearly separable and an inappropriate kernel is used, SVM may not perform well.

3. Explain kernels and their role in classification.

Kernels are functions that transform the input data into a higher-dimensional space where it becomes easier to classify with a linear hyperplane. The main idea is to make non-linearly separable data linearly separable by mapping it to a higher dimension.

- **Linear Kernel:** Used for linearly separable data.
- **Polynomial Kernel:** Suitable for data that is polynomially separable.
- **Radial Basis Function (RBF) Kernel:** Effective in high-dimensional spaces.
- **Sigmoid Kernel:** Acts like a neural network.

4. Explain the difference between hard SVM classifier and soft SVM classifier.

- **Hard SVM Classifier:** Assumes that the data is linearly separable and aims to find a hyperplane that separates the data without any errors.
- **Soft SVM Classifier:** Allows some misclassifications to handle cases where data is not perfectly linearly separable. It introduces a penalty for misclassified points to create a margin that tolerates some level of misclassification.

5. Explain how SVM can be used in regression problems with an illustration.

SVM can be adapted for regression tasks using a technique called Support Vector Regression (SVR). In SVR, the goal is to find a function that deviates from the actual target values by a value no greater than epsilon (ϵ) for all training data, while simultaneously being as flat as possible.

Illustration: In SVR, instead of trying to fit the widest possible street between two classes, we fit a tube of width 2ϵ around the data points. Points outside this tube are considered errors and penalized. The support vectors in this context are the points that lie on the boundary of the tube.

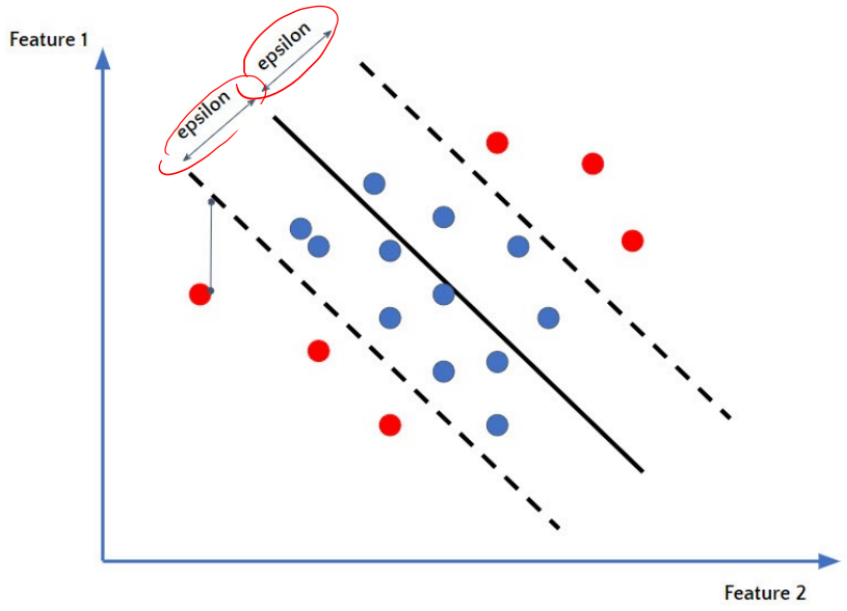

Figure 1: Enter Caption